These are exercises for self-study and exam preparation. All material is examinable unless otherwise mentioned.

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. Since $\sigma > 0$, the function $v = g(\sigma) = \sigma^2$ is invertible, and

the value of v that maximises the likelihood uniquely defines the value of σ that maximises the likelihood, namely

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

This reparameterisation approach holds more generally: if we can express the parameters $\boldsymbol{\theta}$ as $\boldsymbol{\eta} = g(\boldsymbol{\theta})$ where g is invertible, we can maximise $J(\boldsymbol{\eta}) = \ell(g^{-1}(\boldsymbol{\eta}))$ instead of $\ell(\boldsymbol{\theta})$. The optimal value $\hat{\boldsymbol{\eta}} = \operatorname{argmax}_{\boldsymbol{\eta}} J(\boldsymbol{\eta})$ defines the optimal value of $\boldsymbol{\theta}$: $\hat{\boldsymbol{\theta}} = g^{-1}(\hat{\boldsymbol{\eta}})$. The maximum likelihood estimate is said to be invariant to reparameterisation.

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

$$\frac{\partial J}{\partial \mu} = \frac{1}{v}\sum_{i=1}^{n}(x_i - \mu) \tag{S.6}$$

$$= \frac{1}{v}\sum_{i=1}^{n} x_i - \frac{n}{v}\mu \tag{S.7}$$

$$\frac{\partial J}{\partial v} = -\frac{n}{2}\frac{1}{v} + \frac{1}{2v^2}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{S.8}$$

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

$$\frac{1}{v}\sum_{i=1}^{n}(x_i - \mu) = 0 \tag{S.9}$$

$$-\frac{n}{2}\frac{1}{v} + \frac{1}{2v^2}\sum_{i=1}^{n}(x_i - \mu)^2 = 0 \tag{S.10}$$

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

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

$$\frac{1}{2}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 = \frac{n}{2}v, \tag{S.13}$$

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.
Hint: You may use that

$$\mathcal{N}(x; m_1, \sigma_1^2)\mathcal{N}(x; m_2, \sigma_2^2) \propto \mathcal{N}(x; m_3, \sigma_3^2) \tag{4}$$

where

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{5}$$

$$\sigma_3^2 = \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)^{-1} = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \tag{6}$$

$$m_3 = \sigma_3^2\left(\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}\right) = m_1 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}(m_2 - m_1) \tag{7}$$

Solution. We re-use the expression for the likelihood $L(\mu)$ from Exercise 1.

$$L(\mu) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right], \tag{S.18}$$

which we can write as

$$L(\mu) \propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right] \tag{S.19}$$

$$\propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i^2-2\mu x_i+\mu^2)\right] \tag{S.20}$$

$$\propto \exp\left[-\frac{1}{2\sigma^2}\left(-2\mu\sum_{i=1}^{n}x_i+n\mu^2\right)\right] \tag{S.21}$$

$$\propto \exp\left[-\frac{1}{2\sigma^2}\left(-2n\mu\bar{x}+n\mu^2\right)\right] \tag{S.22}$$

$$\propto \exp\left[-\frac{n}{2\sigma^2}(\mu-\bar{x})^2\right] \tag{S.23}$$

$$\propto \mathcal{N}(\mu;\bar{x},\sigma^2/n). \tag{S.24}$$

The posterior is

$$p(\mu|\mathcal{D}) \propto L(\theta)p(\mu;\mu_0,\sigma_0^2) \tag{S.25}$$

$$\propto \mathcal{N}(\mu;\bar{x},\sigma^2/n)\mathcal{N}(\mu;\mu_0,\sigma_0^2) \tag{S.26}$$

so that with (4), we have

$$p(\mu|\mathcal{D}) \propto \mathcal{N}(\mu;\mu_n,\sigma_n^2) \tag{S.27}$$

$$\sigma_n^2 = \left(\frac{1}{\sigma^2/n}+\frac{1}{\sigma_0^2}\right)^{-1} \tag{S.28}$$

