Exercise 1. ***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 $x_1, \ldots, x_d$,

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

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

$$p(x_i|\text{pa}_i; \boldsymbol{\theta}_i) = \prod_{s=1}^{S_i} (\theta_i^s)^{\mathbb{1}(x_i=1,\text{pa}_i=s)}(1-\theta_i^s)^{\mathbb{1}(x_i=0,\text{pa}_i=s)} \tag{2}$$

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 $k \in \{1, \ldots, S_i\}$ of the parents of x_i. Since $\mathbb{1}(x_i = 1, \text{pa}_i = s) = 0$ unless $s = k$, we have indeed that

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

$$= 1 \cdot (\theta_i^k)^{\mathbb{1}(x_i=1,\text{pa}_i=k)}(1-\theta_i^k)^0 \tag{S.2}$$

$$= \theta_i^k. \tag{S.3}$$

(b) *For iid data* $\mathcal{D} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}$ *show that the likelihood can be represented as*

$$p(\mathcal{D}; \boldsymbol{\theta}) = \prod_{i=1}^{d} \prod_{s=1}^{S_i} (\theta_i^s)^{n^s_{x_i=1}}(1-\theta_i^s)^{n^s_{x_i=0}} \tag{3}$$

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

$$p(\mathcal{D};\boldsymbol{\theta}) = \prod_{j=1}^{n} p(\mathbf{x}^{(j)};\boldsymbol{\theta}) \tag{S.4}$$

$$\tag{S.5}$$

where each term $p(\mathbf{x}^{(j)};\boldsymbol{\theta})$ factorises as in (1),

$$p(\mathbf{x}^{(j)};\boldsymbol{\theta}) = \prod_{i=1}^{d} p(x_i^{(j)}|\text{pa}_i^{(j)};\boldsymbol{\theta}_i) \tag{S.6}$$

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 (2),

$$p(x_i^{(j)}|\text{pa}_i^{(j)};\boldsymbol{\theta}_i) = \prod_{s=1}^{S_i} (\theta_i^s)^{\mathbb{1}(x_i^{(j)}=1,\text{pa}_i^{(j)}=s)}(1-\theta_i^s)^{\mathbb{1}(x_i^{(j)}=0,\text{pa}_i^{(j)}=s)}, \tag{S.7}$$

so that

$$\begin{aligned}
p(\mathcal{D};\boldsymbol{\theta}) &= \prod_{j=1}^{n}\prod_{i=1}^{d} p(x_i^{(j)}|\text{pa}_i^{(j)};\boldsymbol{\theta}_i) && \text{(S.8)}\\
&= \prod_{j=1}^{n}\prod_{i=1}^{d}\prod_{s=1}^{S_i} (\theta_i^s)^{\mathbb{1}(x_i^{(j)}=1,\text{pa}_i^{(j)}=s)}(1-\theta_i^s)^{\mathbb{1}(x_i^{(j)}=0,\text{pa}_i^{(j)}=s)} && \text{(S.9)}
\end{aligned}$$

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

$$p(\mathcal{D};\boldsymbol{\theta}) = \prod_{i=1}^{d}\prod_{s=1}^{S_i}\prod_{j=1}^{n} (\theta_i^s)^{\mathbb{1}(x_i^{(j)}=1,\text{pa}_i^{(j)}=s)}(1-\theta_i^s)^{\mathbb{1}(x_i^{(j)}=0,\text{pa}_i^{(j)}=s)} \tag{S.10}$$

