Pattern Recognition and Machine Learning
Cristopher Bishop

Exercise Solutions

Stefan Stefanache

August 8, 2021

Chapter 1

Introduction

TODO: 1.15, 1.16, 1.20, 1.26, 1.27 + CALCULUS OF VARIATIONS: 1.25, 1.34

Exercise 1.1 ⋆

Consider the sum-of-squares error function given by (1.2) in which the function $y(x, \mathbf{w})$ is given by the polynomial (1.1). Show that the coefficients $\mathbf{w} = \{w_i\}$ that minimize this error function are given by the solution to the following set of linear equations

$$\sum_{j=0}^{M} A_{ij} w_j = T_i \tag{1.122}$$

where

$$A_{ij} = \sum_{n=1}^{N} (x_n)^{i+j}, \qquad T_i = \sum_{n=1}^{N} (x_n)^i t_n. \tag{1.123}$$

Here a suffix i or j denotes the index of a component, whereas $(x)^i$ denotes x raised to the power of i.

Proof. The function $y(x, \mathbf{w})$ is given by

$$y(x, \mathbf{w}) = \sum_{j=0}^{M} w_j x^j \tag{1.1}$$

and the error function is given by

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 \tag{1.2}$$

Since we want to find the coefficients $\mathbf{w}$ for which the error function is minimized, we compute its derivative with respect to $\mathbf{w}$:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\mathbf{w}} E(\mathbf{w}) =& \frac{\mathrm{d}}{\mathrm{d}\mathbf{w}} \left(\frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 \right) = \frac{1}{2} \sum_{n=1}^{N} \frac{\mathrm{d}}{\mathrm{d}\mathbf{w}} \{y(x_n, \mathbf{w})^2 - 2 t_n y(x_n, \mathbf{w}) + t_n^2\} \\ =& \sum_{n=1}^{N} y(x_n, \mathbf{w}) \frac{\mathrm{d}}{\mathrm{d}\mathbf{w}} y(x_n, \mathbf{w}) - \sum_{n=1}^{N} t_n \frac{\mathrm{d}}{\mathrm{d}\mathbf{w}} y(x_n, \mathbf{w}) \end{aligned} \tag{1.1.1}$$

We continue by computing the derivative of $y(x_n, \mathbf{w})$ separately and obtain that:

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{w}} y(x_n, \mathbf{w}) = \begin{bmatrix} x_n^1 \\ \vdots \\ x_n^M \end{bmatrix} \tag{1.1.2}$$

By substituting the result of (1.1.2) into (1.1.1) we get that:

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{w}} E(\mathbf{w}) = B - T \tag{1.1.3}$$

where T is given by (1.123) and

$$B_i = \sum_{n=1}^{N} x_n^i y(x_n, \mathbf{w})$$

Now, we easily find that

$$B_i = \sum_{n=1}^{N} \left(x_n^i \sum_{j=0}^{M} w_j x_n^j \right) = \sum_{n=1}^{N} \sum_{j=0}^{M} x_n^{i+j} w_j = A_i \mathbf{w}$$

where A is given by (1.123). Now, the critical point of $E(\mathbf{w})$ is given by the equation:

$$A_i \mathbf{w} = T_i$$

which is equivalent with (1.122). □

Exercise 1.2 ⋆

Write down the set of coupled linear equations, analogous to (1.122), satisfied by the coefficients w_i which minimize the regularized sum-of-squares error function given by (1.4).

Proof. The regularized sum-of-squares error function is given by

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2} ||\mathbf{w}||^2 \tag{1.4}$$

We'll have a similar approach to the previous exercise, i.e. we compute the derivative of the regularized error function and find the associated critical point. We notice that

$$\widetilde{E}(\mathbf{w}) = E(\mathbf{w}) + \frac{\lambda}{2} ||\mathbf{w}||^2$$

so

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{w}} \widetilde{E}(\mathbf{w}) = \frac{\mathrm{d}}{\mathrm{d}\mathbf{w}} E(\mathbf{w}) + \frac{\lambda}{2} \cdot \frac{\mathrm{d}}{\mathrm{d}\mathbf{w}} ||\mathbf{w}||^2$$

One could easily prove that

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{w}} ||\mathbf{w}||^2 = 2\mathbf{w}$$

so by using this and (1.1.3) (where we substitute $B = A\mathbf{w}$), we have that:

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{w}}\widetilde{E}(\mathbf{w}) = A\mathbf{w} + \lambda\mathbf{w} - T = (A + \lambda I)\mathbf{w} - T$$

We obtain the critical point when the derivative is 0, so when

$$(A + \lambda I)\mathbf{w} = T$$

which is equivalent with the system of linear equations

$$\sum_{j=0}^{M} C_{ij} w_j = T_i$$

where

$$C_{ij} = A_{ij} + \lambda I_{ij}$$

□

Exercise 1.3 ⋆⋆

Suppose that we have three coloured boxes r (red), b (blue), and g (green). Box r contains 3 apples, 4 oranges and 3 limes, box b contains 1 apple, 1 orange, and 0 limes, and box g contains 3 apples, 3 oranges, and 4 limes. If a box is chosen at random with probabilities $p(r) = 0.2$, $p(b) = 0.2$, $p(g) = 0.6$, and a picee of fruit is removed from the box (with equal probability of selecting any of the items in the box), then what is the probability of selecting an apple? If we observe that the selected fruit is in fact an orange, what is the probability that it came from the green box?

Proof. The conditional probabilities of obtaining a fruit knowing that we are searching in a certain box are easily found since the fruits are equally likely to be extracted. We also now the probabilities of choosing a specific box, so we can simply apply the sum rule to obtain the probability of getting an apple:

$$p(\text{apple}) = p(\text{apple}|r)p(r) + p(\text{apple}|b)p(b) + p(\text{apple}|g)p(g) = \frac{3}{10}\cdot 0.2 + \frac{1}{2}\cdot 0.2 + \frac{3}{10}\cdot 0.6 = 34\%$$

If we know the selected fruit is an orange, the probability that it came from the green box is given by the Bayes' theorem:

$$p(g|\text{orange}) = \frac{p(g)p(\text{orange}|g)}{p(\text{orange})} \tag{1.3.1}$$

The probability of choosing the green box is known and the probability of getting an orange from the green box is also easily found. We only need to find the probability of extracting an orange in the general case:

$$p(\text{orange}) = p(\text{orange}|r)p(r) + p(\text{orange}|b)p(b) + p(\text{orange}|g)p(g) = \frac{4}{10}\cdot 0.2 + \frac{1}{2}\cdot 0.2 + \frac{3}{10}\cdot 0.6 = 36\%$$

The needed probability is now found by substituting the values in (1.3.1):

$$p(g|\text{orange}) = \frac{0.6 \cdot \frac{3}{10}}{\frac{36}{100}} = \frac{1}{2} = 50\%$$

□

Exercise 1.4 ⋆⋆

Consider a probability density $p_x(x)$ defined over a continuous variable x, and suppose that we make a nonlinear change of variable using $x = g(y)$, so that the density transforms according to (1.27). By differentiating (1.27), show that the location $\widehat{y}$ of the maximum of the density in y is not in general related to the location $\widehat{x}$ of the maximum of the density over x by the simple functional relation $\widehat{x} = g(\widehat{y})$ as a consequence of the Jacobian factor. This shows that the maximum of a probability density (in contrast to a simple function) is dependent of the choice of variable. Verify that, in the case of a linear transformation, the location of the maximum transforms in the same way as the variable itself.

Proof. If we make a nonlinear change of variable $x = g(y)$ in the probbability density $p_x(x)$, it transforms according to

$$p_y(y) = p_x(g(y))|g'(y)| \tag{1.27}$$

We assume that the mode of $p_x(x)$ is given by an unique $\widehat{x}$, i.e.

$$p'_x(x) = 0 \iff x = \widehat{x}$$

Now, let $s \in \{-1, 1\}$ such that $g'(y) = sg'(y)$. The derivative of (1.27) with respect to y is given by:

$$p'_y(y) = sp'x(g(y))\{g'(y)\}^2 + sp_x(g(y))g''(y)$$

For a linear change of variable, we have that $g''(y) = 0$, so the mode of $p_y(y)$ is given by $g'(y) = 0$ and since $x = g(y)$, respectively $x' = g'(y)$ we have that $\widehat{x} = g(\widehat{y})$. Therefore, for a linear change of variable, the location of the maximum transforms in the same way as the variable itself.

For a nonlinear change of variable, the second derivative will not be generally 0, so the mode is not given by $g'(y) = 0$ anymore. As a result, in general $\widehat{x} \neq g(\widehat{y})$, so the location of the mode will transform differently from the variable itself. □

Exercise 1.5 ⋆

Using the definition (1.38) show that $\mathrm{var}[f(x)]$ satisfies (1.39).

Proof. The variance is defined by

$$\mathrm{var}[f] = \mathbb{E}\big[(f(x) - \mathbb{E}[f(x)])^2\big] \tag{1.38}$$

We expand the square and then use the linearity of expectation to obtain:

$$\mathrm{var}[f] = \mathbb{E}\big[f(x)^2 - 2f(x)\mathbb{E}[f(x)] + \mathbb{E}[f(x)]^2\big] = \mathbb{E}[f(x)^2] - 2\mathbb{E}\big[f(x)\mathbb{E}[f(x)]\big] + \mathbb{E}\big[\mathbb{E}[f(x)]^2\big]$$

Since $\mathbb{E}[f(x)]$ is a constant, the expression of the variance becomes:

$$\mathrm{var}[f] = \mathbb{E}[f(x)^2] - 2\mathbb{E}[f(x)]^2 + \mathbb{E}[f(x)]^2 = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2 \tag{1.39}$$

□

Exercise 1.6 ⋆

Show that if two variables x and y are independent, then their covariance is zero.

Proof. The covariance of two random variables is given by:

$$\text{cov}[x, y] = \mathbb{E}_{x,y}[xy] - E[x]E[y] \tag{1.41}$$

We assume that the variables are continuous, but the discrete case result is similarly obtained. If x and y are independent, we have that $p(x, y) = p(x)p(y)$, so

$$E_{x,y}[xy] = \iint p(x, y)xy \ \mathrm{d}x\,\mathrm{d}y = \iint p(x)p(y)xy \ \mathrm{d}x\,\mathrm{d}y = \Big(\int p(x)x \ \mathrm{d}x \Big)\Big(\int p(y)y \ \mathrm{d}y \Big) = E[x]E[y]$$

and (1.41) becomes 0. □

Exercise 1.7 ⋆⋆

In this exercise, we prove the normalization condition (1.48) for the univariate Gaussian. To do this consider the integral

$$I = \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}x^2 \right) \mathrm{d}x \tag{1.124}$$

which we can evaluate by first writing its square in the form

$$I^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}x^2 - \frac{1}{2\sigma^2}y^2 \right) \mathrm{d}x\,\mathrm{d}y \tag{1.125}$$

Now make the transformation from Cartesian coordinates (x, y) to polar coordinates (r, θ) and then substitute $u = r^2$. Show that, by performing the integrals over θ and u, and then taking the square root of both sides, we obtain

$$I = (2\pi\sigma^2)^{1/2} \tag{1.126}$$

Finally, use this result to show that the Gaussian distribution $\mathcal{N}(x|\mu, \sigma^2)$ is normalized.