$$= \frac{\sigma_0^2\sigma^2/n}{\sigma_0^2+\sigma^2/n} \tag{S.29}$$

$$\mu_n = \sigma_n^2\left(\frac{\bar{x}}{\sigma^2/n}+\frac{\mu_0}{\sigma_0^2}\right) \tag{S.30}$$

$$= \frac{1}{\sigma_0^2+\sigma^2/n}\left(\sigma_0^2\bar{x}+(\sigma^2/n)\mu_0\right) \tag{S.31}$$

$$= \frac{\sigma_0^2}{\sigma_0^2+\sigma^2/n}\bar{x}+\frac{\sigma^2/n}{\sigma_0^2+\sigma^2/n}\mu_0. \tag{S.32}$$

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 (7), we also have that

$$\mu_n = \mu_0 + \frac{\sigma_0^2}{\sigma^2/n+\sigma_0^2}(\bar{x}-\mu_0), \tag{S.33}$$

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 $x_1, \ldots, x_d$,

$$p(\mathbf{x}; \boldsymbol{\theta}) = \prod_{i=1}^{d} p(x_i | \mathrm{pa}_i; \boldsymbol{\theta}_i) \qquad x_i \in \{0, 1\} \tag{8}$$

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*

$p(x_3 = 1 \mid x_1, x_2; \theta_3^1, \ldots, \theta_3^4))$	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* $p(x_i|\mathrm{pa}_i; \boldsymbol{\theta}_i)$ *is equivalent to writing* $p(x_i|\mathrm{pa}_i; \boldsymbol{\theta}_i)$ *as*

$$p(x_i|\mathrm{pa}_i; \boldsymbol{\theta}_i) = \prod_{s=1}^{S_i} (\theta_i^s)^{\mathbb{1}(x_i=1, \mathrm{pa}_i = s)} (1 - \theta_i^s)^{\mathbb{1}(x_i=0, \mathrm{pa}_i = s)} \tag{9}$$

where $S_i = 2^{m_i}$ *is the total number of states/configurations that the parents can be in, and* $\mathbb{1}(x_i = 1, \mathrm{pa}_i = s)$ *is one if* $x_i = 1$ *and* $\mathrm{pa}_i = s$*, 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 $p(x_i = 1|\mathrm{pa}_i = k; \boldsymbol{\theta}_i) = \theta_i^k$ for any state $k \in \{1, \ldots, S_i\}$ of the parents of x_i. Since $\mathbb{1}(x_i = 1, \mathrm{pa}_i = s) = 0$ unless $s = k$, we have indeed that

$$p(x_i = 1|\mathrm{pa}_i = k; \boldsymbol{\theta}_i) = \left[\prod_{s \neq k} (\theta_i^s)^0 (1 - \theta_i^s)^0 \right] (\theta_i^k)^{\mathbb{1}(x_i=1, \mathrm{pa}_i = k)} (1 - \theta_i^k)^{\mathbb{1}(x_i=0, \mathrm{pa}_i = k)} \tag{S.34}$$

$$= 1 \cdot (\theta_i^k)^{\mathbb{1}(x_i=1, \mathrm{pa}_i = k)} (1 - \theta_i^k)^0 \tag{S.35}$$

$$= \theta_i^k. \tag{S.36}$$

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

$$p(\mathcal{D}; \boldsymbol{\theta}) = \prod_{i=1}^{d} \prod_{s=1}^{S_i} (\theta_i^s)^{n^s_{x_i=1}} (1 - \theta_i^s)^{n^s_{x_i=0}} \tag{10}$$