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 $x_i^{(j)} = 0$

$$\begin{aligned}
p(\mathcal{D};\boldsymbol{\theta}) &= \prod_{i=1}^{d}\prod_{s=1}^{S_i}\prod_{\substack{j:\\ x_i^{(j)}=1}}\prod_{\substack{j:\\ x_i^{(j)}=0}} (\theta_i^s)^{\mathbb{1}(x_i^{(j)}=1,\text{pa}_i^{(j)}=s)}(1-\theta_i^s)^{\mathbb{1}(x_i^{(j)}=0,\text{pa}_i^{(j)}=s)} && \text{(S.11)}\\
&= \prod_{i=1}^{d}\prod_{s=1}^{S_i}\prod_{\substack{j:\\ x_i^{(j)}=1}} (\theta_i^s)^{\mathbb{1}(x_i^{(j)}=1,\text{pa}_i^{(j)}=s)} \prod_{\substack{j:\\ x_i^{(j)}=0}} (1-\theta_i^s)^{\mathbb{1}(x_i^{(j)}=0,\text{pa}_i^{(j)}=s)} && \text{(S.12)}\\
&= \prod_{i=1}^{d}\prod_{s=1}^{S_i} (\theta_i^s)^{\sum_{j=1}^{n}\mathbb{1}(x_i^{(j)}=1,\text{pa}_i^{(j)}=s)}(1-\theta_i^s)^{\sum_{j=1}^{n}\mathbb{1}(x_i^{(j)}=0,\text{pa}_i^{(j)}=s)} && \text{(S.13)}\\
&= \prod_{i=1}^{d}\prod_{s=1}^{S_i} (\theta_i^s)^{n^s_{x_i=1}}(1-\theta_i^s)^{n^s_{x_i=0}} && \text{(S.14)}
\end{aligned}$$

where

$$n^s_{x_i=1} = \sum_{j=1}^{n}\mathbb{1}(x_i^{(j)}=1,\text{pa}_i^{(j)}=s) \qquad n^s_{x_i=0} = \sum_{j=1}^{n}\mathbb{1}(x_i^{(j)}=0,\text{pa}_i^{(j)}=s) \tag{S.15}$$

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

$$\begin{aligned}
\ell(\boldsymbol{\theta}) &= \log p(\mathcal{D}; \boldsymbol{\theta}) && \text{(S.16)}\\
&= \log \prod_{i=1}^{d} \prod_{s=1}^{S_i} (\theta_i^s)^{n^s_{x_i=1}} (1-\theta_i^s)^{n^s_{x_i=0}} && \text{(S.17)}\\
&= \sum_{i=1}^{d} \sum_{s=1}^{S_i} \log \left[(\theta_i^s)^{n^s_{x_i=1}} (1-\theta_i^s)^{n^s_{x_i=0}} \right] && \text{(S.18)}\\
&= \sum_{i=1}^{d} \sum_{s=1}^{S_i} n^s_{x_i=1} \log(\theta_i^s) + n^s_{x_i=0} \log(1-\theta_i^s) && \text{(S.19)}
\end{aligned}$$

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,

$$\ell_{is}(\theta_i^s) = n^s_{x_i=1} \log(\theta_i^s) + n^s_{x_i=0} \log(1-\theta_i^s). \tag{S.20}$$

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

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

$$\hat{\theta}_i^s = \frac{n^s_{x_i=1}}{n^s_{x_i=1} + n^s_{x_i=0}} = \frac{\sum_{j=1}^n \mathbb{1}(x_i^{(j)} = 1, \mathrm{pa}_i^{(j)} = s)}{\sum_{j=1}^n \mathbb{1}(\mathrm{pa}_i^{(j)} = s)} \tag{4}$$

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

$$\hat{\theta}_i^s = \frac{n^s_{x_i=1}}{n^s_{x_i=1} + n^s_{x_i=0}} \tag{S.21}$$

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

$$\begin{aligned}
n^s_{x_i=1} + n^s_{x_i=0} &= \sum_{j=1}^{n} \mathbb{1}(x_i^{(j)} = 1, \mathrm{pa}_i^{(j)} = s) + \sum_{j=1}^{n} \mathbb{1}(x_i^{(j)} = 0, \mathrm{pa}_i^{(j)} = s) && \text{(S.22)}\\
&= \sum_{j=1}^{n} \mathbb{1}(\mathrm{pa}_i^{(j)} = s), && \text{(S.23)}
\end{aligned}$$

which gives

$$\hat{\theta}_i^s = \frac{\sum_{j=1}^n \mathbb{1}(x_i^{(j)} = 1, \mathrm{pa}_i^{(j)} = s)}{\sum_{j=1}^n \mathbb{1}(\mathrm{pa}_i^{(j)} = s)}. \tag{S.24}$$

