Exercise 1. ***Maximum likelihood estimation for a Gaussian***

The Gaussian pdf parametrised by mean μ and standard deviation σ is given by

$$p(x;\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \qquad \boldsymbol{\theta} = (\mu, \sigma).$$

(a) *Given iid data $\mathcal{D} = \{x_1, \ldots, x_n\}$, what is the likelihood function $L(\boldsymbol{\theta})$ for the Gaussian model?*

Solution. For iid data, the likelihood function is

$$L(\boldsymbol{\theta}) = \prod_i^n p(x_i;\boldsymbol{\theta}) \tag{S.1}$$

$$= \prod_i^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right] \tag{S.2}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\right]. \tag{S.3}$$

(b) *What is the log-likelihood function $\ell(\boldsymbol{\theta})$?*

Solution. Taking the log of the likelihood function gives

$$\ell(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2 \tag{S.4}$$

(c) *Show that the maximum likelihood estimates for the mean μ and standard deviation σ are the sample mean*

$$\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i \tag{1}$$

and the square root of the sample variance

$$S^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2. \tag{2}$$

Solution. Since the logarithm is strictly monotonically increasing, the maximiser of the log-likelihood equals the maximiser of the likelihood. It is easier to take derivatives for the log-likelihood function than for the likelihood function so that the maximum likelihood estimate is typically determined using the log-likelihood.

Given the algebraic expression of $\ell(\boldsymbol{\theta})$, it is simpler to work with the variance $v = \sigma^2$ rather than the standard deviation. (In the lecture notes, we used the variable η to denote the transformed parameters. We could have written $\eta = \sigma^2$, but v is a more natural notation for the variance.) Since $\sigma > 0$ the function $v = g(\sigma) = \sigma^2$ is invertible, and the invariance of the MLE to re-parametrisation guarantees that

$$\hat{\sigma} = \sqrt{\hat{v}}.$$

We now thus maximise the function $J(\mu, v)$,

$$J(\mu, v) = -\frac{n}{2}\log(2\pi v) - \frac{1}{2v}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{S.5}$$

with respect to μ and v.
Taking partial derivatives gives

$$\begin{aligned}\frac{\partial J}{\partial \mu} &= \frac{1}{v}\sum_{i=1}^{n}(x_i - \mu) && \text{(S.6)}\\ &= \frac{1}{v}\sum_{i=1}^{n} x_i - \frac{n}{v}\mu && \text{(S.7)}\\ \frac{\partial J}{\partial v} &= -\frac{n}{2}\frac{1}{v} + \frac{1}{2v^2}\sum_{i=1}^{n}(x_i - \mu)^2 && \text{(S.8)}\end{aligned}$$

A necessary condition for optimality is that the partial derivatives are zero. We thus obtain the conditions

$$\begin{aligned}\frac{1}{v}\sum_{i=1}^{n}(x_i - \mu) &= 0 && \text{(S.9)}\\ -\frac{n}{2}\frac{1}{v} + \frac{1}{2v^2}\sum_{i=1}^{n}(x_i - \mu)^2 &= 0 && \text{(S.10)}\end{aligned}$$

From the first condition it follows that

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{S.11}$$

The second condition thus becomes

$$\begin{aligned}-\frac{n}{2}\frac{1}{v} + \frac{1}{2v^2}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 &= 0 \qquad \text{(multiply with } v^2 \text{ and rearrange)} && \text{(S.12)}\\ \frac{1}{2}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 &= \frac{n}{2}v, && \text{(S.13)}\end{aligned}$$

and hence

$$\hat{v} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2, \tag{S.14}$$

We now check that this solution corresponds to a maximum by computing the Hessian matrix

$$\mathbf{H}(\mu, v) = \begin{pmatrix} \frac{\partial^2 J}{\partial \mu^2} & \frac{\partial^2 J}{\partial \mu \partial v} \\ \frac{\partial^2 J}{\partial \mu \partial v} & \frac{\partial^2 J}{\partial v^2} \end{pmatrix} \tag{S.15}$$