Proof. We transform (1.125) from Cartesian coordinates to polar coordinates and obtain:

$$I^2 = \int_0^{2\pi}\int_0^{\infty} \exp\left(-\frac{r^2\sin^2\theta + r^2\cos^2\theta}{2\sigma^2} \right) r \,\mathrm{d}r\,\mathrm{d}\theta = \int_0^{2\pi}\int_0^{\infty} \exp\left(-\frac{r^2}{2\sigma^2} \right) r \,\mathrm{d}r\,\mathrm{d}\theta$$

We use the substitution $u = r^2$ and then compute the integral to get:

$$I^2 = \frac{1}{2}\int_0^{2\pi}\int_0^{\infty} \exp\left(-\frac{u}{2\sigma^2} \right) du\,\mathrm{d}\theta = \frac{1}{2}\int_0^{2\pi} -2\sigma^2 \exp\left(-\frac{u}{2\sigma^2} \right)\Big|_0^{\infty} \mathrm{d}\theta = \sigma^2 \int_0^{2\pi} \mathrm{d}\theta = 2\pi\sigma^2$$

If we take the square root of this we see that

$$I = (2\pi\sigma^2)^{1/2} \tag{1.126}$$

We can assume without loss of generality that the mean of the Gaussian is 0, as we could make the change of variable $y = x - \mu$. Therefore, by using (1.126) we obtain

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{x}{2\sigma^2} \right) \mathrm{d}x = \frac{I}{\sqrt{2\pi\sigma^2}} = 1$$

which shows that the Gaussian distribution is normailized. □

Exercise 1.8 ⋆⋆

By using a change of variables, verify that the univariate Gaussian given by (1.46) satisfies (1.49). Next, by differentiating both sides of the normalization condition

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2)\, \mathrm{d}x = 1 \tag{1.127}$$

with respect to σ^2, verify that the Gaussian satisfies (1.50). Finally, show that (1.51) holds.

Proof. We start by computing the expected value of the Gaussian:

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) x\, \mathrm{d}x = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} x\, \mathrm{d}x$$

We do a little trick to prepare for the substitution $u = (x-\mu)^2$:

$$\mathbb{E}[x] = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} (x-\mu)\, \mathrm{d}x + \frac{\mu}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} \mathrm{d}x$$

Since the Gaussian is normalized, the second term of the expression will be μ. By using the substitution $u = (x-\mu)^2$, the expected value becomes:

$$\mathbb{E}[x] = \frac{1}{2\sqrt{2\pi\sigma^2}} \int_{\infty}^{\infty} \exp\left(-\frac{u}{2\sigma^2} \right) \mathrm{d}u + \mu$$

We notice that the endpoints of the integral are "equal" (one could rewrite it as a limit of an integral with actual equal endpoints), so its value is 0. Therefore,

$$\mathbb{E}[x] = \mu \tag{1.49}$$

Now, we take the derivative of (1.127) with respect to σ^2 and obtain:

$$\begin{aligned}
\frac{\partial}{\partial \sigma^2} \left(\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} \mathrm{d}x \right) &= 0 \\
-\frac{I}{2\sigma^3\sqrt{2\pi}} + \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \frac{\partial}{\partial \sigma^2} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} \mathrm{d}x &= 0 \\
-\frac{1}{2\sigma^2} + \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{2\sigma^4} \exp\left\{ -\frac{(x+\mu)^2}{2\sigma^2} \right\} \mathrm{d}x &= 0
\end{aligned}$$

We let J be the integral term and compute it separately:

$$\begin{aligned}
J &= \frac{1}{2\sigma^4} \int_{-\infty}^{\infty} (x-\mu)^2 \exp\left\{ -\frac{(x+\mu)^2}{2\sigma^2} \right\} \mathrm{d}x \\
&= \frac{1}{2\sigma^4} \int_{-\infty}^{\infty} x^2 \exp\left\{ -\frac{(x+\mu)^2}{2\sigma^2} \right\} \mathrm{d}x - \frac{2\mu}{2\sigma^4} \int_{-\infty}^{\infty} x \exp\left\{ -\frac{(x+\mu)^2}{2\sigma^2} \right\} \mathrm{d}x + \frac{\mu^2}{2\sigma^4} I
\end{aligned}$$

If we multiply by the normalization constants, the integrals become expected values and the I factor vanishes. Therefore:

$$J = \sqrt{2\pi\sigma^2} \left(\frac{1}{2\sigma^4} \mathbb{E}[x^2] - \frac{2\mu}{2\sigma^4} \mathbb{E}[x] + \frac{\mu^2}{2\sigma^4} \right)$$

We substitute J back in the initial expression to obtain:

$$-\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbb{E}[x^2] - 2\mu^2 + \mu^2) = 0$$

from which is straightforard to show that

$$E[x^2] = \sigma^2 + \mu^2 \tag{1.50}$$

Finally, one can easily see that:

$$\text{var}[x] = E[x^2] - E[x]^2 = \sigma^2 \tag{1.51}$$

□

Exercise 1.9 ⋆

Show that the mode (i.e. the maximum) of the Gaussian distribution (1.46) is given by μ. Similarly, show that the mode of the multivariate Gaussian (1.52) is given by $\boldsymbol{\mu}$.

Proof. In the univariate case, we start by taking the derivative of (1.46) with respect to x :

$$\frac{\partial}{\partial x}\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\left(\frac{\partial}{\partial x}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\frac{(x-\mu)^2}{2\sigma^4}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

We notice that the derivative is 0, for $x = \mu$, so the mode of the univariate Gaussian is given by the mean.

Analogously, we take the derivative of (1.52) with respect to $\mathbf{x}$ and get:

$$\frac{\partial}{\partial \mathbf{x}}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\left(\frac{\partial}{\partial \mathbf{x}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}\right)$$

The covariance matrix $\boldsymbol{\Sigma}$ is both nonsingular and symmetric, so one can easily show that $\boldsymbol{\Sigma}^{-1}$ will be symmetric too. Therefore, we have that (see matrix cookbook):

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = 2\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

As a result, our derivative becomes

$$\frac{\partial}{\partial \mathbf{x}}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = -\frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

and is 0 for $\mathbf{x} = \boldsymbol{\mu}$, so like in the case of the univariate distribution, the mode of the multivariate distribution is given by the mean $\boldsymbol{\mu}$. □

Exercise 1.10 ⋆

Suppose that the two variables x and z are statistically independent. Show that the mean and variance of their sum satisfies

$$\mathbb{E}[x+z] = \mathbb{E}[x] + \mathbb{E}[z] \tag{1.128}$$

$$\text{var}[x+z] = \text{var}[x] + \text{var}[z] \tag{1.129}$$

Proof. Since the variables are independent, we have that $p(x,z) = p(x)p(z)$. Therefore, by using this, the expression of the expected value and the fact that the distributions are normalized, we have that

$$\begin{aligned}
\mathbb{E}[x+z] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,z)(x+z)\,\mathrm{d}x\,\mathrm{d}z \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x)p(z)x + p(x)p(z)z\,\mathrm{d}x\,\mathrm{d}z \\
&= \int_{-\infty}^{\infty} p(z)\left(\int_{-\infty}^{\infty} p(x)x\,\mathrm{d}x\right) + p(z)z\left(\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x\right)\mathrm{d}z \\
&= \int_{-\infty}^{\infty} p(z)\mathbb{E}[x] + p(z)z\,\mathrm{d}z \\
&= \mathbb{E}[x]\int_{-\infty}^{\infty} p(z)\,\mathrm{d}z + \int_{-\infty}^{\infty} p(z)z\,\mathrm{d}z \\
&= \mathbb{E}[x] + \mathbb{E}[z]
\end{aligned} \tag{1.128}$$

Analogously, we can solve the discrete case. Now, by using all the available tools, i.e. (1.39) and (1.128), the linearity of the expectation and the independence of variables, we have that the variance of the sum is given by:

$$\begin{aligned}
\text{var}[x+z] &= \mathbb{E}[(x+z)^2] - \mathbb{E}[x+z]^2 = \mathbb{E}[x^2+2xz+z^2] - (\mathbb{E}[x]+\mathbb{E}[z])^2 \\
&= \mathbb{E}[x^2] + 2\mathbb{E}[x]\mathbb{E}[z] + \mathbb{E}[z^2] - \mathbb{E}[x]^2 - \mathbb{E}[x^2+2xz+z^2] - E[z]^2 \\
&= E[x^2] - E[x]^2 + E[z^2] - E[z]^2 \\
&= \text{var}[x] + \text{var}[z]
\end{aligned} \tag{1.129}$$

□

Exercise 1.11 ⋆

By setting the derivatives of the log likelihood function (1.54) with respect to μ and σ^2 equal to zero, verify the results (1.55) and (1.56).

Proof. The log likelihood of the Gaussian is given by:

$$\ln p(\mathbf{x}|\mu,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi) \tag{1.54}$$

By taking the derivative of (1.54) with respect to μ we get that:

$$\frac{\partial}{\partial \mu} \ln p(\mathbf{x}|\mu, \sigma^2) = -\frac{1}{2\sigma^2}\left\{ \frac{\partial}{\partial \mu} \sum_{n=1}^{N} (x_n - \mu)^2 \right\} = -\frac{1}{2\sigma^2}\left\{ \frac{\partial}{\partial \mu}\left(\sum_{n=1}^{N} x_n^2 - 2\sum_{n=1}^{N} x_n \mu + N\mu^2 \right) \right\}$$
$$= \frac{1}{\sigma^2}\left(\sum_{n=1}^{N} x_n - N\mu \right)$$

which is 0 for the maximum point:

$$\mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n \tag{1.55}$$

Now, we want the variance that maximizes the log likelihood, so we take the derivative of (1.54) (by using μ_{ML}) with respect to σ^2:

$$\frac{\partial}{\partial \sigma^2} \ln p(\mathbf{x}|\mu_{ML}, \sigma^2) = \frac{1}{2\sigma^4} \sum_{n=1}^{N} (x_n - \mu_{ML})^2 - \frac{N}{2\sigma^2} = \frac{1}{2\sigma^4}\left(\sum_{n=1}^{N} (x_n - \mu_{ML})^2 - N\sigma^2 \right)$$

The derivative is 0 for the maximum point

$$\sigma^2_{ML} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{ML})^2 \tag{1.56}$$

□

Exercise 1.12 ⋆⋆

Using the results (1.49) and (1.50), show that

$$\mathbb{E}[x_n x_m] = \mu^2 + I_{nm}\sigma^2 \tag{1.130}$$

where x_n and x_m denote data points sampled from a Gaussian distribution with mean μ and variance σ^2, and I_{nm} satisfies $I_{nm} = 1$ if $n = m$ and $I_{nm} = 0$ otherwise. Hence prove the results (1.57) and (1.58).

Proof. We assume that the data points are i.i.d, so we have that the variables x_n and x_m are not independent for $n \neq m$ and independent for $n = m$. Therefore,

$$\mathbb{E}[x_n x_m] = \begin{cases} \mu^2 & n \neq m \\ \mu^2 + \sigma^2 & n = m \end{cases}$$

which is equivalent with (1.130). Now, the expectation of μ_{ML} is given by:

$$\mathbb{E}[\mu_{ML}] = \mathbb{E}\left[\frac{1}{N} \sum_{n=1}^{N} x_n \right] = \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}[x_n] = \mu \tag{1.57}$$

Similarly, the expectation of σ^2_{ML} is given by:

$$\mathbb{E}[\sigma^2_{ML}] = \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2\right] = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[x_n^2 - 2x_n\mu_{ML} + \mu^2_{ML}]$$
$$= \frac{1}{N}\sum_{n=1}^{N}(\mu^2 + \sigma^2 - 2\mathbb{E}[x_n\mu_{ML}] + \mathbb{E}[\mu^2_{ML}])$$

We compute each expectation separately and get:

$$E[\mu^2_{ML}] = \frac{1}{N^2}\mathbb{E}\left[\sum_{n=1}^{N}x_n^2 + 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}x_ix_j\right] = \frac{1}{N^2}\sum_{n=1}^{N}\mathbb{E}[x_n^2] + \frac{2}{N^2}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\mathbb{E}[x_ix_j] = \frac{\sigma^2}{N} + \mu^2$$

$$E[x_n\mu_{ML}] = \frac{1}{N}\mathbb{E}\left[x_n\sum_{i=1}^{N}x_i\right] = \frac{1}{N}(\sigma^2 + N\mu^2) = \frac{\sigma^2}{N} + \mu^2$$

By putting everything together, we obtain

$$\mathbb{E}[\sigma^2_{ML}] = \left(\frac{N-1}{N}\right)\sigma^2 \tag{1.58}$$

□

Exercise 1.13 ⋆

Suppose that the variance of a Gaussian is estimated using the result (1.56) but with the maximum likelihood estimate μ_{ML} replaced with the true value μ of the mean. Show that this estimator has the property that its expectation is given by the true variance σ^2.