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

Solution. Since the data are iid, we have

$$p(\mathcal{D}; \boldsymbol{\theta}) = \prod_{j=1}^{n} p(\mathbf{x}^{(j)}; \boldsymbol{\theta}) \tag{S.37}$$

$$\tag{S.38}$$

where each term $p(\mathbf{x}^{(j)};\boldsymbol{\theta})$ factorises as in (8),

$$p(\mathbf{x}^{(j)};\boldsymbol{\theta}) = \prod_{i=1}^{d} p(x_i^{(j)}|\mathrm{pa}_i^{(j)};\boldsymbol{\theta}_i) \tag{S.39}$$

with $x_i^{(j)}$ denoting the i-th element of $\mathbf{x}^{(j)}$ and $\mathrm{pa}_i^{(j)}$ the corresponding parents. The conditionals $p(x_i^{(j)}|\mathrm{pa}_i^{(j)};\boldsymbol{\theta}_i)$ factorise further according to (9),

$$p(x_i^{(j)}|\mathrm{pa}_i^{(j)};\boldsymbol{\theta}_i) = \prod_{s=1}^{S_i} (\theta_i^s)^{\mathbb{1}(x_i^{(j)}=1,\mathrm{pa}_i^{(j)}=s)} (1-\theta_i^s)^{\mathbb{1}(x_i^{(j)}=0,\mathrm{pa}_i^{(j)}=s)}, \tag{S.40}$$

so that

$$p(\mathcal{D};\boldsymbol{\theta}) = \prod_{j=1}^{n}\prod_{i=1}^{d} p(x_i^{(j)}|\mathrm{pa}_i^{(j)};\boldsymbol{\theta}_i) \tag{S.41}$$

$$= \prod_{j=1}^{n}\prod_{i=1}^{d}\prod_{s=1}^{S_i} (\theta_i^s)^{\mathbb{1}(x_i^{(j)}=1,\mathrm{pa}_i^{(j)}=s)} (1-\theta_i^s)^{\mathbb{1}(x_i^{(j)}=0,\mathrm{pa}_i^{(j)}=s)} \tag{S.42}$$

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

$$p(\mathcal{D};\boldsymbol{\theta}) = \prod_{i=1}^{d}\prod_{s=1}^{S_i}\prod_{j=1}^{n} (\theta_i^s)^{\mathbb{1}(x_i^{(j)}=1,\mathrm{pa}_i^{(j)}=s)} (1-\theta_i^s)^{\mathbb{1}(x_i^{(j)}=0,\mathrm{pa}_i^{(j)}=s)} \tag{S.43}$$

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 $x_i^{(j)}=0$

$$p(\mathcal{D};\boldsymbol{\theta}) = \prod_{i=1}^{d}\prod_{s=1}^{S_i} \prod_{\substack{j:\\ x_i^{(j)}=1}} \prod_{\substack{j:\\ x_i^{(j)}=0}} (\theta_i^s)^{\mathbb{1}(x_i^{(j)}=1,\mathrm{pa}_i^{(j)}=s)} (1-\theta_i^s)^{\mathbb{1}(x_i^{(j)}=0,\mathrm{pa}_i^{(j)}=s)} \tag{S.44}$$

$$= \prod_{i=1}^{d}\prod_{s=1}^{S_i} \prod_{\substack{j:\\ x_i^{(j)}=1}} (\theta_i^s)^{\mathbb{1}(x_i^{(j)}=1,\mathrm{pa}_i^{(j)}=s)} \prod_{\substack{j:\\ x_i^{(j)}=0}} (1-\theta_i^s)^{\mathbb{1}(x_i^{(j)}=0,\mathrm{pa}_i^{(j)}=s)} \tag{S.45}$$

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

$$= \prod_{i=1}^{d}\prod_{s=1}^{S_i} (\theta_i^s)^{n_{x_i=1}^s} (1-\theta_i^s)^{n_{x_i=0}^s} \tag{S.47}$$

where

$$n_{x_i=1}^s = \sum_{j=1}^{n} \mathbb{1}(x_i^{(j)}=1, \mathrm{pa}_i^{(j)}=s) \qquad n_{x_i=0}^s = \sum_{j=1}^{n} \mathbb{1}(x_i^{(j)}=0, \mathrm{pa}_i^{(j)}=s) \tag{S.48}$$

