Sheet 2 Theory

Exercise 1: Maximum-Likelihood Estimation

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 $\mathcal{D} = ((x_1,y_1), \dots, (x_N,y_N))$ composed of N independent draws from this distribution.

a.) Show that x and y are independent.

The variables x and y are statistically independent if the following condition is fulfilled:

$$p(x,y) = p(x) \cdot p(y) \quad , \forall (x,y) \in \mathbb{R}_+^2$$

To calculate $p(x)$ we have to integrate the given probability distribution $p(x,y)$ over the y and vise versa.

$$p(x) = \int_0^\infty \lambda\eta \, e^{-\lambda x - \eta y} \, dy = \lambda\eta \, e^{-\lambda x} \left(-\frac{1}{\eta}\right) \left[e^{-\eta y}\right]_0^\infty = -\lambda \, e^{-\lambda x} \left[0 - 1\right]$$

$$= \lambda \, e^{-\lambda x}$$

$$p(y) = \int_0^\infty \lambda\eta \, e^{-\lambda x - \eta y} \, dx = \lambda\eta \, e^{-\eta y} \, \frac{1}{-\lambda} \, (-1) = \eta \, e^{-\eta y}$$

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

$\rightarrow$ x and y are independent! ☺

b) Derive a Maximum Likelihood estimator of the parameter λ based on $\mathcal{D}$

General Maximum Likelihood

- Dataset: $\mathcal{D} = (x_1, \dots, x_n)$
- Likelihood under iid. assumption: $p(\mathcal{D} \mid \theta) = \prod_{k=1}^{n} p(x_k, \theta)$
- Log-Likelihood: $\ell(\theta) = \log p(\mathcal{D} \mid \theta)$
- Maximum Likelihood parameter: $\hat{\theta} = \arg\max_{\theta} \ell(\theta)$

We define a log-likelihood function based on the Dataset $\mathcal{D} = ((x_1, y_1), \dots, (x_N, y_N))$

$$\mathcal{L}(\theta) = \ln p(\mathcal{D}|\theta) = \ln \prod_{k=1}^{n} p((x_n, y_n), \theta) = \sum_{k=1}^{n} \ln p((x_n, y_n), \theta)$$

$$p(\mathcal{D}|\theta) = \prod_{k=1}^{n} p(x_n|\theta) \qquad \ln(\prod \dots) = \sum \ln$$

$$\hat{\lambda} = \arg\max_{\lambda} \mathcal{L}(\lambda) = \arg\max_{\lambda} \sum_{k=1}^{n} \ln(\lambda\eta e^{-\lambda x_k - \eta y_k})$$

$$= \arg\max_{\lambda} N \ln(\lambda\eta) - \sum_{k=1}^{N} (\lambda x_k - \eta y_k)$$

$$\frac{\partial \mathcal{L}(\lambda)}{\partial \lambda} = \frac{N}{\lambda} - \sum_{k=1}^{N} x_k \overset{!}{=} 0$$

$$\hookrightarrow \hat{\lambda} = \sum_{k=1}^{N} \frac{N}{x_k}$$

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

$$\mathcal{L}(\lambda) = \sum_{k=1}^{N} \ln\left(e^{-\lambda x_k - \eta y_k}\right) = -\sum_{k=1}^{N} \lambda x_k - \frac{1}{\lambda} y_k$$

$$\frac{\partial \mathcal{L}(\lambda)}{\partial \lambda} = -\sum_{k=1}^{N} x_k + \frac{1}{\lambda^2} y_k \overset{!}{=} 0$$

$$\frac{1}{\lambda^2} \sum_{k=1}^{N} y_k = \sum_{k=1}^{N} x_k \longrightarrow \hat{\lambda} = \sqrt{\sum_{k=1}^{N} \frac{y_k}{x_k}}$$

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

$$\mathcal{L}(\lambda) = \sum_{k=1}^{N} \ln \lambda(1-\lambda)\, e^{-\lambda x_k - (1-\lambda) y_k}$$

$$= N \ln(\lambda - \lambda^2) - \sum_{k=1}^{N} \lambda x_k + y_k - \lambda y_k$$

$$\frac{\partial \mathcal{L}(\lambda)}{\partial \lambda} = \frac{N(1-2\lambda)}{\lambda - \lambda^2} - \sum_{k=1}^{N} (x_k - y_k)$$

$$0 = N - 2N\lambda - \sum_{k=1}^{N} (x_k - y_k)\cdot\lambda + \sum_{k=1}^{N} (x_k - y_k)\lambda^2 \qquad \text{p/q-Formel}$$

$$= \underbrace{\sum_{k=1}^{N} (x_k - y_k)}_{a} \lambda^2 \underbrace{- \sum_{k=1}^{N} (x_k - y_k) - 2N}_{b} \lambda + \underbrace{N}_{c}$$

$$\lambda_{1/2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} = \sum_{k=1}^{N} \frac{(x_k - y_k) + 2N \pm \sqrt{(x_k - y_k)^2 + 4N^2}}{2(x_k - y_k)}$$