Proof. Let

$$\sigma^*_{ML}{}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2$$

be the estimator described in the hypothesis. It's straightforward to show that the expectation of the estimator is the actual variance:

$$\mathbb{E}[\sigma^*_{ML}{}^2] = \frac{1}{N}\sum_{n=1}^{N}\Big(\mathbb{E}[x_n^2] - 2\mathbb{E}[x_n\mu] + \mathbb{E}[\mu^2]\Big) = \frac{1}{N}\sum_{n=1}^{N}(\sigma^2 + \mu^2 - 2\mu^2 + \mu^2) = \sigma^2$$

□

Exercise 1.14 ⋆⋆

Show that an arbitrary square matrix with elements w_{ij} can be written in the form $w_{ij} = w^S_{ij} + w^A_{ij}$ where w^S_{ij} and w^A_{ij} are symmetric and anti-symmetric matrices, respectively, satisfying $w^S_{ij} = w^S_{ji}$

and $w_{ij}^A = -w_{ji}^A$ for all i and j. Now consider the second order term in a higher order polynomial in D dimensions, given by

$$\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}x_i x_j \tag{1.131}$$

Show that

$$\sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}x_i x_j = \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}^S x_i x_j \tag{1.132}$$

so that the contribution from the anti-symmetric vanishes. We therefore see that, without loss of generality, the matrix of coefficients w_{ij} can be chosen to be symmetric, and so not all of the D^2 elements of this matrix can be chosen independently. Show that the number of independent parameters in the matrix w_{ij}^S is given by $D(D+1)/2$.

Proof. If we consider the system of equations

$$w_{ij} = w_{ij}^S + w_{ij}^A \qquad w_{ji} = w_{ij}^S - w_{ij}^A$$

we quickly reach the conclusion that the solutions are given by

$$w_{ij}^S = \frac{w_{ij} + w_{ji}}{2} \qquad w_{ij}^A = \frac{w_{ij} - w_{ji}}{2} \tag{1.14.1}$$

such that for all i and j,

$$w_{ij} = w_{ij}^S + w_{ij}^A$$

The coefficient matrix w associated with the second order higher order polynomial in D dimensions is actually a $D \times D$ *symmetric* matrix. Therefore, from (1.14.1) we'd have that $w^S = w$ and $w_A = 0_D$, where 0_D is the null matrix of dimension D, so (1.132) definitely holds as the anti-symmetric contribution vanishes.

We consider as independent parameters of the matrix w the elements on and above the diagonal, since the ones under the diagonal are reflections of the ones above. There are

$$\sum_{i=1}^{D}(D-i+1) = D^2 + D - \sum_{i=1}^{D} i = D^2 + D - \frac{D(D+1)}{2} = \frac{D(D+1)}{2}$$

such independent parameters □

Exercise 1.15 ⋆⋆⋆

In this exercise and the next, we explore how the number of independent parameters in a polynomial grows with the order M of the polynomial and with the dimensionality D of the input space. We start by writing down the M^{th} order term for a polynomial in D dimensions in the form

$$\sum_{i_1=1}^{D}\sum_{i_2=1}^{D}\cdots\sum_{i_M=1}^{D} w_{i_1,i_2,\ldots,i_M}x_{i_1}x_{i_2}\cdot\ldots x_{i_M} \tag{1.133}$$

The coefficients $w_{i_1,i_2,\ldots i_M}$ compromise D^M elements, but the number of independent parameters is significantly fewer due to the many interchange symmetries of the factor $x_{i_1}, x_{i_2}\ldots x_{i_M}$. Begin

by showing that the redundancy in the coefficients can be removed by rewriting the M^{th} order term in the form

$$\sum_{i_1=1}^{D}\sum_{i_2=1}^{i_1}\cdots\sum_{i_M=1}^{i_{M-1}}\widetilde{w}_{i_1,i_2,\ldots,i_M}x_{i_1}x_{i_2}\cdot\ldots x_{i_M} \tag{1.134}$$

Note that the precise relationship between the $\widetilde{w}$ coefficients and w coefficients need not be made explicit. Use this result to show that the number of *independent* parameters $n(D, M)$, which appear at order M, satisfies the following recursion relation

$$n(D, M) = \sum_{i=1}^{D} n(i, M-1) \tag{1.135}$$

Next use proof by induction to show that the following result holds

$$\sum_{i=1}^{D}\frac{(i+M-2)!}{(i-1)!(M-1)!} = \frac{(D+M-1)!}{(D-1)!M!} \tag{1.136}$$

which can be done by first proving the result for $D = 1$ and arbitrary M by making use of the result $0! = 1$, then assuming it is correct for dimension D and verifying that it is correct for dimension $D + 1$. Finally, use the two previous results, together with proof by induction, to show

$$n(D, M) = \frac{(D+M-1)!}{(D-1)!M!} \tag{1.137}$$

To do this, first show that the result is true for $M = 2$, and any value of $D \geq 1$, by comparison with the result of Exercise 1.14. Then make use of (1.135), together with (1.136), to show that, if the result holds at order $M - 1$, then it will also hold at order M.

Proof. □

Exercise 1.17 ⋆⋆

The gamma function is defined by

$$\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}\,\mathrm{d}u \tag{1.141}$$

Using integration by parts, prove the relation $\Gamma(x + 1) = x\Gamma(x)$. Show also that $\Gamma(1) = 1$ and hence that $\Gamma(x + 1) = x!$ when x is an integer.

Proof. Knowing that $-u^x e^{-u} \to 0$ as $u \to \infty$, we integrate $\Gamma(x + 1)$ by parts and obtain:

$$\Gamma(x+1) = \int_0^\infty u^x(-e^{-u})'\,\mathrm{d}u = -u^x e^{-u}\Big|_0^\infty + x\int_0^\infty u^{x-1}e^{-u}\,\mathrm{d}u = x\Gamma(x)$$

Computing $\Gamma(1)$ is also easily done by integrating by parts:

$$\Gamma(1) = \int_0^\infty ue^{-u}\,\mathrm{d}u = \int_0^\infty u(-e^{-u})'\,\mathrm{d}u = -ue^{-u}\Big|_0^\infty + \int_0^\infty e^{-u}\,\mathrm{d}u = 1$$

We can prove by induction that $\Gamma(x+1) = x!$ when x is an integer. This is obviously valid for $x = 0$, since $0! = 1$. Now, assume that $\Gamma(k) = (k-1)!$, for $k \in \mathbb{N}$. Then,

$$\Gamma(k+1) = k\Gamma(k) = k \cdot (k-1)! = k!$$

Therefore, $\Gamma(n+1) = n!$ for all $n \in \mathbb{N}$. □

Exercise 1.18 ⋆⋆

We can use the result (1.126) to derive an expression for the surface area S_D and the volume V_D, of a sphere of unit radius in D dimensions. To do this, consider the following result, which is obtained by transforming from Cartesian to polar coordinates

$$\prod_{i=1}^{D} \int_{-\infty}^{\infty} e^{-x_i^2} \, \mathrm{d}x_i = S_D \int_0^{\infty} e^{-r^2} r^{D-1} \, \mathrm{d}r \tag{1.142}$$

Using the definition (1.141) of the Gamma function, together with (1.126), evaluate both sides of this equation, and hence show that

$$S_D = \frac{2\pi^{D/2}}{\Gamma(D/2)} \tag{1.143}$$

Next, by integrating with respect to radius from 0 to 1, show that the volume of the unit sphere in D dimensions is given by

$$V_D = \frac{S_D}{D} \tag{1.144}$$

Finally, use the results $\Gamma(1) = 1$ and $\Gamma(3/2) = \sqrt{\pi}/2$ to show that (1.143) and (1.144) reduce to the usual expressions for $D = 2$ and $D = 3$.

Proof. We observe that the left side factor of (1.142) looks like (1.126) for $\sigma^2 = 1/2$. Therefore,

$$\prod_{i=1}^{D} \int_{-\infty}^{\infty} e^{-x_i^2} \, \mathrm{d}x_i = \prod_{i=1}^{D} \pi^{1/2} = \pi^{D/2}$$

One can easily notice that the integral in the right side of (1.142) can be written as:

$$\int_0^{\infty} e^{-r^2} r^{D-1} \, \mathrm{d}r = \int_0^{\infty} e^{-r^2} (r^2)^{(D-2)/2} r \, \mathrm{d}r = \frac{1}{2} \int_0^{\infty} e^{-u} u^{(D-2)/2} \, \mathrm{d}u = \frac{1}{2}\Gamma(D/2) \, \mathrm{d}u$$

where we made the substitution $u = r^2$.

Therefore, from those results and from (1.142), we find that

$$S_D = \frac{2\pi^{D/2}}{\Gamma(D/2)} \tag{1.143}$$

The volume of the unit hypersphere is now given by the integral

$$V_D = \int_0^1 S_D r^{D-1} \, \mathrm{d}r = \frac{S_D}{D} \tag{1.144}$$

Now, we get the expected results for $D = 2$ and $D = 3$:

$$S_2 = \frac{2\pi}{\Gamma(1)} = 2\pi \qquad V_2 = \pi \qquad S_3 = \frac{2\pi^{3/2}}{\Gamma(\frac{3}{2})} = 4\pi \qquad V_3 = \frac{4\pi}{3}$$

□

Exercise 1.19 ⋆⋆

Consider a sphere of radius a in D-dimensions together with the concentric hypercube of side $2a$, so that the sphere touches the hypercube at the centres of each of its sides. By using the results of Exercise 1.18, show that the ratio of the volume of the sphere to the volume of the cube is given by

$$\frac{\text{volume of sphere}}{\text{volume of cube}} = \frac{\pi^{D/2}}{D2^{D-1}\Gamma(D/2)} \tag{1.145}$$

Now, make use of Stirling's formula in the form

$$\Gamma(x+1) \simeq (2\pi)^{1/2} e^{-x} x^{x+1/2} \tag{1.146}$$

which is valid for $x \gg 1$, to show that, as $D \to \infty$, the ratio (1.145) goes to zero. Show also that the ratio of the distance from the centre of the hypercube to one of the corners, divided by the perpendicular distance to one of the sides, is $\sqrt{D}$, which therefore goes to ∞ as $D \to \infty$. From these results we see that, in a space of high dimensionality, most of the volume of a cube is concentrated in a large number of corners, which themselves become very lone 'spikes'!

Proof. Using the results of Exercise 1.18, we have that the volume of D-dimensional hypersphere of radius a is

$$V_{D_{\text{sphere}}}(a) = \frac{2\pi^{D/2} a^D}{D\Gamma(D/2)}$$

We also know that the volume of the $D-$hypercube of size $2a$ is given by:

$$V_{D_{\text{cube}}(2a)} = (2a)^D = 2^D a^D$$

Therefore the ratio of the volumes is given by

$$\frac{V_{D_{\text{sphere}}(a)}}{V_{D_{\text{cube}}}(a)} = \frac{\pi^{D/2}}{D2^{D-1}\Gamma(D/2)} \tag{1.145}$$

By using Stirling's approximation, we have that

$$\begin{aligned}
\lim_{D\to\infty} \frac{\pi^{D/2}}{D2^{D-1}\Gamma(D/2)} &= \lim_{D\to\infty} \frac{\pi^{D/2}}{D2^{D-1}(2\pi)^{1/2} e^{1-D/2}(D/2-1)^{D/2-1/2}} \\
&= \lim_{D\to\infty} \left\{ \left(\frac{\pi}{4}\right)^{D/2} \cdot \left(\frac{e}{D/2-1}\right)^{D/2-1} \cdot \frac{\sqrt{D-2}}{D\sqrt{\pi}} \right\} = 0
\end{aligned}$$

Now, we want to find the ratio between the distance from the centre of the hypercube to one of the corners and the distance from the centre to a side. We can consider without loss of generality a D-dimensional hypercube of length 2α, centered in the origin 0_D of the $\mathbb{R}^D$ Cartesian system. The center of a hypercube side takes the form $\mathbf{s} = (\alpha_1, \alpha_2, \ldots, \alpha_D)$, where $\alpha_i \in \{0, a\}$ such that $||\mathbf{s}|| = a$, i.e. only one coordinate is equal to a and the rest are 0. On the other hand, the corners of the hypercube take the form $\mathbf{c} = (\beta_1, \beta_2, \ldots, \beta_D)$, where $\beta_i \in \{\pm a\}$. We'll then have that $||\mathbf{c}|| = a\sqrt{D}$. As a result, our ratio looks like expected:

$$\frac{\text{distance from center to corner}}{\text{distance from center to side}} = \frac{||\mathbf{s}||}{||\mathbf{c}||} = \frac{a\sqrt{D}}{a} = \sqrt{D}$$

□

Exercise 1.21 ⋆⋆

Consider two nonnegative numbers a and b, and show that, if $a \leq b$, then $a \leq (ab)^{1/2}$. Use this result to show that, if the decision regions of a two-class classification problem are chosen to minimize the probability of misclassification, this probability will satisfy

$$p(\text{mistake}) \leq \int \{p(\mathbf{x}, \mathcal{C}_1)p(\mathbf{x}, \mathcal{C}_2)\}^{1/2} \, \mathrm{d}\mathbf{x} \tag{1.150}$$

Proof. We start by proving the identity. We have that

$$a \leq (ab)^{1/2} \iff a^2 \leq ab \iff a^2 - ab \leq 0 \iff a(a-b) \leq 0$$

which is true since $a \leq b$.