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

$$\ell(\boldsymbol{\theta}) = \log p(\mathcal{D}; \boldsymbol{\theta}) \tag{S.49}$$

$$= \log \prod_{i=1}^{d} \prod_{s=1}^{S_i} (\theta_i^s)^{n^s_{x_i=1}} (1-\theta_i^s)^{n^s_{x_i=0}} \tag{S.50}$$

$$= \sum_{i=1}^{d} \sum_{s=1}^{S_i} \log \left[(\theta_i^s)^{n^s_{x_i=1}} (1-\theta_i^s)^{n^s_{x_i=0}} \right] \tag{S.51}$$

$$= \sum_{i=1}^{d} \sum_{s=1}^{S_i} n^s_{x_i=1} \log(\theta_i^s) + n^s_{x_i=0} \log(1-\theta_i^s) \tag{S.52}$$

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,

$$\ell_{is}(\theta_i^s) = n^s_{x_i=1} \log(\theta_i^s) + n^s_{x_i=0} \log(1-\theta_i^s). \tag{S.53}$$

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

$$\hat{\theta}_i^s = \frac{n^s_{x_i=1}}{n^s_{x_i=1} + n^s_{x_i=0}} = \frac{\sum_{j=1}^{n} \mathbb{1}(x_i^{(j)} = 1, \mathrm{pa}_i^{(j)} = s)}{\sum_{j=1}^{n} \mathbb{1}(\mathrm{pa}_i^{(j)} = s)} \tag{11}$$

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

$$\hat{\theta}_i^s = \frac{n^s_{x_i=1}}{n^s_{x_i=1} + n^s_{x_i=0}} \tag{S.54}$$

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

$$n^s_{x_i=1} + n^s_{x_i=0} = \sum_{j=1}^{n} \mathbb{1}(x_i^{(j)} = 1, \mathrm{pa}_i^{(j)} = s) + \sum_{j=1}^{n} \mathbb{1}(x_i^{(j)} = 0, \mathrm{pa}_i^{(j)} = s) \tag{S.55}$$

$$= \sum_{j=1}^{n} \mathbb{1}(\mathrm{pa}_i^{(j)} = s), \tag{S.56}$$

which gives

$$\hat{\theta}_i^s = \frac{\sum_{j=1}^{n} \mathbb{1}(x_i^{(j)} = 1, \mathrm{pa}_i^{(j)} = s)}{\sum_{j=1}^{n} \mathbb{1}(\mathrm{pa}_i^{(j)} = s)}. \tag{S.57}$$

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

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

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

$$\mathcal{B}(\theta;\alpha,\beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \qquad \theta \in [0,1] \tag{12}$$

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

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

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

Solution. This follows from

$$p(\theta|\mathcal{D}) \propto L(\theta)p(\theta;\boldsymbol{\alpha}_0) \tag{S.58}$$

and from the expression for the likelihood function of the Bernoulli model, which is

$$L(\theta) = \prod_{i=1}^n p(x_i|\theta) \tag{S.59}$$
$$= \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} \tag{S.60}$$
$$= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{\sum_{i=1}^n (1-x_i)} \tag{S.61}$$
$$= \theta^{n_{x=1}}(1-\theta)^{n_{x=0}}, \tag{S.62}$$

where $n_{x=1} = \sum_{i=1}^n x_i$ denotes the number of 1's in the data, and $n_{x=0} = \sum_{i=1}^n (1-x_i) = n - n_{x=1}$ the number of 0's.

Inserting the expressions for the likelihood and prior into (S.58) gives

$$p(\theta|\mathcal{D}) \propto \theta^{n_{x=1}}(1-\theta)^{n_{x=0}}\theta^{\alpha_0-1}(1-\theta)^{\beta_0-1} \tag{S.63}$$
$$\propto \theta^{\alpha_0+n_{x=1}-1}(1-\theta)^{\beta_0+n_{x=0}-1} \tag{S.64}$$
$$\propto \mathcal{B}(\theta,\alpha_0+n_{x=1},\beta_0+n_{x=0}), \tag{S.65}$$

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) Show that the mean of a Beta random variable $f \sim \mathcal{B}(f;\alpha,\beta)$ *is*

$$\mathbb{E}[f] = \frac{\alpha}{\alpha+\beta}. \tag{13}$$

You may use that

$$\int_0^1 f^{\alpha-1}(1-f)^{\beta-1}\mathrm{d}f = B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{14}$$

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

$$\Gamma(t) = \int_o^\infty f^{t-1}\exp(-f)\mathrm{d}f \tag{15}$$

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

Hint: Represent the partition function of the Beta distribution in terms of the Beta function.