Nebenrechnung:

$$\sum_{k=1}^{N} \left[(x_k - y_k)^2 + 4N(x_k - y_k) + 4N^2 - 4(x_k - y_k)N \right]$$

We know $\lambda > 0$:

$$(x_k - y_k) + 2N - \sqrt{(x_k - y_k)^2 + 4N^2} > 0$$

$$(x_k - y_k) + 2N > \sqrt{(x_k + y_k)^2 + 4N^2}$$

$$\sqrt{((x_k - y_k) + 2N)^2} = \sqrt{(x_k + y_k)^2 + 4N(x_k - y_k) + 4N^2} > \sqrt{(x_k + y_k)^2 + 4N^2}$$

$$4N(x_k + y_k) > 0 \quad \checkmark$$

→ both solutions fulfil the condition $\lambda > 0$ ☺

Exercise 2 Maximum Likelihood vs. Bayes

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

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

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

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

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

a.) State the likelihood function $P(\mathcal{D}|\theta)$, that depends on the parameter θ
↳ not log-likelihood ☺

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

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

$$P(x_8 = \text{head}, x_9 = \text{head} \,|\, \hat{\theta})$$

→ Maximum Likelihood estimator

$$\mathcal{L}(\theta) = \ln \theta^5 (1-\theta)^2$$

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{5\theta^4(1-\theta)^2 - 2\theta^5(1-\theta)}{\theta^5(1-\theta)^2} = \frac{5 - 7\theta}{\theta - \theta^2} \stackrel{!}{=} 0 \qquad 5 = 7\theta \rightarrow \underline{\underline{\theta = \frac{5}{7}}}$$

From this we can evaluate $\underbrace{P(x_8 = head, x_9 = head \mid \hat{\theta})}_{P(D|\theta) = \prod_k P(x_k|\theta)} \rightarrow D = \{x_8, x_9\}$

$P(x_8 = head, x_9 = head \mid \hat{\theta}) = \hat{\theta} \cdot \hat{\theta} = \hat{\theta} = \frac{25}{49}$.

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

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

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

$$\int P(x_8 = head, x_9 = head \mid \theta)\, p(\theta|D)\, d\theta$$

General Bayesian Estimation

- Bayes formula: $p(\theta|D) = \frac{P(D|\theta)\, p(\theta)}{p(D)} = \frac{p(D|\theta)\, p(\theta)}{\int p(D|\theta)\, p(\theta)\, d\theta}$
- Probability of new data point: $p(x|D) = \int p(x|\theta)\, p(\theta|D)\, d\theta$

posterior distribution: $p(\theta|D) = \frac{P(D|\theta) \cdot p(\theta)}{P(D)}$ $\qquad P(D|\theta) = \theta^5 (1-\theta)^2$

$$p(D) = \int p(D|\theta) \cdot p(\theta)\, d\theta = \int_0^1 \theta^5 (1-\theta)^2 \cdot 1\, d\theta = \int_0^1 \theta^5 (1 - 2\theta + \theta^2)\, d\theta$$

$$= \left[\frac{1}{6}\theta^6 - \frac{2}{7}\theta^7 + \frac{1}{8}\theta^8 \right]_0^1 = \frac{1}{168}$$

$$p(\theta|D) = 168\, \theta^5 (1-\theta)^2$$

Our new probability for the next two tosses is

$$P(x_8 = head, x_9 = head) = \int P(x_8 = head, x_9 = head \mid \theta)\, p(\theta|D)\, d\theta$$

$$= \int_0^1 \theta^2\, 168\, \theta^5 (1-\theta)^2\, d\theta = 168 \int_0^1 \theta^7 - 2\theta^8 + \theta^9\, d\theta$$

$$= 168 \left[\frac{1}{8}\theta^8 - \frac{2}{9}\theta^9 + \frac{1}{10}\theta^{10} \right]_0^1$$

$$= 168 \cdot \frac{1}{360} = \underline{\underline{\frac{7}{15}}}$$

Exercise 3 Convergence of Bayes Parameter Estimation

We consider Section 3.4.1 of Duda et. al., where the data is generated according to the univariate probability density $p(x|\mu) \sim N(\mu, \theta^2)$, where θ^2 is known and where μ is unknown with prior distribution $p(\mu) \sim N(\mu_0, \theta_0^2)$. Having sampled a dataset $\mathcal{D}$ from the data-generating distribution, the posterior probability distribution over the unknown parameter μ becomes $p(\mu|\mathcal{D}) \sim N(\mu_n, \theta_n^2)$, where

$$\frac{1}{\theta_n^2} = \frac{n}{\theta^2} + \frac{1}{\theta_0^2} \qquad \frac{\mu_n}{\theta_n^2} = \frac{n}{\theta^2}\hat{\mu}_n + \frac{\mu_0}{\theta_0^2} \qquad \hat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} x_k$$

a.) Show that the variance of the posterior can be upper-bounded as follows:

$$\theta_n^2 \leq \min\left(\frac{\theta^2}{n}, \theta_0^2\right)$$

that is, the variance of the posterior is contained both by the uncertainty of the data mean and of the prior.