If the Hessian negative definite at $(\hat{\mu}, \hat{v})$, the point is a (local) maximum. Since we only have one critical point, $(\hat{\mu}, \hat{v})$, the local maximum is also a global maximum. Taking second derivatives gives

$$\mathbf{H}(\mu, v) = \begin{pmatrix} -\frac{n}{v} & -\frac{1}{v^2}\sum_{i=1}^{n}(x_i - \mu) \\ -\frac{1}{v^2}\sum_{i=1}^{n}(x_i - \mu) & \frac{n}{2}\frac{1}{v^2} - \frac{1}{v^3}\sum_{i=1}^{n}(x_i - \mu)^2 \end{pmatrix}. \tag{S.16}$$

Substituting the values for $(\hat{\mu}, \hat{v})$ gives

$$\mathbf{H}(\hat{\mu}, \hat{v}) = \begin{pmatrix} -\frac{n}{\hat{v}} & 0 \\ 0 & -\frac{n}{2}\frac{1}{\hat{v}^2} \end{pmatrix}, \tag{S.17}$$

which is negative definite. Note that the the (negative) curvature increases with n, which means that $J(\mu, v)$, and hence the log-likelihood becomes more and more peaked as the number of data points n increases.

Exercise 2. ***Posterior of the mean of a Gaussian with known variance***

Given iid data $\mathcal{D} = \{x_1, \ldots, x_n\}$, compute $p(\mu|\mathcal{D}, \sigma^2)$ for the Bayesian model

$$p(x|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \qquad p(\mu; \mu_0, \sigma_0^2) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left[-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right] \tag{3}$$

where σ^2 is a fixed known quantity.

Solution. Recall the following result from Tutorial 6:

$$\mathcal{N}(x; m_1, \sigma_1^2)\mathcal{N}(x; m_2, \sigma_2^2) \propto \mathcal{N}(x; m_3, \sigma_3^2) \tag{S.18}$$

where

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{S.19}$$

$$\sigma_3^2 = \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)^{-1} = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \tag{S.20}$$

$$m_3 = \sigma_3^2\left(\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}\right) = m_1 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}(m_2 - m_1) \tag{S.21}$$

We can further re-use the expression for the likelihood $L(\mu)$ from the previous exercise,

$$L(\mu) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right], \tag{S.22}$$

which we can write as

$$L(\mu) \propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right] \tag{S.23}$$

$$\propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i^2 - 2\mu x_i + \mu^2)\right] \tag{S.24}$$

$$\propto \exp\left[-\frac{1}{2\sigma^2}\left(-2\mu\sum_{i=1}^{n}x_i + n\mu^2\right)\right] \tag{S.25}$$

$$\propto \exp\left[-\frac{1}{2\sigma^2}\left(-2n\mu\bar{x} + n\mu^2\right)\right] \tag{S.26}$$

$$\propto \exp\left[-\frac{n}{2\sigma^2}(\mu - \bar{x})^2\right] \tag{S.27}$$

$$\propto \mathcal{N}(\mu; \bar{x}, \sigma^2/n). \tag{S.28}$$

The posterior is

$$p(\mu|\mathcal{D}) \propto L(\theta)p(\mu;\mu_0,\sigma_0^2) \tag{S.29}$$
$$\propto \mathcal{N}(\mu;\bar{x},\sigma^2/n)\mathcal{N}(\mu;\mu_0,\sigma_0^2) \tag{S.30}$$

so that with (S.18), we have

$$p(\mu|\mathcal{D}) \propto \mathcal{N}(\mu;\mu_n,\sigma_n^2) \tag{S.31}$$
$$\sigma_n^2 = \left(\frac{1}{\sigma^2/n} + \frac{1}{\sigma_0^2}\right)^{-1} \tag{S.32}$$
$$= \frac{\sigma_0^2\sigma^2/n}{\sigma_0^2 + \sigma^2/n} \tag{S.33}$$
$$\mu_n = \sigma_n^2\left(\frac{\bar{x}}{\sigma^2/n} + \frac{\mu_0}{\sigma_0^2}\right) \tag{S.34}$$
$$= \frac{1}{\sigma_0^2 + \sigma^2/n}\left(\sigma_0^2\bar{x} + (\sigma^2/n)\mu_0\right) \tag{S.35}$$
$$= \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2/n}\bar{x} + \frac{\sigma^2/n}{\sigma_0^2 + \sigma^2/n}\mu_0 \tag{S.36}$$