Now, since the regions are chosen to minimize the probability of misclassification, for an individual value of $\mathbf{x}$, the region $\mathcal{R}_k$ with the higher joint/posterior probability associated to $\mathcal{C}_k$ is chosen, so:

$$p(\mathbf{x}, \mathcal{C}_2) \leq p(\mathbf{x}, \mathcal{C}_1), \forall \mathbf{x} \in \mathcal{R}_1 \qquad p(\mathbf{x}, \mathcal{C}_1) \leq p(\mathbf{x}, \mathcal{C}_2), \forall \mathbf{x} \in \mathcal{R}_2$$

By applying the $a \leq (ab)^{1/2}$ identity above, we get that

$$p(\mathbf{x}, \mathcal{C}_2) \leq \{p(\mathbf{x}, \mathcal{C}_1)p(\mathbf{x}, \mathcal{C}_2)\}^{1/2}, \forall \mathbf{x} \in \mathcal{R}_1 \qquad p(\mathbf{x}, \mathcal{C}_1) \leq \{p(\mathbf{x}, \mathcal{C}_1)p(\mathbf{x}, \mathcal{C}_2)\}^{1/2}, \forall \mathbf{x} \in \mathcal{R}_2$$

If we integrate the inequalities over the associated regions, we have that:

$$\int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2) \, \mathrm{d}\mathbf{x} \leq \int_{\mathcal{R}_1} \{p(\mathbf{x}, \mathcal{C}_1)p(\mathbf{x}, \mathcal{C}_2)\}^{1/2} \, \mathrm{d}\mathbf{x}$$

$$\int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1) \, \mathrm{d}\mathbf{x} \leq \int_{\mathcal{R}_2} \{p(\mathbf{x}, \mathcal{C}_1)p(\mathbf{x}, \mathcal{C}_2)\}^{1/2} \, \mathrm{d}\mathbf{x}$$

By summing the above inequalities, we find that:

$$\int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2) \, \mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1) \, \mathrm{d}\mathbf{x} \leq \int \{p(\mathbf{x}, \mathcal{C}_1)p(\mathbf{x}, \mathcal{C}_2)\}^{1/2}$$

which is equivalent to (1.150). □

Exercise 1.22 ⋆

Given a loss matrix with elements L_{kj}, the expected risk is minimized, if for each $\mathbf{x}$, we choose the class that minimizes (1.81). Verify that, when the loss matrix is given by $L_{kj} = 1 - I_{kj}$, where I_{kj} are the elements of the identity matrix, this reduces to the criterion of choosing the class having the largest posterior probability. What is the interpretation of this form of loss matrix?

Proof. The expectation is minimized if for each $\mathbf{x}$ we choose the class $\mathcal{C}_j$ such that the quantity

$$\sum_k L_{kj} p(\mathcal{C}_k | \mathbf{x}) \tag{1.81}$$

is minimized. For $L_{kj} = 1 - I_{kj}$ the quantity becomes

$$\sum_k (1 - I_{kj}) p(\mathcal{C}_k|\mathbf{x}) = \sum_k p(\mathcal{C}_k|\mathbf{x}) - p(\mathcal{C}_j|\mathbf{x}) = 1 - p(\mathcal{C}_j|\mathbf{x})$$

and it's obviously minimised by choosing the class $\mathcal{C}_j$ having the largest posterior probability $p(\mathcal{C}_j|\mathbf{x})$

This form of loss matrix makes each mistake have the same "weight", no mistake is worse than another. □

Exercise 1.23 ⋆

Derive the criterion for minimizing the expected loss when there is a general loss matrix and general prior probabilities for the classes.

Proof. Minimizing the expected loss

$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_j) \, \mathrm{d}\mathbf{x} \tag{1.80}$$

is equivalent with minimizing

$$\sum_k L_{kj} p(\mathbf{x}, \mathcal{C}_k)$$

for each $\mathbf{x}$. Therefore, by using Bayes' theorem, we have that the criterion of minimizing the expected loss is the class $\mathcal{C}_j$ for each $\mathbf{x}$ such that

$$\sum_k L_{kj} p(\mathbf{x}|\mathcal{C}_k)$$

is minimized. □

Exercise 1.24 ⋆⋆

Consider a classification problem in which the loss incurred when an input vector from class $\mathcal{C}_k$ is classified as belonging to class $\mathcal{C}_j$ is given by the loss matrix L_{kj}, and for which the loss incurred in selecting the reject option is λ. Find the decision criterion that will give the minimum expected loss. Verify that this reduces to the reject criterion discussed in Section 1.5.3 when the loss matrix is given by $L_{kj} = 1 - I_{kj}$. What is the relationship between λ and the rejection threshold θ?

Proof. The decision criterion reduces to choosing the minimum between the loss of choosing the best class and the reject loss λ. Therefore, if

$$\alpha = \operatorname*{argmin}_j \sum_k L_{kj} p(\mathbf{x}|\mathcal{C}_k)$$

we choose the class α if the above quantity is less than λ and use the reject option otherwise. If the loss matrix is given by $L_{kj} = 1 - I_{kj}$, then

$$\alpha = \operatorname*{argmin}_j \{1 - p(\mathcal{C}_j|\mathbf{x})\}$$

which makes $\mathcal{C}_\alpha$ the class with the highest posterior probability. Therefore the criterion reduces to the one discussed in Section 1.5.3. If the highest posterior probability is smaller than $1 - \lambda$, then we use the reject option. This is equivalent with using $\theta = 1 - \lambda$ in Section 1.5.3. □

Exercise 1.25 ⋆ CALCULUS OF VARIATIONS

Consider the generalization of the squared loss function (1.87) for a single target variable t to the case of multiple target variables described by the vector $\mathbf{t}$ given by

$$\mathbb{E}[L(\mathbf{t}, \mathbf{y}(\mathbf{x})] = \iint ||\mathbf{y}(\mathbf{x}) - \mathbf{t}||^2 p(\mathbf{x}, \mathbf{t}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{t} \tag{1.151}$$

Using the calculus of the variations, show that the function $\mathbf{y}(\mathbf{x})$ for which this expected loss is minimized is given by $\mathbf{y}(\mathbf{x}) = \mathbb{E}_{\mathbf{t}}[\mathbf{t}|\mathbf{x}]$. Show that this result reduces to (1.89) for the case of a single target variable t.

Exercise 1.26 ⋆ TODO

By expansion of the square in (1.151), derive a result analogous to (1.90), and hence show that the function $\mathbf{y}(\mathbf{x})$ that minimizes the expected square loss for the case of a vector $\mathbf{t}$ of target variables is again given by the conditional expectation of $\mathbf{t}$.

Exercise 1.27 ⋆⋆ TODO

Consider the expected loss for regression problems under the L_q loss function given by (1.91). Write down the condition that $y(\mathbf{x})$ must satisfy in order to minimize $\mathbb{E}[L_q]$. Show that, for $q = 1$, this solution represents the conditional median, i.e., the function $y(\mathbf{x})$ such that the probability mass for $t < y(\mathbf{x})$ is the same for $t \geq y(\mathbf{x})$. Also show that the minimum expected L_q loss for $q \to 0$ is given by the conditional mode, i.e., by the function $y(\mathbf{x})$ equal to the value t that maximizes $p(t|\mathbf{x})$ for each $\mathbf{x}$.

Proof. □

Exercise 1.28 ⋆

In Section 1.6, we introduced the idea of entropy $h(x)$ as the information gained on observing the value of a random variable x having distribution $p(x)$. We saw that, for independent variables x and y for which $p(x, y) = p(x)p(y)$, the entropy functions are additive, so that $h(x, y) = h(x) + h(y)$. In this exercise, we derive that the relation between h and p in the form of a function $h(p)$. First show that $h(p^2) = 2h(p)$, and hence by induction that $h(p^n) = nh(p)$ where n is a positive integer. Hence show that $h(p^{n/m}) = n/mh(p)$ where m is also a positive integer. This implies that $h(p^x) = xh(p)$ where x is a positive rational number, and hence by continuity when it is a positive real number. Finally, show that this implies $h(p)$ must take the form $h(p) \propto \ln p$.

Proof. For independent variables x and y we have that:

$$h(x,y) = -\log_2 p(x,y) = -\log_2 p(x)p(y) = -\log_2 p(x) - \log_2 p(y) = h(x) + h(y)$$

Next, we show that:

$$h(p^2) = -\log_2 p^2 = -2\log_2 p = 2h(p)$$

and more generally for a positive integer n:

$$h(p^n) = -\log_2 p^n = -n\log_2 p = nh(p)$$

This can be extended to rational number by letting $n, m \in \mathbb{N}$ and showing that:

$$h(p^{n/m}) = -\log_2 p^{n/m} = -\frac{n}{m}\log_2 p = \frac{n}{m}h(p)$$

Finally, since

$$h(p) = -\log_2 p = -\frac{1}{\ln 2}\ln p$$

we have that $h(p) \propto \ln p$. □

Exercise 1.29

Consider an M-state discrete random variable x, and use Jensen's inequality in the form (1.115) to show that the entropy of the distribution $p(x)$ satisfies $H[x] \leq \ln M$.

Proof. The entropy of the distribution $p(x)$ is given by:

$$H[x] = -\sum_{i=1}^{M} p(x_i)\ln p(x_i)$$

We apply Jensen's inequality with $\lambda_i = p(x_i)$ and the convex function $f(x) = \ln(x)$ to obtain:

$$H[x] \leq -\ln\left(\sum_{i=1}^{M} p(x)^2\right) \tag{1.29.1}$$

One can prove by using Lagrange multipliers that

$$\sum_{i=1}^{M} p(x)^2 \leq \frac{1}{M}$$

Therefore, by substituting into (1.29.1) and using the fact that $\ln x$ is strictly increasing on $(0,\infty)$, we have that

$$H[x] \leq \ln M$$

□

Exercise 1.30 ⋆⋆

Evaluate the Kullback-Leibler divergence (1.113) between two Gaussians $p(x) = \mathcal{N}(x|\mu, \sigma^2)$ and $q(x) = \mathcal{N}(x|m, s^2)$.