Solution: $\frac{1}{\theta_n^2} = \frac{\theta_0^2 n + \theta^2}{\theta^2\theta_0^2} \rightarrow$ variance of the posterior $\theta_n^2 = \frac{\theta^2\theta_0^2}{n\theta_0^2 + \theta^2}$

Posterior can be upper bounded:

$$\frac{\theta^2\theta_0^2}{n\theta_0^2 + \theta^2} \leq \min\left(\frac{\theta^2}{n}, \theta_0^2\right)$$

Case 1: $\frac{\theta^2}{n} < \theta_0^2 \rightarrow \min\left(\frac{\theta^2}{n}, \theta_0^2\right) = \frac{\theta^2}{n}$

$$\frac{\theta^2\theta_0^2}{n\theta_0^2 + \theta^2} \leq \frac{\theta^2}{n}$$

$$n\theta^2\theta_0^2 \leq \theta^2(n\theta_0^2 + \theta^2)$$

$$0 \leq \frac{\theta^2}{n} \rightarrow \text{with } \theta^2 \geq 0 \text{ and } n \in \mathbb{N}^+ \text{ always true} \smile$$

Case 2: $\theta_0^2 < \frac{\theta^2}{n} \rightarrow \min(\ldots) = \theta_0^2$

$$\frac{\theta^2\theta_0^2}{n(\theta_0^2 + \frac{\theta^2}{n})} \leq \theta_0^2$$

$$\theta^2 \leq n\theta_0^2 + \theta^2 \rightarrow 0 \leq n\theta_0^2 \text{ with } n \in \mathbb{N}^+ \text{ and } \theta_0^2 \geq 0 \text{ always true}$$

$\Rightarrow$ The variance of the posterior can be upper bounded by $\theta_n^2 \leq \min\left(\frac{\theta^2}{n}, \theta_0^2\right)$

Good to know! For the Gaussian distribution $N(\mu, \theta^2)$ we know

$\mu \in \mathbb{R}$!
$\theta^2 > 0$.

b.) Show that the mean of the posterior can be lower- and upper bounded as follows:

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

that is, the mean of the posterior distribution lies somewhere on the segment between the mean of the prior distribution and the sample mean.

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

↳ from a.) we know $\sigma_n^2 = \frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2}$

$$\mu_n = \frac{\sigma^2\sigma_0^2}{n\sigma_0^2+\sigma^2} \cdot \frac{n\hat{\mu}_n\sigma_0^2 + \mu_0\sigma^2}{\sigma^2\sigma_0^2} = \frac{n\hat{\mu}_n\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2}$$

Case 1: $\mu_0 < \hat{\mu}_n$

$$\mu_0 \leq \frac{n\hat{\mu}_n\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2} \leq \hat{\mu}_n$$

$\mu_0(n\sigma_0^2 + \sigma^2) \leq n\hat{\mu}_n\sigma_0^2 + \mu_0\sigma^2$

$\mu_0 n\sigma_0^2 + \mu_0\sigma^2 \leq n\hat{\mu}_n\sigma_0^2 + \mu_0\sigma^2$

$\mu_0 \leq \hat{\mu}_n$

$n\hat{\mu}_n\sigma_0^2 + \mu_0\sigma^2 \leq \hat{\mu}_n(n\sigma_0^2 + \sigma^2)$

$n\hat{\mu}_n\sigma_0^2 + \mu_0\sigma^2 \leq n\hat{\mu}_n\sigma_0^2 + \hat{\mu}_n\sigma^2$

$\mu_0 \leq \hat{\mu}_n$

↳ both true, because of the assumption $\mu_0 < \hat{\mu}_n$ ☺

Case 2: $\hat{\mu}_n < \mu_0$

$$\hat{\mu}_n \leq \frac{n\hat{\mu}_n\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2} \leq \mu_0$$

$\hat{\mu}_n n\sigma_0^2 + \hat{\mu}_n\sigma^2 \leq n\hat{\mu}_n\sigma_0^2 + \mu_0\sigma^2$

$\hat{\mu}_n \leq \mu_0$

$n\hat{\mu}_n\sigma_0^2 + \mu_0\sigma^2 \leq \mu_0 n\sigma_0^2 + \mu_0\sigma^2$

$\hat{\mu}_n \leq \mu_0$

↳ both true, because of the assumption $\hat{\mu}_n < \mu_0$ ☺