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 2. ***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*

$$\mathcal{B}(\theta;\alpha,\beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \qquad \theta \in [0,1] \tag{5}$$

(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

$$p(\theta|\mathcal{D}) \propto L(\theta)p(\theta;\boldsymbol{\alpha}_0) \tag{S.25}$$

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

$$L(\theta) = \theta^{n_{x=1}}(1-\theta)^{n_{x=0}}. \tag{S.26}$$

Inserting all expressions into (S.25) gives

$$p(\theta|\mathcal{D}) \propto \theta^{n_{x=1}}(1-\theta)^{n_{x=0}}\theta^{\alpha_0-1}(1-\theta)^{\beta_0-1} \tag{S.27}$$
$$\propto \theta^{\alpha_0+n_{x=1}-1}(1-\theta)^{\beta_0+n_{x=0}-1} \tag{S.28}$$
$$\propto \mathcal{B}(\theta, \alpha_0+n_{x=1}, \beta_0+n_{x=0}), \tag{S.29}$$

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,

$$p(f;\alpha,\beta) = \mathcal{B}(f;\alpha,\beta) \qquad f \in [0,1], \tag{6}$$

using that

$$\int_0^1 f^{\alpha-1}(1-f)^{\beta-1}\mathrm{d}f = B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{7}$$

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

$$\Gamma(t) = \int_o^\infty f^{t-1}\exp(-f)\mathrm{d}f \tag{8}$$

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

Hint: It will be useful to represent the partition function in terms of the Beta function.

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

$$Z(\alpha,\beta) = \int_0^1 f^{\alpha-1}(1-f)^{\beta-1} \tag{S.30}$$
$$= B(\alpha,\beta). \tag{S.31}$$

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

$$\mathbb{E}[f] = \int_0^1 f p(f;\alpha,\beta)\mathrm{d}f \tag{S.32}$$
$$= \frac{1}{B(\alpha,\beta)}\int_0^1 f f^{\alpha-1}(1-f)^{\beta-1}\mathrm{d}f \tag{S.33}$$
$$= \frac{1}{B(\alpha,\beta)}\int_0^1 f^{\alpha+1-1}(1-f)^{\beta-1}\mathrm{d}f \tag{S.34}$$
$$= \frac{B(\alpha+1,\beta)}{B(\alpha,\beta)} \tag{S.35}$$
$$= \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+1+\beta)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \tag{S.36}$$
$$= \frac{\alpha\Gamma(\alpha)\Gamma(\beta)}{(\alpha+\beta)\Gamma(\alpha+\beta)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \tag{S.37}$$
$$= \frac{\alpha}{\alpha+\beta} \tag{S.38}$$

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*

$$\mathbb{E}(\theta|\mathcal{D}) = \frac{\alpha_0 + n_{x=1}}{\alpha_0+\beta_0+n}. \tag{9}$$

Solution. We obtain

$$p(x=1|\mathcal{D}) = \int_0^1 p(x=1,\theta|\mathcal{D})\mathrm{d}\theta \qquad \text{(sum rule)} \tag{S.39}$$
$$= \int_0^1 p(x=1|\theta,\mathcal{D})p(\theta|\mathcal{D})\mathrm{d}\theta \qquad \text{(product rule)} \tag{S.40}$$
$$= \int_0^1 p(x=1|\theta)p(\theta|\mathcal{D})\mathrm{d}\theta \qquad (x \perp\!\!\!\perp \mathcal{D}|\theta) \tag{S.41}$$
$$= \int_0^1 \theta p(\theta|\mathcal{D})\mathrm{d}\theta \tag{S.42}$$
$$= \mathbb{E}[\theta|\mathcal{D}] \tag{S.43}$$