Proof. The Kullback-Leibler divergence is given by

$$\mathrm{KL}(p||q) = -\int p(x) \ln \left\{ \frac{q(x)}{p(x)} \right\} \mathrm{d}x \tag{1.113}$$

We start by splitting the integral into:

$$\mathrm{KL}(p||q) = -\int p(x) \ln q(x)\, \mathrm{d}x + \int p(x) \ln p(x)\, \mathrm{d}x$$

The negation of the second term will be equal to the entropy of the Gaussian, that is:

$$H_p[x] = \frac{1}{2}\{1 + \ln\big(2\pi\sigma^2\big)\} \tag{1.110}$$

We have that

$$\ln q(x) = \ln \mathcal{N}(x|m, s^2) = \frac{1}{2} \ln\big(2\pi s^2\big) - \frac{(x-m)^2}{s^2}$$

so by using the fact that the Gaussian is normalized and by noticing the expected values, the KL divergence becomes:

$$\begin{aligned}
\mathrm{KL}(p||q) &= \frac{1}{s^2} \int p(x) x^2\, \mathrm{d}x - \frac{2m}{s^2} \int p(x) x\, \mathrm{d}x + \left\{ \frac{1}{2} \ln\big(2\pi s^2\big) + \frac{m^2}{s^2} \right\} \int p(x)\, \mathrm{d}x - H_p[x] \\
&= \frac{1}{s^2} \mathbb{E}[x^2] - \frac{2m}{s^2} E[x] + \frac{m^2}{s^2} + \ln \frac{s}{\sigma} + \frac{1}{2} \\
&= \frac{1}{2} + \ln \frac{s}{\sigma} + \frac{\sigma^2 + (\mu - m)^2}{s^2}
\end{aligned}$$

□

Exercise 1.31 ⋆⋆

Consider two variables $\mathbf{x}$ and $\mathbf{y}$ having joint distribution $p(\mathbf{x}, \mathbf{y})$. Show that the differential entropy of this pair of variables satisfies

$$H[\mathbf{x}, \mathbf{y}] \leq H[\mathbf{x}] + H[\mathbf{y}] \tag{1.152}$$

with equality if, and only if $\mathbf{x}$ and $\mathbf{y}$ are statistically independent.

Proof. The differential entropy of two variables $\mathbf{x}$ and $\mathbf{y}$ is given by

$$H[\mathbf{x}, \mathbf{y}] = H[\mathbf{y}|\mathbf{x}] + H[\mathbf{x}] \tag{1.112}$$

so (1.152) becomes equivalent with

$$H[\mathbf{y}|\mathbf{x}] - H[\mathbf{y}] \leq 0 \tag{1.31.1}$$

which we're going to prove now.

We start by rewriting the entropy $H[\mathbf{y}]$ as

$$H[\mathbf{y}] = -\int p(\mathbf{y}) \ln p(\mathbf{y}) \, \mathrm{d}\mathbf{y} = -\iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{y}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y}$$

Therefore, since the differential entropy is given by

$$H[\mathbf{y}|\mathbf{x}] = \iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{y}|\mathbf{x}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y} \tag{1.111}$$

we have that

$$\begin{aligned} H[\mathbf{y}|\mathbf{x}] - H[\mathbf{y}] &= -\iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{y}|\mathbf{x}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y} + \iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{y}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y} \\ &= \iint p(\mathbf{x}, \mathbf{y}) \ln \left\{ \frac{p(\mathbf{y})}{p(\mathbf{y}|\mathbf{x})} \right\} \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y} \end{aligned}$$

By using the inequality $\ln \alpha \leq \alpha - 1$, for all $\alpha > 0$, we obtain:

$$\begin{aligned} H[\mathbf{y}|\mathbf{x}] - H[\mathbf{y}] &\leq \iint p(\mathbf{x}) p(\mathbf{y}|\mathbf{x}) \left\{ \frac{p(\mathbf{y})}{p(\mathbf{y}|\mathbf{x})} - 1 \right\} \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y} \\ &\leq \iint p(\mathbf{x}) p(\mathbf{y}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y} - \iint p(\mathbf{x}) p(\mathbf{y}) \, \mathrm{d}\mathbf{x}\mathbf{y} \\ &\leq \iint p(\mathbf{x}) p(\mathbf{y}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y} - 1 \\ &\leq 0 \end{aligned}$$

which proves (1.31.1), respectively (1.152). □

Exercise 1.32 ⋆

Consider a vector $\mathbf{x}$ of continuous variables with distribution $p(\mathbf{x})$ and corresponding entropy $H[\mathbf{x}]$. Suppose that we make a nonsingular linear transformation of $\mathbf{x}$ to obtain a new variable $\mathbf{y} = \mathbf{A}\mathbf{x}$. Show that the coresponding entropy is given by $H[\mathbf{y}] = H[\mathbf{x}] + \ln |\mathbf{A}|$ where $|\mathbf{A}|$ denotes the determinant of $\mathbf{A}$.

Proof. By generalizing (1.27) for the multivariate case, we have that:

$$p_{\mathbf{y}}(\mathbf{y}) = p_{\mathbf{x}}(\mathbf{x}) \left| \frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right| = p_{\mathbf{x}}(\mathbf{x}) \left| \frac{\partial \mathbf{A}^{-1}\mathbf{y}}{\partial \mathbf{y}} \right| = p_{\mathbf{x}}(\mathbf{x}) |\mathbf{A}^{-1}| = p_{\mathbf{x}}(\mathbf{x}) |\mathbf{A}|^{-1}$$

where $J = \left| \frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right| = |\mathbf{A}|^{-1}$ is the Jacobian determinant.

Now, the entropy of $\mathbf{y}$ is given by:

$$\begin{aligned} H[\mathbf{y}] &= -\int p_{\mathbf{y}}(\mathbf{y}) \ln p_{\mathbf{y}}(\mathbf{y}) \, \mathrm{d}\mathbf{y} = -\int \frac{p_{\mathbf{x}}(\mathbf{x})}{|\mathbf{A}|} \ln \frac{p_{\mathbf{x}}(\mathbf{x})}{|\mathbf{A}|} \left| \frac{\mathrm{d}\mathbf{y}}{\mathrm{d}\mathbf{x}} \right| \mathrm{d}\mathbf{x} = -\int p_{\mathbf{x}}(\mathbf{x}) \ln \frac{p_{\mathbf{x}}(\mathbf{x})}{|\mathbf{A}|} \, \mathrm{d}\mathbf{x} \\ &= -\int p_{\mathbf{x}}(\mathbf{x}) \ln p_{\mathbf{x}} \, \mathrm{d}\mathbf{x} + \ln |\mathbf{A}| \int p_{\mathbf{x}}(\mathbf{x}) \, \mathrm{d}\mathbf{x} \\ &= H[\mathbf{y}] + \ln |\mathbf{A}| \end{aligned}$$

□

Exercise 1.33 ⋆⋆

Suppose that the conditional entropy $H[y|x]$ between two discrete random variables x and y is zero. Show that, for all values of x such that $p(x) > 0$, the variable y must be a function of x, in other words for each x there is only one value of y such that $p(y|x) \neq 0$.

Proof. Assuming x, y have N respectively M outcomes, we can rewrite the conditional entropy as:

$$H[y|x] = -\sum_{i}^{N}\sum_{j}^{M} p(x_i, y_j) \ln p(y_j|x_i) = -\sum_{i}^{N} p(x_i) \sum_{j}^{M} p(y_j|x_i) \ln p(y_j|x_i)$$

Since all the sum terms have the same sign, the entropy is 0 if each term is 0. Therefore, the entropy is 0 if for each $p(x_i) > 0$, the inner sum terms are 0. This happens only for $p(y_j|x_i) = 0$ or $\ln p(y_j|x_i) = 0$, which means that $p(y_j|x_i) \in \{0, 1\}$. Since $\sum_{j=1}^{M} p(y_j|x_i) = 1$, we have that for each x_i there is an unique y_j such that $p(y_j|x_i) = 1$, which proves our hypothesis. □

Exercise 1.34 ⋆⋆ CALCULUS OF VARIATIONS

Use the calculus of variations to show that the stationary point of the functional (1.108) is given by (1.108). Then use the constraints (1.105), (1.106) and (1.107) to eliminate de Lagrange multipliers and hence show that the maximum entropy solution is given by the Gaussian (1.109).

Exercise 1.35 ⋆

Use the results (1.106) and (1.107) to show that the entropy of the univariate Gaussian (1.109) is given by (1.110).

Proof. The entropy of the univariate Gaussian is given by:

$$\begin{aligned} H[x] =& -\int \mathcal{N}(x|\mu,\sigma^2) \ln \mathcal{N}(x|\mu,\sigma^2)\,\mathrm{d}x = -\frac{1}{2}\ln\big(2\pi\sigma^2\big)\int \mathcal{N}(x|\mu,\sigma^2)\,\mathrm{d}x + \int \mathcal{N}(x|\mu,\sigma^2)\frac{(x-\mu)^2}{\sigma^2}\,\mathrm{d}x \\ =& \left\{ -\frac{1}{2}\ln\big(2\pi\sigma^2\big) + \frac{\mu^2}{\sigma^2} \right\} \int \mathcal{N}(x|\mu,\sigma^2)\,\mathrm{d}x + \frac{1}{2\sigma^2}\int \mathcal{N}(x|\mu,\sigma^2)x^2\,\mathrm{d}x - \frac{2\mu}{2\sigma^2}\int \mathcal{N}(x|\mu,\sigma^2)x\,\mathrm{d}x \end{aligned}$$

By using the fact that the Gaussian is normalized and by noticing the expression of the expected value, we have that

$$H[x] = -\frac{1}{2}\ln\big(2\pi\sigma^2\big) + \frac{\mu^2}{2\sigma^2} + \frac{1}{2\sigma^2}\mathbb{E}[x^2] - \frac{2\mu}{2\sigma^2}\mathbb{E}[x] = \frac{1}{2}\left\{1 - \frac{1}{2}\ln\big(2\pi\sigma^2\big)\right\} \tag{1.110}$$

□

Exercise 1.36 ⋆

A strictly convex function is defined as one for which every chord lies above the function. Show that this is equivalent to the condition that the second derivative of the function be positive.

Proof. Suppose that f is a twice differentiable function. By summing the Taylor expansions of $f(x+h)$ and $f(x-h)$, one can show that

$$f''(x) = \lim_{h \to 0} \frac{f(x+h) + f(x-h) - 2f(x)}{h^2}$$

Therefore, we have that

$$\begin{aligned} f''(x) > 0 &\iff f(x+h) + f(x-h) - 2f(x) > 0 \\ &\iff \frac{1}{2}f(x+h) + \frac{1}{2}f(x-h) - f(x) > 0 \end{aligned}$$

If f is strictly convex, we can apply (1.114) in a strict form to obtain

$$\frac{1}{2}f(x+h) + \frac{1}{2}f(x-h) - f(x) > f\left(\frac{1}{2}(x+h) + \frac{1}{2}(x-h)\right) - f(x) = 0$$

Therefore, the second derivative of a strictly convex function is positive. □

Exercise 1.37 ⋆

Using the definition (1.111) together with the product rule of probability, prove the result (1.112).

Proof. Using the product rule of probability, one could rewrite the entropy of $\mathbf{x}$ as:

$$H[\mathbf{x}] = -\int p(\mathbf{x}) \ln p(\mathbf{x}) \, \mathrm{d}\mathbf{x} = -\iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y}$$

Now, by summing this with (1.111) we see that:

$$\begin{aligned} H[\mathbf{y}|\mathbf{x}] + H[\mathbf{x}] &= -\iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{y}|\mathbf{x}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y} - \iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y} \\ &= -\iint p(\mathbf{x}, \mathbf{y}) \ln\{p(\mathbf{y}|\mathbf{x}) p(\mathbf{x})\} \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y} \\ &= -\iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}, \mathbf{y}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{y} \\ &= H[\mathbf{x}, \mathbf{y}] \end{aligned} \tag{1.112}$$

□

Exercise 1.38 ⋆⋆

Using proof by induction, show that the inequality (1.114) for convex functions implies the result (1.115).

