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

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

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

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

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 will need the result from Tutorial 5 for taking the product of Gaussians.

Solution. Recall the following result from Tutorial 5:

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

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 Exercise 1 in the main tutorial sheet,

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

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 $\bar{x}$.