Solution. We first write the partition function of the Beta distribution in terms of the Beta function

$$Z(\alpha, \beta) = \int_0^1 f^{\alpha-1}(1-f)^{\beta-1} \tag{S.66}$$

$$= B(\alpha, \beta). \tag{S.67}$$

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

$$\mathbb{E}[f] = \int_0^1 f p(f; \alpha, \beta) \mathrm{d}f \tag{S.68}$$

$$= \frac{1}{B(\alpha, \beta)} \int_0^1 f f^{\alpha-1}(1-f)^{\beta-1} \mathrm{d}f \tag{S.69}$$

$$= \frac{1}{B(\alpha, \beta)} \int_0^1 f^{\alpha+1-1}(1-f)^{\beta-1} \mathrm{d}f \tag{S.70}$$

$$= \frac{B(\alpha+1, \beta)}{B(\alpha, \beta)} \tag{S.71}$$

$$= \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+1+\beta)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \tag{S.72}$$

$$= \frac{\alpha\Gamma(\alpha)\Gamma(\beta)}{(\alpha+\beta)\Gamma(\alpha+\beta)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \tag{S.73}$$

$$= \frac{\alpha}{\alpha+\beta} \tag{S.74}$$

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

$$\mathbb{E}(\theta|\mathcal{D}) = \frac{\alpha_0 + n_{x=1}}{\alpha_0 + \beta_0 + n}. \tag{16}$$

Solution. We obtain

$$p(x = 1|\mathcal{D}) = \int_0^1 p(x = 1, \theta|\mathcal{D}) \mathrm{d}\theta \qquad \text{(sum rule)} \tag{S.75}$$

$$= \int_0^1 p(x = 1|\theta, \mathcal{D}) p(\theta|\mathcal{D}) \mathrm{d}\theta \qquad \text{(product rule)} \tag{S.76}$$

$$= \int_0^1 p(x = 1|\theta) p(\theta|\mathcal{D}) \mathrm{d}\theta \qquad (x \perp\!\!\!\perp \mathcal{D}|\theta) \tag{S.77}$$

$$= \int_0^1 \theta p(\theta|\mathcal{D}) \mathrm{d}\theta \qquad \text{(Bernoulli model)} \tag{S.78}$$

$$= \mathbb{E}[\theta|\mathcal{D}] \tag{S.79}$$

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

$$p(x = 1|\mathcal{D}) = \mathbb{E}[\theta|\mathcal{D}] \tag{S.80}$$

$$= \frac{\alpha_n}{\alpha_n + \beta_n} \tag{S.81}$$

$$= \frac{\alpha_0 + n_{x=1}}{\alpha_0 + n_{x=1} + \beta_0 + n_{x=0}} \tag{S.82}$$

$$= \frac{\alpha_0 + n_{x=1}}{\alpha_0 + \beta_0 + n} \tag{S.83}$$

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

$$p(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^{d} p(x_i|\mathrm{pa}_i, \boldsymbol{\theta}_i) \qquad x_i \in \{0, 1\} \tag{17}$$

$$p(\boldsymbol{\theta}; \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) = \prod_{i=1}^{d} \prod_{s=1}^{S_i} \mathcal{B}(\theta_i^s; \alpha_{i,0}^s, \beta_{i,0}^s) \tag{18}$$

where $p(x_i|\mathrm{pa}_i, \boldsymbol{\theta}_i)$ is defined via (9), $\boldsymbol{\alpha}_0$ 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.