Proof. We'll prove Jensen's inequality by induction, i.e. if we have N points $x_1, \dots x_n$, f is a convex function and $\lambda_i \geq 0$, $\sum_{i=1}^{N} \lambda_i = 1$, then

$$f\left(\sum_{i=1}^{N} \lambda_i x_i\right) \leq \sum_{i=1}^{N} \lambda_i f(x_i) \tag{1.115}$$

We consider the base case of the induction to be given by

$$f(\lambda a + (1-\lambda)b) \leq \lambda f(a) + (1-\lambda) f(b) \tag{1.114}$$

Now, we assume that Jensen's inequality is true for a set of N points and want to prove that it's also true for $N+1$ points. Since $\sum_{i=1}^{N} \lambda_i = 1$, there exists at least one $\lambda_i \leq 1$. We can assume without loss of generality that this is λ_1. Therefore, we have that

$$f\left(\sum_{i=1}^{N+1} \lambda_i x_i\right) = f\left(\lambda_1 x_1 + (1-\lambda_1) \sum_{i=2}^{N+1} \frac{\lambda_i}{1-\lambda_1} x_i\right)$$

Since λ_1 and $1-\lambda_1$ are both nonnegative and sum to 1, we can apply (1.115) to the right-hand side of the equality to obtain:

$$\begin{aligned} f\left(\sum_{i=1}^{N+1} \lambda_i x_i\right) &\leq \lambda_1 f(x_1) + (1-\lambda_1) \sum_{i=2}^{N+1} \frac{\lambda_i}{1-\lambda_1} f(x_i) \\ &\leq \lambda_1 f(x_1) + \sum_{i=2}^{N+1} f(x_i) \\ &\leq \sum_{i=1}^{N+1} \lambda_i f(x_i) \end{aligned} \tag{1.115}$$

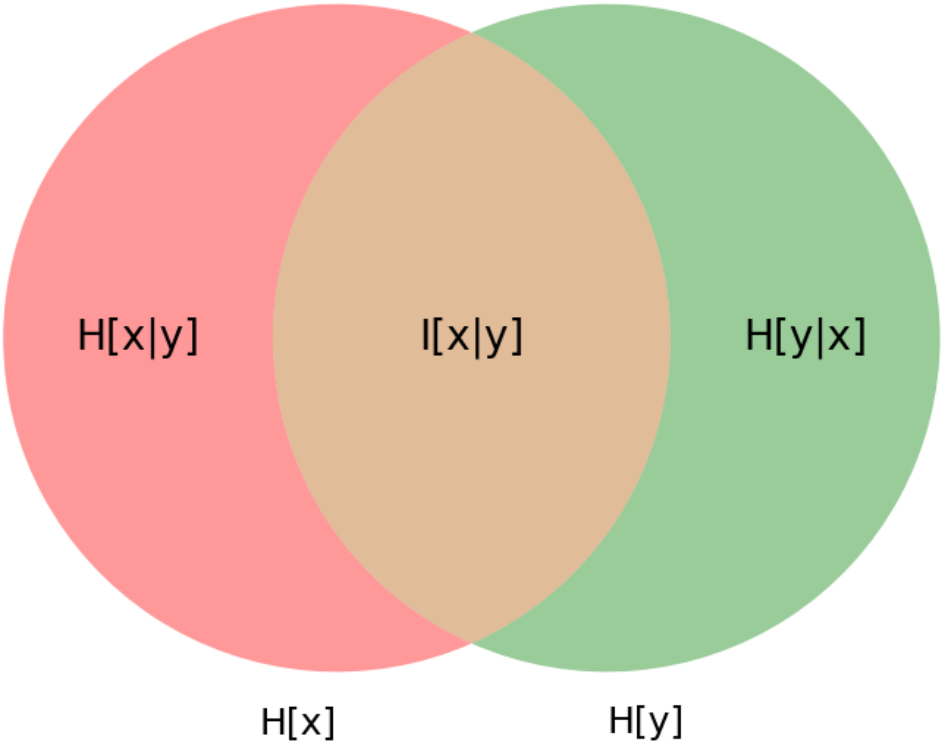
Therefore, we proved Jensen's inequality by induction. □

Exercise 1.39 ⋆⋆⋆

Consider two binary variables x and y having the joint distribution given in Table 1.3. Evaluate the following quantities:

(a) $H[x]$ (c) $H[y|x]$ (e) $H[x, y]$

(b) $H[y]$ (d) $H[x|y]$ (f) $I[x, y]$



Exercise 1.39 Diagram

Draw a diagram to show the relationship between these various quantities.

		y 0	y 1
x	0	1/3	1/3
x	1	0	1/3

Table 1.3 The joint distribution $p(x, y)$ used in Exercise 1.39.

Proof. Through straightforward computations using the discrete formula for the entropy, we have

(a) $H[x] = -2/3 \ln 2 + \ln 3$ (c) $H[x|y] = 2/3 \ln 2$ (e) $H[x, y] = \ln 3$

(b) $H[y] = -2/3 \ln 2 + \ln 3$ (d) $H[y|x] = 2/3 \ln 2$ (f) $I[x, y] = -4/3 \ln 2 + \ln 3$

The diagram shows the relationship between the entropies. Note that the joint entropy $H[x, y]$ occupies all three colored areas.

□

Exercise 1.40 ⋆

By applying Jensen's inequality (1.115) with $f(x) = \ln x$, show that the arithmetic mean of a set of real numbers is never less than their geometric mean.

Proof. Let N be the cardinality of the considered set of real numbers. By considering $f(x) = \ln x$ (which is convex) and $\lambda_i = 1/N$, we use Jensen's inequality to obtain:

$$\ln\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) \leq \frac{1}{N}\sum_{i=1}^{N} \ln x_i = \frac{1}{N}\ln\left(\prod_{i=1}^{N} x_i\right) = \ln\left\{\left(\prod_{i=1}^{N} x_i\right)^{1/N}\right\}$$

Since $\ln x$ is increasing, the above inequality is equivalent with:

$$\frac{1}{N}\sum_{i=1}^{N} x_i \leq \left(\prod_{i=1}^{n} x_i\right)^{1/N}$$

which proves that the arithmetic mean of a set of real numbers is never less than their geometric mean. □

Exercise 1.41 ⋆

Using the sum and product rules of probability, show that the mutual information $I(\mathbf{x}, \mathbf{y})$ satisfies the relation (1.121).

Proof. The mutual information between the variables $\mathbf{x}$ and $\mathbf{y}$ is given by:

$$I[\mathbf{x}, \mathbf{y}] = -\iint p(\mathbf{x}, \mathbf{y}) \ln\left(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})}\right) \mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} \tag{1.120}$$

We split the integral and use the product and sum rules of probability to obtain the desired result:

$$\begin{aligned} I[\mathbf{x}, \mathbf{y}] &= -\iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} + \iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}|\mathbf{y})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} \\ &= -\int p(\mathbf{x}) \ln p(\mathbf{x})\,\mathrm{d}\mathbf{x} + \iint p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}|\mathbf{y})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} \\ &= H[\mathbf{x}] - H[\mathbf{x}|\mathbf{y}] \end{aligned} \tag{1.121}$$

Analogously, one could easily show that also $I[\mathbf{x}, \mathbf{y}] = H[\mathbf{y}] - H[\mathbf{y}|\mathbf{x}]$ □

Chapter 2

Probability Distributions

Exercise 2.1 ⋆

Verify that the Bernoulli distribution (2.2) satisfies the following properties

$$\sum_{x=0}^{1} p(x|\mu) = 1 \tag{2.257}$$

$$\mathbb{E}[x] = \mu \tag{2.258}$$

$$\text{var}[x] = \mu(1-\mu) \tag{2.259}$$

Show that the entropy $H[x]$ of a Bernoulli distributed random binary variable x is given by

$$H[x] = -\mu \ln \mu - (1-\mu)\ln(1-\mu) \tag{2.260}$$

Proof. The Bernoulli distribution is given by

$$\text{Bern}(x|\mu) = \mu^x (1-\mu)^{1-x} \tag{2.2}$$

The properties are easily verified:

$$\sum_{x=0}^{1} p(x|\mu) = p(x=0|\mu) + p(x=1|\mu) = \mu^0(1-\mu)^1 + \mu^1(1-\mu)^0 = 1 \tag{2.257}$$

$$\mathbb{E}[x] = \sum_{x=0}^{1} x p(x|\mu) = 0 \cdot p(x=0|\mu) + 1 \cdot p(x=1|\mu) = \mu \tag{2.258}$$

$$\text{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sum_{x=0}^{1} x^2 p(x|\mu) - \mu^2 = 0^2 \cdot p(x=0|\mu) + 1^2 \cdot p(x=1|\mu) - \mu^2 = \mu(1-\mu) \tag{2.259}$$

The entropy is also straightforward to derive:

$$\begin{aligned} H[x] = -\sum_{x=0}^{1} p(x|\mu) \ln p(x|\mu) &= -p(x=0|\mu)\ln p(x=0|\mu) - p(x=1|\mu)\ln p(x=1|\mu) \\ &= -\mu \ln \mu - (1-\mu)\ln(1-\mu) \end{aligned}$$

□

Exercise 2.2 ⋆⋆

The form of the Bernoulli distribution given by (2.2) is not symmetric between the two values of x. In some situations, it will be more convenient to use an equivalent formulation for which $x \in \{-1, 1\}$, in which case the distribution can be written

$$p(x|\mu) = \left(\frac{1-\mu}{2}\right)^{(1-x)/2} \left(\frac{1+\mu}{2}\right)^{(1+x)/2} \tag{2.261}$$

where $\mu \in [-1, 1]$. Show that the distribution (2.261) is normalized, and evaluate its mean, variance and entropy.

Proof. The distribution is normalized since

$$\sum_x p(x|\mu) = p(x = -1|\mu) + p(x = 1|\mu) = \frac{1-\mu}{2} + \frac{1+\mu}{2} = 1$$

The other properties are also easily derived:

$$\mathbb{E}[x] = \sum_x x p(x|\mu) = p(x = 1|\mu) - p(x = -1|\mu) = \frac{1+\mu}{2} - \frac{1-\mu}{2} = \mu$$

$$\begin{aligned} \mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sum_x x^2 p(x|\mu) - \mu^2 &= p(x = -1|\mu) + p(x = 1|\mu) - \mu^2 \\ &= \frac{1+\mu}{2} + \frac{1-\mu}{2} - \mu^2 = (1-\mu)(1+\mu) \end{aligned}$$

$$\begin{aligned} H[x] = -\sum_x p(x|\mu) \ln p(x|\mu) &= -p(x = -1|\mu) \ln p(x = -1|\mu) - p(x = 1|\mu) \ln p(x = 1|\mu) \\ &= -\frac{1-\mu}{2} \ln\left(\frac{1-\mu}{2}\right) - \frac{1+\mu}{2} \ln\left(\frac{1+\mu}{2}\right) \end{aligned}$$

□

Exercise 2.3 ⋆⋆

In this exercise, we prove that the binomial distribution (2.9) is normalized. First use the definition (2.10) of the number of combinations of m identical objects chosen from a total of N to show that

$$\binom{N}{m} + \binom{N}{m-1} = \binom{N+1}{m} \tag{2.262}$$

Use this result to prove by induction the following result

$$(1+x)^N = \sum_{m=0}^{N} \binom{N}{m} x^m \tag{2.263}$$

which is known as the *binomial theorem*, and which is valid for all real values of x. Finally, show that the binomial distribution is normalized, so that

$$\sum_{m=0}^{N} \binom{N}{m} \mu^m (1-\mu)^{N-m} = 1 \tag{2.264}$$

which can be done by first pulling out a factor $(1-\mu)^N$ out of the summation and then making use of the binomial theorem.

Proof. The binomial distribution is given by

$$\mathrm{Bin}(m|N,\mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m} \tag{2.9}$$

By using (2.10), we prove (2.262)

$$\begin{aligned}
\binom{N}{m} + \binom{N}{m-1} &= \frac{N!}{(N-m)!m!} + \frac{N!}{(N-m+1)!(m-1)!} \\
&= \frac{(N-m+1)N!}{(N-m+1)!m!} + \frac{mN!}{(N-m+1)!m!} \\
&= \frac{(N+1)!}{(N-m+1)!m!} \\
&= \binom{N+1}{m}
\end{aligned} \tag{2.262}$$

We aim to prove (2.263) by induction. The base case for $N = 1$ is obviously true since

$$1 + x = \binom{1}{0} + \binom{1}{1} x$$

Now, suppose that the case for $N = k \in \mathbb{N}^*$ is true, i.e.

$$(1+x)^k = \sum_{m=0}^{k} \binom{k}{m} x^m$$

By using this and (2.262), we show that

$$\begin{aligned}
(1+x)^{k+1} &= (1+x) \sum_{m=0}^{k} \binom{k}{m} x^m \\
&= \sum_{m=0}^{k} \binom{k}{m} x^m + \sum_{m=0}^{k} \binom{k}{m} x^{m+1} \\
&= 1 + \sum_{m=1}^{k} \binom{k}{m} x^m + \sum_{m=1}^{k+1} \binom{k}{m-1} x^m \\
&= \binom{k+1}{0} + \binom{k+1}{k+1} x^{k+1} + \sum_{m=1}^{k} \left\{ \binom{k}{m} + \binom{k}{m-1} \right\} x^m \\
&= \sum_{m=0}^{k+1} \binom{k+1}{m} x^m
\end{aligned}$$

which by induction proves that (2.263) is indeed true.