which are the expressions given in the lecture slides. As n increases, σ^2/n goes to zero so that $\sigma_n^2 \to 0$ and $\mu_n \to \bar{x}$. This means that with an increasing amount of data, the posterior of the mean tends to be concentrated around the maximum likelihood estimate $\bar{x}$.

From (S.21), we also have that

$$\mu_n = \mu_0 + \frac{\sigma_0^2}{\sigma^2/n + \sigma_0^2}(\bar{x} - \mu_0), \tag{S.37}$$

which shows more clearly that the value of μ_n lies on a line with end-points μ_0 (for $n = 0$) and $\bar{x}$ (for $n \to \infty$). As the amount of data increases, μ_n moves form the mean under the prior, μ_0, to the average of the observed sample, that is the MLE $\bar{x}$.

Exercise 3. ***Maximum likelihood estimation of probability tables in fully observed directed graphical models of binary variables***

We assume that we are given a parametrised directed graphical model for variables $x_1, \ldots, x_d$,

$$p(\mathbf{x};\boldsymbol{\theta}) = \prod_{i=1}^{d} p(x_i|\mathrm{pa}_i;\boldsymbol{\theta}_i) \qquad x_i \in \{0,1\} \tag{4}$$

where the conditionals are represented by parametrised probability tables, For example, if $\mathrm{pa}_3 = \{x_1, x_2\}$, $p(x_3|\mathrm{pa}_3;\boldsymbol{\theta}_3)$ is represented as

$p(x_3 = 1\|x_1, x_2; \theta_3^1, \ldots, \theta_3^4))$	x_1	x_2
θ_3^1	*0*	*0*
θ_3^2	*1*	*0*
θ_3^3	*0*	*1*
θ_3^4	*1*	*1*

with $\boldsymbol{\theta}_3 = (\theta_3^1, \theta_3^2, \theta_3^3, \theta_3^4)$, and where the superscripts j of θ_3^j enumerate the different states that the parents can be in.

(a) *Assuming that x_i has m_i parents, verify that the table parametrisation of $p(x_i|\text{pa}_i;\boldsymbol{\theta}_i)$ is equivalent to writing $p(x_i|\text{pa}_i;\boldsymbol{\theta}_i)$ as*

$$p(x_i|\text{pa}_i;\boldsymbol{\theta}_i) = \prod_{s=1}^{S_i} (\theta_i^s)^{\mathbb{1}(x_i=1,\text{pa}_i=s)}(1-\theta_i^s)^{\mathbb{1}(x_i=0,\text{pa}_i=s)} \tag{5}$$

where $S_i = 2^{m_i}$ is the total number of states/configurations that the parents can be in, and $\mathbb{1}(x_i = 1, \text{pa}_i = s)$ is one if $x_i = 1$ and $\text{pa}_i = s$, and zero otherwise.

Solution. The number of configurations that m binary parents can be in is given by S_i. The questions thus boils down to showing that $p(x_i = 1|\text{pa}_i = k;\boldsymbol{\theta}_i) = \theta_i^k$ for any state $k \in \{1,\dots,S_i\}$ of the parents of x_i. Since $\mathbb{1}(x_i = 1, \text{pa}_i = s) = 0$ unless $s = k$, we have indeed that

$$p(x_i = 1|\text{pa}_i = k;\boldsymbol{\theta}_i) = \left[\prod_{s\neq k}(\theta_i^s)^0(1-\theta_i^s)^0\right](\theta_i^k)^{\mathbb{1}(x_i=1,\text{pa}_i=k)}(1-\theta_i^k)^{\mathbb{1}(x_i=0,\text{pa}_i=k)} \tag{S.38}$$