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

$$p(\boldsymbol{\theta}|\mathcal{D}) = \prod_{i=1}^{d} \prod_{s=1}^{S_i} \mathcal{B}(\theta_i^s, \alpha_{i,n}^s, \beta_{i,n}^s) \tag{19}$$

where

$$\alpha_{i,n}^s = \alpha_{i,0}^s + n_{x_i=1}^s \qquad \beta_{i,n}^s = \beta_{i,0}^s + n_{x_i=0}^s \tag{20}$$

and that the parameters are also independent under the posterior.

Solution. We start with

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta}; \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0). \tag{S.84}$$

Inserting the expression for $p(\mathcal{D}|\boldsymbol{\theta})$ given in (10) and the assumed form of the prior gives

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto \prod_{i=1}^{d}\prod_{s=1}^{S_i}(\theta_i^s)^{n^s_{x_i=1}}(1-\theta_i^s)^{n^s_{x_i=0}} \prod_{i=1}^{d}\prod_{s=1}^{S_i}\mathcal{B}(\theta_i^s;\alpha^s_{i,0},\beta^s_{i,0}) \tag{S.85}$$

$$\propto \prod_{i=1}^{d}\prod_{s=1}^{S_i}(\theta_i^s)^{n^s_{x_i=1}}(1-\theta_i^s)^{n^s_{x_i=0}}\mathcal{B}(\theta_i^s;\alpha^s_{i,0},\beta^s_{i,0}) \tag{S.86}$$

$$\propto \prod_{i=1}^{d}\prod_{s=1}^{S_i}(\theta_i^s)^{n^s_{x_i=1}}(1-\theta_i^s)^{n^s_{x_i=0}}(\theta_i^s)^{\alpha^s_{i,0}-1}(1-\theta_i^s)^{\beta^s_{i,0}-1} \tag{S.87}$$

$$\propto \prod_{i=1}^{d}\prod_{s=1}^{S_i}(\theta_i^s)^{\alpha^s_{i,0}+n^s_{x_i=1}-1}(1-\theta_i^s)^{\beta^s_{i,0}+n^s_{x_i=0}-1} \tag{S.88}$$

$$\propto \prod_{i=1}^{d}\prod_{s=1}^{S_i}\mathcal{B}(\theta_i^s;\alpha^s_{i,0}+n^s_{x_i=1},\beta^s_{i,0}+n^s_{x_i=0}) \tag{S.89}$$

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.

(b) For a variable x_i with parents pa_i, compute the posterior predictive probability $p(x_i=1|\mathrm{pa}_i,\mathcal{D})$ where $n^s = n^s_{x_i=0}+n^s_{x_i=1}$ denotes the number of times the parent configuration s occurs in the observed data $\mathcal{D}$.

Solution. The solution is analogue to the solution for question (c), using the sum rule, independencies, and properties of beta random variables:

$$p(x_i=1|\mathrm{pa}_i=s,\mathcal{D}) = \int p(x_i=1,\theta_i^s|\mathrm{pa}_i=s,\mathcal{D})\mathrm{d}\theta_i^s \tag{S.90}$$

$$= \int p(x_i=1|\theta_i^s,\mathrm{pa}_i=s,\mathcal{D})p(\theta_i^s|\mathrm{pa}_i=s,\mathcal{D}) \tag{S.91}$$

$$= \int p(x_i=1|\theta_i^s,\mathrm{pa}_i=s)p(\theta_i^s|\mathcal{D}) \tag{S.92}$$

$$= \int \theta_i^s p(\theta_i^s|\mathcal{D}) \tag{S.93}$$

$$= \mathbb{E}[\theta_i^s|\mathcal{D})] \tag{S.94}$$

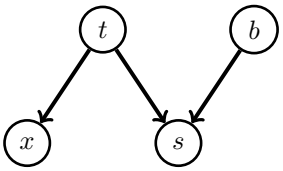
$$\overset{(S.74)}{=} \frac{\alpha^s_{i,n}}{\alpha^s_{i,n}+\beta^s_{i,n}} \tag{S.95}$$

$$= \frac{\alpha^s_{i,0}+n^s_{x_i=1}}{\alpha^s_{i,0}+\beta^s_{i,0}+n^s} \tag{S.96}$$

Exercise 6. *Learning parameters of a directed graphical model*

We consider the directed graphical model shown below on the left for the four binary variables t,b,s,x, each being either zero or one. Assume that we have observed the data shown in the table on the right.

