The purpose of this tutorial sheet is to help you better understand the lecture material. Start early and do as many as you have time for. Even if you are unable to make much progress, you should still attend your tutorial.

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.

(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 $n^s_{x_i=0}$.

(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.

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

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

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.

(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.

(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

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

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.