$$= 1\cdot(\theta_i^k)^{\mathbb{1}(x_i=1,\text{pa}_i=k)}(1-\theta_i^k)^0 \tag{S.39}$$

$$= \theta_i^k. \tag{S.40}$$

(b) *For iid data $\mathcal{D} = \{\mathbf{x}^{(1)},\dots,\mathbf{x}^{(n)}\}$ show that*

$$p(\mathcal{D};\boldsymbol{\theta}) = \prod_{i=1}^{d}\prod_{s=1}^{S_i}(\theta_i^s)^{n^s_{x_i=1}}(1-\theta_i^s)^{n^s_{x_i=0}} \tag{6}$$

where $n^s_{x_i=1}$ is the number of times the pattern $(x_i = 1, \text{pa}_i = s)$ occurs in the data $\mathcal{D}$, and equivalently for $n^s_{x_i=0}$.

Solution. Since the data are iid, we have

$$p(\mathcal{D};\boldsymbol{\theta}) = \prod_{j=1}^{n} p(\mathbf{x}^{(j)};\boldsymbol{\theta}) \tag{S.41}$$

$$\tag{S.42}$$

where each term $p(\mathbf{x}^{(j)};\boldsymbol{\theta})$ factorises as in (4),

$$p(\mathbf{x}^{(j)};\boldsymbol{\theta}) = \prod_{i=1}^{d} p(x_i^{(j)}|\text{pa}_i^{(j)};\boldsymbol{\theta}_i) \tag{S.43}$$

with $x_i^{(j)}$ denoting the i-th element of $\mathbf{x}^{(j)}$ and $\text{pa}_i^{(j)}$ the corresponding parents. The conditionals $p(x_i^{(j)}|\text{pa}_i^{(j)};\boldsymbol{\theta}_i)$ factorise further according to (5),

$$p(x_i^{(j)}|\text{pa}_i^{(j)};\boldsymbol{\theta}_i) = \prod_{s=1}^{S_i}(\theta_i^s)^{\mathbb{1}(x_i^{(j)}=1,\text{pa}_i^{(j)}=s)}(1-\theta_i^s)^{\mathbb{1}(x_i^{(j)}=0,\text{pa}_i^{(j)}=s)}, \tag{S.44}$$

so that

$$p(\mathcal{D};\boldsymbol{\theta}) = \prod_{j=1}^{n}\prod_{i=1}^{d} p(x_i^{(j)}|\text{pa}_i^{(j)};\boldsymbol{\theta}_i) \tag{S.45}$$

$$= \prod_{i=1}^{d}\prod_{s=1}^{S_i}(\theta_i^s)^{\mathbb{1}(x_i^{(j)}=1,\text{pa}_i^{(j)}=s)}(1-\theta_i^s)^{\mathbb{1}(x_i^{(j)}=0,\text{pa}_i^{(j)}=s)} \tag{S.46}$$

Swapping the order of the products so that the product over the data points comes first, we obtain

$$p(\mathcal{D}; \boldsymbol{\theta}) = \prod_{i=1}^{d} \prod_{s=1}^{S_i} \prod_{j=1}^{n} (\theta_i^s)^{\mathbb{1}(x_i^{(j)}=1, \mathrm{pa}_i^{(j)}=s)} (1-\theta_i^s)^{\mathbb{1}(x_i^{(j)}=0, \mathrm{pa}_i^{(j)}=s)} \tag{S.47}$$

We next split the product over j into two products, one for all j where $x_i^{(j)} = 1$, and one for all j where $x_i^{(j)} = 0$