Model:



$t = 1$ *has tuberculosis*
$b = 1$ *has bronchitis*
$s = 1$ *has shortness of breath*
$x = 1$ *has positive x-ray*

Observed data:

x	s	t	b
0	*1*	*0*	*1*
0	*0*	*0*	*0*
0	*1*	*0*	*1*
0	*1*	*0*	*1*
0	*0*	*0*	*0*
0	*0*	*0*	*0*
0	*1*	*0*	*1*
0	*1*	*0*	*1*
0	*0*	*0*	*1*
1	*1*	*1*	*0*

We assume the (conditional) pmf of $s|t, b$ is specified by the following parametrised probability table:

$p(s = 1\|t, b; \theta_s^1, \ldots, \theta_s^4))$	t	b
θ_s^1	*0*	*0*
θ_s^2	*1*	*0*
θ_s^3	*0*	*1*
θ_s^4	*1*	*1*

(a) What are the maximum likelihood estimates for $p(s = 1|b = 0, t = 0)$ and $p(s = 1|b = 0, t = 1)$, i.e. the parameters θ_s^1 and θ_s^2?

Solution. The maximum likelihood estimates (MLEs) are equal to the fraction of occurrences of the relevant events.

$$\hat{\theta}_s^1 = \frac{\sum_{i=1}^n \mathbb{1}(s_i = 1, b_i = 0, t_i = 0)}{\sum_{i=1}^n \mathbb{1}(b_i = 0, t_i = 0)} = \frac{0}{3} = 0 \tag{S.97}$$

$$\hat{\theta}_s^2 = \frac{\sum_{i=1}^n \mathbb{1}(s_i = 1, b_i = 0, t_i = 1)}{\sum_{i=1}^n \mathbb{1}(b_i = 0, t_i = 1)} = \frac{1}{1} = 1 \tag{S.98}$$

(b) Assume each parameter in the table for $p(s|t, b)$ has a uniform prior on $(0, 1)$. Compute the posterior mean of the parameters of $p(s = 1|b = 0, t = 0)$ and $p(s = 1|b = 0, t = 1)$ and explain the difference to the maximum likelihood estimates.

Solution. A uniform prior corresponds to a Beta distribution with hyperparameters $\alpha_0 = \beta_0 = 1$. With Exercise 5 question (b), we have

$$\mathbb{E}(\theta_s^1|\mathcal{D}) = \frac{\alpha_0 + 0}{\alpha_0 + \beta_0 + 3} = \frac{1}{5} \tag{S.99}$$

$$\mathbb{E}(\theta_s^2|\mathcal{D}) = \frac{\alpha_0 + 1}{\alpha_0 + \beta_0 + 1} = \frac{2}{3} \tag{S.100}$$

Compared to the MLE, the posterior mean is less extreme. It can be considered a "smoothed out" or regularised estimate, where $\alpha_0 > 0$ and $\beta_0 > 0$ provides regularisation (see `https://en.wikipedia.org/wiki/Additive_smoothing`). We can see a pull of the parameters towards the prior predictive mean, which equals 1/2.

Exercise 7. ***Maximum likelihood estimation and unnormalised models***

Consider the Ising model for two binary random variables (x_1, x_2)*,*

$$p(x_1, x_2; \theta) \propto \exp\left(\theta x_1 x_2 + x_1 + x_2\right), \qquad x_i \in \{-1, 1\},$$