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

$$p(x = 1|\mathcal{D}) = \mathbb{E}[\theta|\mathcal{D}] \tag{S.44}$$
$$= \frac{\alpha_n}{\alpha_n + \beta_n} \tag{S.45}$$
$$= \frac{\alpha_0 + n_{x=1}}{\alpha_0 + n_{x=1} + \beta_0 + n_{x=0}} \tag{S.46}$$
$$= \frac{\alpha_0 + n_{x=1}}{\alpha_0 + \beta_0 + n} \tag{S.47}$$

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

This is the Bayesian analogue of Exercise 1 and the notation follows that exercise. We consider the Bayesian model

$$p(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^{d} p(x_i|\mathrm{pa}_i, \boldsymbol{\theta}_i) \qquad x_i \in \{0, 1\} \tag{10}$$
$$p(\boldsymbol{\theta}; \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) = \prod_{i=1}^{d} \prod_{s=1}^{S_i} \mathcal{B}(\theta_i^s; \alpha_{i,0}^s, \beta_{i,0}^s) \tag{11}$$

where $p(x_i|\mathrm{pa}_i, \boldsymbol{\theta}_i)$ is defined via (2), $\boldsymbol{\alpha}_0$ 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

$$p(\boldsymbol{\theta}|\mathcal{D}) = \prod_{i=1}^{d} \prod_{s=1}^{S_i} \mathcal{B}(\theta_i^s, \alpha_{i,n}^s, \beta_{i,n}^s) \tag{12}$$

where

$$\alpha_{i,n}^s = \alpha_{i,0}^s + n_{x_i=1}^s \qquad \beta_{i,n}^s = \beta_{i,0}^s + n_{x_i=0}^s \tag{13}$$

and that the parameters are also independent under the posterior.

Solution. We start with

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta}; \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0). \tag{S.48}$$

Inserting the expression for $p(\mathcal{D}|\boldsymbol{\theta})$ given in (3) and the assumed form of the prior gives

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto \prod_{i=1}^{d}\prod_{s=1}^{S_i}(\theta_i^s)^{n^s_{x_i=1}}(1-\theta_i^s)^{n^s_{x_i=0}}\prod_{i=1}^{d}\prod_{s=1}^{S_i}\mathcal{B}(\theta_i^s;\alpha^s_{i,0},\beta^s_{i,0}) \tag{S.49}$$

$$\propto \prod_{i=1}^{d}\prod_{s=1}^{S_i}(\theta_i^s)^{n^s_{x_i=1}}(1-\theta_i^s)^{n^s_{x_i=0}}\mathcal{B}(\theta_i^s;\alpha^s_{i,0},\beta^s_{i,0}) \tag{S.50}$$

$$\propto \prod_{i=1}^{d}\prod_{s=1}^{S_i}(\theta_i^s)^{n^s_{x_i=1}}(1-\theta_i^s)^{n^s_{x_i=0}}(\theta_i^s)^{\alpha^s_{i,0}-1}(1-\theta_i^s)^{\beta^s_{i,0}-1} \tag{S.51}$$

$$\propto \prod_{i=1}^{d}\prod_{s=1}^{S_i}(\theta_i^s)^{\alpha^s_{i,0}+n^s_{x_i=1}-1}(1-\theta_i^s)^{\beta^s_{i,0}+n^s_{x_i=0}-1} \tag{S.52}$$

$$\propto \prod_{i=1}^{d}\prod_{s=1}^{S_i}\mathcal{B}(\theta_i^s;\alpha^s_{i,0}+n^s_{x_i=1},\beta^s_{i,0}+n^s_{x_i=0}) \tag{S.53}$$

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