$$\begin{aligned}
p(\mathcal{D}; \boldsymbol{\theta}) &= \prod_{i=1}^{d} \prod_{s=1}^{S_i} \prod_{\substack{j: \\ x_i^{(j)}=1}} \prod_{\substack{j: \\ x_i^{(j)}=0}} (\theta_i^s)^{\mathbb{1}(x_i^{(j)}=1, \mathrm{pa}_i^{(j)}=s)} (1-\theta_i^s)^{\mathbb{1}(x_i^{(j)}=0, \mathrm{pa}_i^{(j)}=s)} && \text{(S.48)} \\
&= \prod_{i=1}^{d} \prod_{s=1}^{S_i} \prod_{\substack{j: \\ x_i^{(j)}=1}} (\theta_i^s)^{\mathbb{1}(x_i^{(j)}=1, \mathrm{pa}_i^{(j)}=s)} \prod_{\substack{j: \\ x_i^{(j)}=0}} (1-\theta_i^s)^{\mathbb{1}(x_i^{(j)}=0, \mathrm{pa}_i^{(j)}=s)} && \text{(S.49)} \\
&= \prod_{i=1}^{d} \prod_{s=1}^{S_i} (\theta_i^s)^{\sum_{j=1}^{n} \mathbb{1}(x_i^{(j)}=1, \mathrm{pa}_i^{(j)}=s)} (1-\theta_i^s)^{\sum_{j=1}^{n} \mathbb{1}(x_i^{(j)}=0, \mathrm{pa}_i^{(j)}=s)} && \text{(S.50)} \\
&= \prod_{i=1}^{d} \prod_{s=1}^{S_i} (\theta_i^s)^{n_{x_i=1}^s} (1-\theta_i^s)^{n_{x_i=0}^s} && \text{(S.51)}
\end{aligned}$$

where

$$n_{x_i=1}^s = \sum_{j=1}^{n} \mathbb{1}(x_i^{(j)} = 1, \mathrm{pa}_i^{(j)} = s) \qquad n_{x_i=0}^s = \sum_{j=1}^{n} \mathbb{1}(x_i^{(j)} = 0, \mathrm{pa}_i^{(j)} = s) \tag{S.52}$$

is the number of times $x_i = 1$ and $x_i = 0$, respectively, with its parents being in state s.

(c) *Show that the log-likelihood decomposes into sums of terms that can be independently optimised, and that each term corresponds to the log-likelihood for a Bernoulli model.*

Solution. The log-likelihood $\ell(\boldsymbol{\theta})$ equals

$$\begin{aligned}
\ell(\boldsymbol{\theta}) &= \log p(\mathcal{D}; \boldsymbol{\theta}) && \text{(S.53)} \\
&= \log \prod_{i=1}^{d} \prod_{s=1}^{S_i} (\theta_i^s)^{n_{x_i=1}^s} (1-\theta_i^s)^{n_{x_i=0}^s} && \text{(S.54)} \\
&= \sum_{i=1}^{d} \sum_{s=1}^{S_i} \log \left[(\theta_i^s)^{n_{x_i=1}^s} (1-\theta_i^s)^{n_{x_i=0}^s} \right] && \text{(S.55)} \\
&= \sum_{i=1}^{d} \sum_{s=1}^{S_i} n_{x_i=1}^s \log(\theta_i^s) + n_{x_i=0}^s \log(1-\theta_i^s) && \text{(S.56)}
\end{aligned}$$

Since the parameters θ_i^s are not coupled in any way, maximising $\ell(\boldsymbol{\theta})$ can be achieved by maximising each term $\ell_{is}(\theta_i^s)$ individually,

$$\ell_{is}(\theta_i^s) = n_{x_i=1}^s \log(\theta_i^s) + n_{x_i=0}^s \log(1-\theta_i^s). \tag{S.57}$$

Moreover, $\ell_{is}(\theta_i^s)$ corresponds to the log-likelihood for a Bernoulli model with success probability θ_i^s and data with $n_{x_i=1}^s$ number of ones and $n_{x_i=0}^s$ number of zeros.