(a) Compute the partition function $Z(\theta)$*.*

Solution. The definition of the partition function is

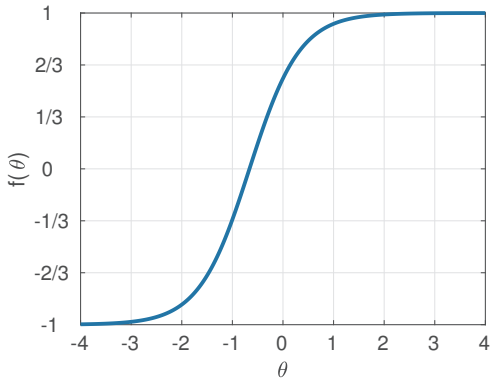
$$Z(\theta) = \sum_{\{-1,1\}^2} \exp\left(\theta x_1 x_2 + x_1 + x_2\right). \tag{S.101}$$

where have have to sum over $(x_1, x_2) \in \{-1, 1\}^2 = \{(-1, 1),\, (1, 1),\, (1, -1)\,(-1 - 1)\}$. This gives

$$\begin{aligned} Z(\theta) &= \exp(-\theta - 1 + 1) + \exp(\theta + 2) + \exp(-\theta + 1 - 1) + \exp(\theta - 2) && \text{(S.102)} \\ &= 2\exp(-\theta) + \exp(\theta + 2) + \exp(\theta - 2) && \text{(S.103)} \end{aligned}$$

(b) The figure below shows the graph of $f(\theta) = \frac{\partial \log Z(\theta)}{\partial \theta}$*.*

Assume you observe three data points (x_1, x_2) *equal to* $(-1, -1)$*,* $(-1, 1)$*, and* $(1, -1)$*. Using the figure, what is the maximum likelihood estimate of* θ*? Justify your answer.*



Solution. Denoting the i-th observed data point by (x_1^i, x_2^i), the log-likelihood is

$$\ell(\theta) = \sum_{i=1}^{n} \log p(x_1^i, x_2^i; \theta) \tag{S.104}$$

Inserting the definition of the $p(x_1, x_2; \theta)$ yields

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^{n} \left[\theta x_1^i x_2^i + x_1^i + x_2^i\right] - n \log Z(\theta) && \text{(S.105)} \\ &= \theta \sum_{i=1}^{n} \left[x_1^i x_2^i\right] + \sum_{i=1}^{n} \left[x_1^i + x_2^i\right] - n \log Z(\theta) && \text{(S.106)} \end{aligned}$$

Its derivative with respect to the θ is

$$\frac{\partial \ell(\theta)}{\partial \theta} = \sum_{i=1}^{n} \left[x_1^i x_2^i\right] - n \frac{\partial \log Z(\theta)}{\partial \theta} \tag{S.107}$$

$$= \sum_{i=1}^{n} \left[x_1^i x_2^i\right] - n f(\theta) \tag{S.108}$$

Setting it to zero yields

$$\frac{1}{n} \sum_{i=1}^{n} \left[x_1^i x_2^i\right] = f(\theta) \tag{S.109}$$

An alternative approach is to start with the more general relationship that relates the gradient of the partition function to the gradient of the log unnormalised model. For example, if

$$p(\mathbf{x}, \boldsymbol{\theta}) = \frac{\phi(\mathbf{x}; \boldsymbol{\theta})}{Z(\boldsymbol{\theta})}$$

we have

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(\mathbf{x}_i; \boldsymbol{\theta}) \tag{S.110}$$

$$= \sum_{i=1}^{n} \log \phi(\mathbf{x}_i; \boldsymbol{\theta}) - n \log Z(\boldsymbol{\theta}) \tag{S.111}$$

Setting the derivative to zero gives,

$$\frac{1}{n} \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \log \phi(\mathbf{x}_i; \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \log Z(\boldsymbol{\theta})$$

In either case, numerical evaluation of $1/n \sum_{i=1}^{n} x_1^i x_2^i$ gives

$$\frac{1}{n} \sum_{i=1}^{n} \left[x_1^i x_2^i\right] = \frac{1}{3} (1 - 1 - 1) \tag{S.112}$$

$$= -\frac{1}{3} \tag{S.113}$$

From the graph, we see that $f(\theta)$ takes on the value $-1/3$ for $\theta = -1$, which is the desired MLE.