Finally, we use this result to show that the Binomial distribution is normalized:

$$\begin{aligned}\sum_{m=0}^{N}\binom{N}{m}\mu^m(1-\mu)^{N-m} &= (1-\mu)^N\sum_{m=0}^{N}\binom{N}{m}\left(\frac{\mu}{1-\mu}\right)^m \\ &= (1-\mu)^N\left(1+\frac{\mu}{1-\mu}\right)^N \\ &= 1 \end{aligned} \tag{2.264}$$

□

Exercise 2.4 ⋆⋆

Show that the mean of the binomial distribution is given by (2.11). To do this, differentiate both sides of the normalization condition (2.264) with respect to μ and then rearrange to obtain an expression for the mean of m. Similarly, by differentiating (2.264) twice with respect to μ and making use of the result (2.11) for the mean of the binomial distribution prove the result (2.12) for the variance of the binomial.

Proof. We start by differentiating both sides of (2.264) with respect to μ:

$$\frac{\partial}{\partial\mu}\sum_{m=0}^{N}\binom{N}{m}\mu^m(1-\mu)^{N-m} = 0$$

$$\sum_{m=0}^{N}\binom{N}{m}\mu^m(1-\mu)^{N-m}\left(\frac{m}{\mu}+\frac{m-N}{1-\mu}\right) = 0$$

$$\left(\frac{1}{\mu}+\frac{1}{1-\mu}\right)\sum_{m=0}^{N}m\binom{N}{m}\mu^m(1-\mu)^{N-m} - \frac{N}{1-\mu}\sum_{m=0}^{N}\binom{N}{m}\mu^m(1-\mu)^{N-m} = 0$$

We recognize the expression of the binomial distribution and use the fact that it is normalized, to obtain:

$$\left(\frac{1}{\mu}+\frac{1}{1-\mu}\right)\sum_{m=0}^{N}m\mathrm{Bin}(m|N,\mu) = \frac{N}{1-\mu}\sum_{m=0}^{N}\mathrm{Bin}(m|N,\mu)$$

$$\left(\frac{1-\mu}{\mu}+1\right)\mathbb{E}[m] = N$$

which directly gives us the desired result, that is

$$\mathbb{E}[m] = \sum_{m=0}^{N}m\mathrm{Bin}(m|N,\mu) = N\mu \tag{2.11}$$

To derive the variance, we differentiate twice both sides of (2.264), so

$$\frac{\partial^2}{\partial \mu^2} \sum_{m=0}^{N} \binom{N}{m} \mu^m (1-\mu)^{N-m} = 0$$

$$\frac{1}{\mu^2(1-\mu)^2} \sum_{m=0}^{N} \mathrm{Bin}(m|N,\mu)\{m^2 + m(2\mu - 2N\mu - 1) + (N-1)N\mu^2\} = 0$$

$$\sum_{m=0}^{N} \mathrm{Bin}(m|N,\mu)(m-N\mu)^2 + (2\mu-1)\sum_{m=0}^{N} m\mathrm{Bin}(m|N,\mu) - N\mu^2 \sum_{m=0}^{N} \mathrm{Bin}(m|N,\mu) = 0$$

$$\mathrm{var}[m] + (2\mu - 1)\mathbb{E}[m] - N\mu^2 = 0$$

which gives us the desired result, i.e.

$$\mathrm{var}[m] \equiv \sum_{m=0}^{N} (m - \mathbb{E}[m])^2 \mathrm{Bin}(m|N,\mu) = N\mu(1-\mu) \tag{2.12}$$

□

Exercise 2.5 ⋆⋆

In this exercise, we prove that the beta distribution, given by (2.13), is correctly normalized, so that (2.14) holds. This is equivalent to showing that

$$\int_0^1 \mu^{a-1}(1-\mu)^{b-1}\,\mathrm{d}\mu = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{2.265}$$

From the definition (1.141) of the gamma function, we have

$$\Gamma(a)\Gamma(b) = \int_0^\infty \exp(-x)x^{a-1}\,\mathrm{d}x + \int_0^\infty \exp(-y)y^{b-1}\,\mathrm{d}y \tag{2.266}$$

Use this expression to prove (2.265) as follows. First bring the integral over y inside the integrand of the integral over x, next make the change of variable $t = y + x$, where x is fixed, then interchange the order of the x and t integrations, and finally make the change of variable $x = t\mu$ where t is fixed.

Proof. The problem is easily solved by following the provided steps. By bringing the integral over y inside the integrand of the integral over x we obtain that

$$\Gamma(a)\Gamma(b) = \int_0^\infty \int_0^\infty \exp\{-(x+y)\}x^{a-1}y^{b-1}\,\mathrm{d}y\,\mathrm{d}x$$

We know use the change of variable $t = y + x$ with x fixed to get

$$\Gamma(a)\Gamma(b) = \int_0^\infty \int_0^\infty \exp(-t)x^{a-1}(x-t)^{b-1}\,\mathrm{d}t\,\mathrm{d}x$$

Interchanging the order of integrations yields

$$\Gamma(a)\Gamma(b) = \int_0^\infty \int_0^\infty \exp(-t)x^{a-1}(x-t)^{b-1}\,\mathrm{d}x\,\mathrm{d}t$$

which by making the change of variable $x = t\mu$ with t fixed becomes

$$\Gamma(a)\Gamma(b) = \int_0^\infty \int_0^\infty \exp(-t)(t\mu)^{a-1}(t\mu-t)^{b-1}t\,\mathrm{d}\mu\,\mathrm{d}t$$

By separating the t terms from the first integral, we have that

$$\Gamma(a)\Gamma(b) = \int_0^\infty \exp(-t)t^{a+b-1}\,\mathrm{d}t \int_0^\infty \mu^{a-1}(1-\mu)^{b-1}\,\mathrm{d}\mu$$

Finally, we notice that the first integral is equal to $\Gamma(a+b)$ and by noting the fact that μ is a probability, so its range is $[0,1]$, we obtain the desired result:

$$\int_0^1 \mu^{a-1}(1-\mu)^{b-1}\,\mathrm{d}\mu = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{2.265}$$

□

Exercise 2.6 ⋆

Make use of the result (2.265) to show that the mean, variance, and mode of the beta distribution (2.13) are given respectively by

$$\mathbb{E}[\mu] = \frac{a}{a+b} \tag{2.267}$$

$$\mathrm{var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)} \tag{2.268}$$

$$\mathrm{mode}[\mu] = \frac{a-1}{a+b-2} \tag{2.269}$$

Proof. The beta distribution is given by

$$\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \tag{2.13}$$

By using (2.265) and the fact that $\Gamma(x+1) = x\Gamma(x)$, we obtain the mean of the Beta distribution:

$$\mathbb{E}[\mu] = \int_0^1 \mu\mathrm{Beta}(\mu|a,b)\,\mathrm{d}\mu = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^1 \mu^a(1-\mu)^{b-1}\,\mathrm{d}\mu = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\cdot\frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} = \frac{a}{a+b} \tag{2.267}$$

From this result, we can also easily get the variance:

$$\begin{aligned}
\text{var}[\mu] &= \int_0^1 \left(\mu - \frac{a}{a+b}\right)^2 \text{Beta}(\mu|a,b)\,\mathrm{d}\mu \\
&= \int_0^1 \mu^2 \text{Beta}(\mu|a,b)\,\mathrm{d}\mu - \frac{2a}{a+b}\int_0^1 \mu \text{Beta}(\mu|a,b)\,\mathrm{d}\mu + \frac{a^2}{(a+b)^2}\int_0^1 \text{Beta}(\mu|a,b)\,\mathrm{d}\mu \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)} - \frac{2a}{a+b} \cdot \frac{a}{a+b} + \frac{a^2}{(a+b)^2} \\
&= \frac{a(a+1)}{(a+b)(a+b+1)} - \frac{2a^2}{a+b} + \frac{a^2}{(a+b)^2} \\
&= \frac{ab}{(a+b)^2(a+b+1)}
\end{aligned} \tag{2.267}$$

Finally, the mode of the distribution is given by getting the value of μ for which the derivative of the distribution is 0,

$$\begin{aligned}
\frac{\partial}{\partial\mu}\text{Beta}(\mu|a,b) = 0 &\iff \frac{\partial}{\partial\mu}\mu^{a-1}(1-\mu)^{b-1} = 0 \\
&\iff (a-1)\mu^{a-2}(1-\mu)^{b-1} + (b-1)\mu^{a-1}(1-\mu)^{b-2} = 0 \\
&\iff \mu^{a-2}(1-\mu)^{b-2}\{(a-1)(1-\mu) + (b-1)\mu\} = 0 \\
&\iff (a-1)(1-\mu) + (b-1)\mu = 0 \\
&\iff \mu = \frac{a-1}{a+b-2}
\end{aligned}$$

so indeed

$$\text{mode}[\mu] = \frac{a-1}{a+b-2} \tag{2.268}$$

□

Exercise 2.7 ⋆⋆

Consider a binomial random variable x given by (2.9), with prior distribution for μ given by the beta distribution (2.13), and suppose we have observed m occurences of $x = 1$ and l occurences of $x = 0$. Show that the posterior mean value of μ lies between the prior mean and the maximum likelihood estimate for μ. To do this, show that the posterior mean can be written as λ times the prior mean plus $(1-\lambda)$ times the maximum likelihood estimate, where $0 \leq \lambda \leq 1$. This illustrates the concept of the posterior distribution being a compromise between the prior distribution and the maximum likelihood solution.

Proof. The prior mean is $\frac{a}{a+b}$, the posterior mean is $\frac{a+m}{a+m+b+l}$ and the maximum likelihood estimate is $\frac{m}{m+l}$. Suppose that our hypothesis is true, i.e. there exists a λ such that we can have

our equality and $0 \leq \lambda \leq 1$. Then we'd have that:

$$\begin{aligned}
\frac{a+m}{a+m+b+l} &= \frac{\lambda m}{m+l} + \frac{(1-\lambda)a}{a+b} \\
\frac{a+m}{a+m+b+l} - \frac{a}{a+b} &= \lambda\left(\frac{m}{m+l} + \frac{a}{a+b}\right) \\
\lambda &= \frac{bm-al}{(a+b)(a+m+b+l)} \cdot \frac{(a+b)(a+m)}{bm-al} \\
\lambda &= \frac{l+m}{a+m+b+l}
\end{aligned}$$

This λ obviously exists and $0 \leq \lambda \leq 1$, so our hypothesis is true and the posterior mean value of x lies between the prior mean and the maximum likelihood estimate for μ. □

Exercise 2.8 ⋆

Consider two variables x and y with joint distribution $p(x, y)$. Prove the following two results

$$\mathbb{E}[x] = \mathbb{E}_y[\mathbb{E}_x[x|y]] \tag{2.270}$$

$$\text{var}[x] = \mathbb{E}_y[\text{var}_x[x|y]] + \text{var}_y[\mathbb{E}_x[x|y]] \tag{2.271}$$

Here $\mathbb{E}_x[x|y]$ denotes the expectation of x under the conditional distribution $p(x|y)$, with a similar notation for the conditional variance.