(d) *Referring to the lecture material, conclude that the maximum likelihood estimates are given by*

$$\hat{\theta}_i^s = \frac{n^s_{x_i=1}}{n^s_{x_i=1} + n^s_{x_i=0}} = \frac{\sum_{j=1}^n \mathbb{1}(x_i^{(j)} = 1, \mathrm{pa}_i^{(j)} = s)}{\sum_{j=1}^n \mathbb{1}(\mathrm{pa}_i^{(j)} = s)} \tag{7}$$

Solution. Given the result from the previous question, we can optimise each term $\ell_{is}(\theta_i^s)$ separately. Furthermore, each term formally corresponds to a log-likelihood for a Bernoulli model, so that we can immediately use the results derived in the lecture, which gives

$$\hat{\theta}_i^s = \frac{n^s_{x_i=1}}{n^s_{x_i=1} + n^s_{x_i=0}} \tag{S.58}$$

Since $n^s_{x_i=1} = \sum_{j=1}^n \mathbb{1}(x_i^{(j)} = 1, \mathrm{pa}_i^{(j)} = s)$ and

$$n^s_{x_i=1} + n^s_{x_i=0} = \sum_{j=1}^n \mathbb{1}(x_i^{(j)} = 1, \mathrm{pa}_i^{(j)} = s) + \sum_{j=1}^n \mathbb{1}(x_i^{(j)} = 0, \mathrm{pa}_i^{(j)} = s) \tag{S.59}$$

$$= \sum_{j=1}^n \mathbb{1}(\mathrm{pa}_i^{(j)} = s), \tag{S.60}$$

which gives

$$\hat{\theta}_i^s = \frac{\sum_{j=1}^n \mathbb{1}(x_i^{(j)} = 1, \mathrm{pa}_i^{(j)} = s)}{\sum_{j=1}^n \mathbb{1}(\mathrm{pa}_i^{(j)} = s)}. \tag{S.61}$$

Hence, to determine $\hat{\theta}_i^s$, we first count the number of times the parents of x_i are in state s, which gives the denominator, and then among them, count the number of times $x_i = 1$, which gives the numerator.

Exercise 4. *Bayesian inference for the Bernoulli model*

Consider the Bayesian model

$$p(x|\theta) = \theta^x (1-\theta)^{1-x} \qquad\qquad p(\theta; \boldsymbol{\alpha}_0) = \mathcal{B}(\theta; \alpha_0, \beta_0)$$

where $x \in \{0, 1\}$, $\theta \in [0, 1]$, $\boldsymbol{\alpha}_0 = (\alpha_0, \beta_0)$, *and*

$$\mathcal{B}(\theta; \alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \qquad \theta \in [0, 1] \tag{8}$$

(a) *Given iid data* $\mathcal{D} = \{x_1, \ldots, x_n\}$ *show that the posterior of* θ *given* $\mathcal{D}$ *is*

$$p(\theta|\mathcal{D}) = \mathcal{B}(\theta; \alpha_n, \beta_n)$$
$$\alpha_n = \alpha_0 + n_{x=1} \qquad\qquad \beta_n = \beta_0 + n_{x=0}$$

where $n_{x=1}$ *denotes the number of ones and* $n_{x=0}$ *the number of zeros in the data.*

Solution. This follows immediately from

$$p(\theta|\mathcal{D}) \propto L(\theta) p(\theta; \boldsymbol{\alpha}_0) \tag{S.62}$$

and from the expression for the likelihood function of the Bernoulli model (see above or the lecture slides)

$$L(\theta) = \theta^{n_{x=1}}(1-\theta)^{n_{x=0}}. \tag{S.63}$$

Inserting all expressions into (S.62) gives

$$\begin{aligned} p(\theta|\mathcal{D}) &\propto \theta^{n_{x=1}}(1-\theta)^{n_{x=0}}\theta^{\alpha_0-1}(1-\theta)^{\beta_0-1} && \text{(S.64)}\\ &\propto \theta^{\alpha_0+n_{x=1}-1}(1-\theta)^{\beta_0+n_{x=0}-1} && \text{(S.65)}\\ &\propto \mathcal{B}(\theta, \alpha_0+n_{x=1}, \beta_0+n_{x=0}), && \text{(S.66)}\end{aligned}$$

which is the desired result. Since α_0 and β_0 are updated by the counts of ones and zeros in the data, these hyperparameters are also referred to as "pseudo-counts". Alternatively, one can think that they are the counts that are observed in another iid data set which has been previously analysed and used to determine the prior.

(b) *Compute the mean of a Beta random variable f,*

$$p(f;\alpha,\beta) = \mathcal{B}(f;\alpha,\beta) \qquad f\in[0,1], \tag{9}$$

using that

$$\int_0^1 f^{\alpha-1}(1-f)^{\beta-1}\mathrm{d}f = B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{10}$$

where $B(\alpha,\beta)$ denotes the Beta function and where the Gamma function $\Gamma(t)$ is defined as

$$\Gamma(t) = \int_o^\infty f^{t-1}\exp(-f)\mathrm{d}f \tag{11}$$

and satisfies $\Gamma(t+1) = t\Gamma(t)$.

Solution. We first write the partition function of $p(f;\alpha,\beta)$ in terms of the Beta function

$$\begin{aligned} Z(\alpha,\beta) &= \int_0^1 f^{\alpha-1}(1-f)^{\beta-1} && \text{(S.67)}\\ &= B(\alpha,\beta). && \text{(S.68)}\end{aligned}$$

We then have that the mean $\mathbb{E}[f]$ is given by

$$\begin{aligned} \mathbb{E}[f] &= \int_0^1 fp(f;\alpha,\beta)\mathrm{d}f && \text{(S.69)}\\ &= \frac{1}{B(\alpha,\beta)}\int_0^1 ff^{\alpha-1}(1-f)^{\beta-1} && \text{(S.70)}\\ &= \frac{1}{B(\alpha,\beta)}\int_0^1 f^{\alpha+1-1}(1-f)^{\beta-1} && \text{(S.71)}\\ &= \frac{B(\alpha+1,\beta)}{B(\alpha,\beta)} && \text{(S.72)}\\ &= \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+1+\beta)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} && \text{(S.73)}\\ &= \frac{\alpha\Gamma(\alpha)\Gamma(\beta)}{(\alpha+\beta)\Gamma(\alpha+\beta)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} && \text{(S.74)}\\ &= \frac{\alpha}{\alpha+\beta} && \text{(S.75)}\end{aligned}$$

where we have used the definition of the Beta function in terms of the Gamma function and the property $\Gamma(t+1) = t\Gamma(t)$.

(c) *Show that the predictive posterior probability $p(x = 1|\mathcal{D})$ for a new independently observed data point x equals the posterior mean of $p(\theta|\mathcal{D})$, which in turn is given by*

$$\mathbb{E}(\theta|\mathcal{D}) = \frac{\alpha_0 + n_{x=1}}{\alpha_0 + \beta_0 + n}. \tag{12}$$

Solution. We obtain

$$\begin{aligned}
p(x = 1|\mathcal{D}) &= \int_0^1 p(x = 1, \theta|\mathcal{D})\mathrm{d}\theta \qquad \text{(sum rule)} && \text{(S.76)}\\
&= \int_0^1 p(x = 1|\theta, \mathcal{D})p(\theta|\mathcal{D}) \qquad \text{(product rule)} && \text{(S.77)}\\
&= \int_0^1 p(x = 1|\theta)p(\theta|\mathcal{D}) \qquad (x \perp\!\!\!\perp \mathcal{D}|\theta) && \text{(S.78)}\\
&= \int_0^1 \theta p(\theta|\mathcal{D}) && \text{(S.79)}\\
&= \mathbb{E}[\theta|\mathcal{D}] && \text{(S.80)}
\end{aligned}$$

From the previous question we know the mean of a Beta random variable. Since $\theta \sim \mathcal{B}(\theta; \alpha_n, \beta_n)$, we obtain

$$\begin{aligned}
p(x = 1|\mathcal{D}) &= \mathbb{E}[\theta|\mathcal{D}] && \text{(S.81)}\\
&= \frac{\alpha_n}{\alpha_n + \beta_n} && \text{(S.82)}\\
&= \frac{\alpha_0 + n_{x=1}}{\alpha_0 + n_{x=1} + \beta_0 + n_{x=0}} && \text{(S.83)}\\
&= \frac{\alpha_0 + n_{x=1}}{\alpha_0 + \beta_0 + n} && \text{(S.84)}
\end{aligned}$$

where the last equation follows from the fact that $n = n_{x=0} + n_{x=1}$. Note that for $n \to \infty$, the posterior mean tends to the MLE $n_{x=1}/n$.

Exercise 5. *Bayesian inference of probability tables in fully observed directed graphical models of binary variables*

This is the Bayesian analogue of Exercise 3 and the notation follows that exercise. We consider the Bayesian model

$$p(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^{d} p(x_i|\mathrm{pa}_i, \boldsymbol{\theta}_i) \qquad x_i \in \{0, 1\} \tag{13}$$

$$p(\boldsymbol{\theta}; \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) = \prod_{i=1}^{d}\prod_{s=1}^{S_i} \mathcal{B}(\theta_i^s; \alpha_{i,0}^s, \beta_{i,0}^s) \tag{14}$$

where $p(x_i|\mathrm{pa}_i, \boldsymbol{\theta}_i)$ is defined via (5), $\boldsymbol{\alpha}_0$ is a vector of hyperparameters containing all $\alpha_{i,0}^s$, $\boldsymbol{\beta}_0$ the vector containing all $\beta_{i,0}^s$, and as before $\mathcal{B}$ denotes the Beta distribution. Under the prior, all parameters are independent.

For iid data $\mathcal{D} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}$ show that

$$p(\boldsymbol{\theta}|\mathcal{D}) = \prod_{i=1}^{d}\prod_{s=1}^{S_i} \mathcal{B}(\theta_i^s, \alpha_{i,n}^s, \beta_{i,n}^s) \tag{15}$$

where

$$\alpha_{i,n}^s = \alpha_{i,0}^s + n_{x_i=1}^s \qquad\qquad \beta_{i,n}^s = \beta_{i,0}^s + n_{x_i=0}^s \tag{16}$$

and that the parameters are also independent under the posterior.

Solution. We start with

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}; \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0). \tag{S.85}$$

Inserting the expression for $p(\mathcal{D}|\boldsymbol{\theta})$ given in (6) and the assumed form of the prior gives

$$\begin{aligned}
p(\boldsymbol{\theta}|\mathcal{D}) &\propto \prod_{i=1}^{d}\prod_{s=1}^{S_i} (\theta_i^s)^{n_{x_i=1}^s}(1-\theta_i^s)^{n_{x_i=0}^s} \prod_{i=1}^{d}\prod_{s=1}^{S_i} \mathcal{B}(\theta_i^s; \alpha_{i,0}^s, \beta_{i,0}^s) && \text{(S.86)}\\
&\propto \prod_{i=1}^{d}\prod_{s=1}^{S_i} (\theta_i^s)^{n_{x_i=1}^s}(1-\theta_i^s)^{n_{x_i=0}^s} \mathcal{B}(\theta_i^s; \alpha_{i,0}^s, \beta_{i,0}^s) && \text{(S.87)}\\
&\propto \prod_{i=1}^{d}\prod_{s=1}^{S_i} (\theta_i^s)^{n_{x_i=1}^s}(1-\theta_i^s)^{n_{x_i=0}^s} (\theta_i^s)^{\alpha_{i,0}^s-1}(1-\theta_i^s)^{\beta_{i,0}^s-1} && \text{(S.88)}\\
&\propto \prod_{i=1}^{d}\prod_{s=1}^{S_i} (\theta_i^s)^{\alpha_{i,0}^s+n_{x_i=1}^s-1}(1-\theta_i^s)^{\beta_{i,0}^s+n_{x_i=0}^s-1} && \text{(S.89)}\\
&\propto \prod_{i=1}^{d}\prod_{s=1}^{S_i} \mathcal{B}(\theta_i^s; \alpha_{i,0}^s + n_{x_i=1}^s, \beta_{i,0}^s + n_{x_i=0}^s) && \text{(S.90)}
\end{aligned}$$

It can be immediately verified that $\mathcal{B}(\theta_i^s; \alpha_{i,0}^s + n_{x_i=1}^s, \beta_{i,0}^s + n_{x_i=0}^s)$ is proportional to the marginal $p(\theta_i^s|\mathcal{D})$ so that the parameters are independent under the posterior too.