Proof. The first is straightforward to derive:

$$\begin{aligned}
\mathbb{E}[x] = \iint xp(x,y)\,\mathrm{d}x\,\mathrm{d}y = \iint xp(x|y)p(y)\,\mathrm{d}x\,\mathrm{d}y &= \int \left(\int xp(x|y)\,\mathrm{d}x\right)p(y)\,\mathrm{d}y \\
&= \int \mathbb{E}_x[x|y]p(y)\,\mathrm{d}y = \mathbb{E}_y[\mathbb{E}_x[x|y]] \qquad (2.270)
\end{aligned}$$

However, proving (2.271) is slightly more complicated. We'll compute each term separately:

□

Exercise 2.10 ⋆⋆

Using the property $\Gamma(x+1) = x\Gamma(x)$ of the gamma function, derive the following results for the mean, variance, and covariance of the Dirichlet distribution given by (2.38)

$$\mathbb{E}[\mu_j] = \frac{\alpha_j}{\alpha_j} \tag{2.273}$$

$$\text{var}[\mu_j] = \frac{\alpha_j(\alpha_0 - \alpha_j)}{\alpha_0^2(\alpha_0+1)} \tag{2.274}$$

$$\text{cov}[\mu_j\mu_l] = -\frac{\alpha_j\alpha_l}{\alpha_0^2(\alpha_0+1)}, \qquad j \neq l \tag{2.275}$$

where α_0 is defined by (2.39).

Proof. The Dirichlet distribution is given by

$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_K)}\prod_{k=1}^{K}\mu_k^{\alpha_k-1} \tag{2.38}$$

Besides the property that $\Gamma(x+1) = x\Gamma(x)$, we'll be using the fact that the distribution is normalized, specifically that

$$\int\prod_{k=1}^{K}\mu_k^{\alpha_k-1}\,\mathrm{d}\boldsymbol{\mu} = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_K)}{\Gamma(\alpha_0)},$$

where α_0 is defined by (2.39).

The expected value is then given by

$$\begin{aligned}
\mathbb{E}[\mu_j] &= \int \mu_j \mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{\mu} \\
&= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_K)}\int \mu_1^{\alpha_1-1}\ldots\mu_j^{\alpha_j}\ldots\mu_K^{\alpha_K-1}\,\mathrm{d}\boldsymbol{\mu} \\
&= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_K)}\cdot\frac{\Gamma(\alpha_1)\ldots\Gamma(\alpha_j+1)\ldots\Gamma(\alpha_K)}{\Gamma(\alpha_0+1)} \\
&= \frac{\alpha_j}{\alpha_0}
\end{aligned} \tag{2.273}$$

This can now be used to derive the variance:

$$\begin{aligned}
\mathrm{var}[\mu_j] &= \int (\mu_j - \mathbb{E}[\mu_j])^2 \mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{\mu} \\
&= \int \left(\mu_j - \frac{\alpha_j}{\alpha_0}\right)^2 \mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{\mu} \\
&= \int \mu_j^2 \mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{\mu} - \frac{2\alpha_j}{\alpha_0}\int \mu_j \mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{\mu} + \frac{\alpha_j^2}{\alpha_0^2}\int \mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{\mu} \\
&= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_K)}\int \mu_1^{\alpha_1-1}\ldots\mu_j^{\alpha_j+1}\ldots\mu_K^{\alpha_K-1}\,\mathrm{d}\boldsymbol{\mu} - \frac{2\alpha_j}{\alpha_0}\mathbb{E}[\mu_j] + \frac{\alpha_j^2}{\alpha_0^2} \\
&= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_K)}\cdot\frac{\Gamma(\alpha_1)\ldots\Gamma(\alpha_j+2)\ldots\Gamma(\alpha_K)}{\Gamma(\alpha_0+2)} - \frac{\alpha_j^2}{\alpha_0^2} \\
&= \frac{\alpha_j(\alpha_j+1)}{\alpha_0(\alpha_0+1)} - \frac{\alpha_j^2}{\alpha_0^2} \\
&= \frac{\alpha_j(\alpha_0-\alpha_j)}{\alpha_0^2(\alpha_0+1)}
\end{aligned} \tag{2.275}$$

The covariance is given by

$$\mathrm{cov}[\mu_j\mu_l] = \mathbb{E}[\mu_j\mu_l] - \mathbb{E}[\mu_j]\mathbb{E}[\mu_l] = \mathbb{E}[\mu_j\mu_l] - \frac{\alpha_j\alpha_l}{\alpha_0^2}$$

By computing the expectation separately, we find that

$$\begin{aligned}\mathbb{E}[\mu_j\mu_l] &= \int \mu_j\mu_l \mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{\mu} \\ &= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_K)} \int \mu_1^{\alpha_1-1}\ldots\mu_j^{\alpha_j}\ldots\mu_l^{\alpha_l}\ldots\mu_K^{\alpha_K-1}\,\mathrm{d}\boldsymbol{\mu} \\ &= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_K)} \cdot \frac{\Gamma(\alpha_1)\ldots\Gamma(\alpha_j+1)\ldots\Gamma(\alpha_l+1)\ldots\Gamma(\alpha_K)}{\Gamma(\alpha_0+2)} \\ &= \frac{\alpha_j\alpha_l}{\alpha_0(\alpha_0+1)}\end{aligned}$$

Finally, the covariance becomes

$$\mathrm{cov}[\mu_j\mu_l] = \frac{\alpha_j\alpha_l}{\alpha_0(\alpha_0+1)} - \frac{\alpha_j\alpha_l}{\alpha_0^2} = -\frac{\alpha_j\alpha_l}{\alpha_0^2(\alpha_0+1)} \tag{2.275}$$

□

Exercise 2.11 ⋆

By expressing the expectation of $\ln\mu_j$ under the Dirichlet distribution (2.38) as a derivative with respect to α_j, show that

$$\mathbb{E}[\ln\mu_j] = \psi(\alpha_j) - \psi(\alpha_0) \tag{2.276}$$

where α_0 is given by (2.39) and

$$\psi(a) \equiv \frac{\mathrm{d}}{\mathrm{d}a}\ln\Gamma(a) \tag{2.277}$$

is the *digamma* function.

Proof. We start by taking the partial derivative of the Dirichlet distribution with respect to α_j:

$$\begin{aligned}\frac{\partial}{\partial\alpha_j}\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) &= \frac{\partial}{\partial\alpha_j}\left(\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_K)}\prod_{k=1}^{K}\mu_k^{\alpha_k-1}\right) \\ &= \left(\frac{\partial}{\partial\alpha_j}\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_K)}\right)\prod_{k=1}^{K}\mu_k^{\alpha_k-1} + \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_K)}\left(\frac{\partial}{\partial\alpha_j}\prod_{k=1}^{K}\mu_k^{\alpha_k-1}\right)\end{aligned}$$

Our goal is to compute both terms separately. Firstly, since a small change in one of the sum terms is equivalent to a small change in the sum itself, i.e.

$$\frac{\partial}{\partial\alpha_j}\Gamma(\alpha_0) = \frac{\partial}{\partial\alpha_0}\Gamma(\alpha_0)$$

we have that

$$\begin{aligned}\frac{\partial}{\partial\alpha_j}\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_j)} &= \frac{\frac{\partial}{\partial\alpha_j}\Gamma(\alpha_0) - \frac{\partial}{\partial\alpha_j}\Gamma(\alpha_j)}{\Gamma(\alpha_j)^2} \\ &= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_j)}\left(\frac{\frac{\partial}{\partial\alpha_j}\Gamma(\alpha_0)}{\Gamma(\alpha_0)} - \frac{\frac{\partial}{\partial\alpha_j}\Gamma(\alpha_j)}{\Gamma(\alpha_j)}\right) \\ &= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_j)}\left(\frac{\partial}{\partial\alpha_0}\ln\Gamma(\alpha_0) - \frac{\partial}{\partial\alpha_j}\ln\Gamma(\alpha_j)\right) \\ &= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_j)}(\psi(\alpha_0) - \psi(\alpha_j))\end{aligned}$$

and therefore, that

$$\left(\frac{\partial}{\partial\alpha_j}\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_K)}\right)\prod_{k=1}^{K}\mu_k^{\alpha_k-1} = (\psi(\alpha_0) - \psi(\alpha_j))\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha})$$

Now, since

$$\frac{\partial}{\partial\alpha_j}\prod_{k=1}^{K}\mu_k^{\alpha_k-1} = (\mu_1^{\alpha_1-1}\ldots\mu_{j-1}^{\alpha_{j-1}-1}\mu_{j+1}^{\alpha_{j+1}-1}\ldots\mu_K^{\alpha_K-1})\frac{\partial}{\partial\alpha_j}\mu_j^{\alpha_j-1} = \ln\mu_j\prod_{k=1}^{K}\mu_k^{\alpha_k-1}$$

it follows that

$$\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_K)}\left(\frac{\partial}{\partial\alpha_j}\prod_{k=1}^{K}\mu_k^{\alpha_k-1}\right) = \ln\mu_j\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha})$$

By substituting into the initial expression,

$$\frac{\partial}{\partial\alpha_j}\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha})(\ln\mu_j + \psi(\alpha_0) - \psi(\alpha_j))$$

and then integrating with respect to $\boldsymbol{\mu}$, we obtain the desired result:

$$\mathbb{E}[\ln\mu_j] = \psi(\alpha_j) - \psi(\alpha_0) \tag{2.276}$$

□

Chapter 3

Linear Models for Regression

Exercise 3.1 ⋆

Show that the tanh function and the logistic sigmoid function (3.6) are related by

$$\tanh(a) = 2\sigma(2a) - 1 \tag{3.100}$$

Hence show that a general linear combination of logistic sigmoid functions of the form

$$y(x, \mathbf{w}) = w_0 + \sum_{j=1}^{M} w_j \sigma\left(\frac{x - \mu_j}{s}\right) \tag{3.101}$$

is equivalent to a linear combination of tanh functions of the form

$$y(x, \mathbf{u}) = u_0 + \sum_{j=1}^{M} u_j \tanh\left(\frac{x - \mu_j}{2s}\right) \tag{3.102}$$

and find expressions to relate the new parameters $\{u_0, \ldots, u_M\}$ to the original parameters $\{w_0, \ldots, w_M\}$.

Proof. The logistic sigmoid function is given by

$$\sigma(x) = \frac{1}{1 + \exp(-x)} \tag{3.6}$$

and the tanh function is given by

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{e^{2x} - 1}{e^{2x} + 1} \tag{3.1}$$

By starting from the right-hand side of (3.100) and then using the fact that tanh is odd, we obtain

$$2\sigma(2a) - 1 = \frac{2}{e^{-2a}} - 1 = \frac{1 - e^{-2a}}{1 + e^{-2a}} = -\tanh(-a) = \tanh(a) \tag{3.100}$$

Now, we can express the logistic sigmoid functions as

$$\sigma(x) = \frac{1}{2}\tanh\frac{x}{2} + \frac{1}{2}$$

By substituting this in (3.101), we have that

$$y(x, \mathbf{w}) = w_0 + \frac{M}{2} + \sum_{j=1}^{M} \frac{w_j}{2} \tanh\left(\frac{x - \mu_j}{2s}\right) = y(x, \mathbf{u})$$

where

$$u_0 = w_0 + \frac{M}{2} \qquad\qquad u_j = \frac{1}{2} w_j, j \geq 1$$

Therefore, we proved that (3.101) is equivalent to (3.102). □

Exercise 3.2 ⋆⋆

Show that the matrix

$$\mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T \tag{3.103}$$

takes any vector $\mathbf{v}$ and projects it onto the space spanned by the columns of $\mathbf{\Phi}$. Use this result to show that the least-squares solution (3.15) corresponds to an orthogonal projection of the vector $\mathbf{t}$ onto the manifold $\mathcal{S}$ as shown in Figure 3.2.

Proof. Let $\mathbf{p}$ be the projection of $\mathbf{v}$ onto the space spanned by the columns of $\mathbf{\Phi}$. We then have that $\mathbf{p}$ is contained by the space, so $\mathbf{p}$ can be written as a linear combination of the columns of $\mathbf{\Phi}$, i.e. there exists $\mathbf{x}$ such that $\mathbf{p} = \mathbf{\Phi}\mathbf{x}$. By using this and the fact that $\mathbf{p} - \mathbf{v}$ is orthogonal to the space, we have that

$$\begin{aligned}
\mathbf{\Phi}^T(\mathbf{p} - \mathbf{v}) &= \mathbf{0} \\
\mathbf{\Phi}^T(\mathbf{\Phi}\mathbf{x} - \mathbf{v}) &= \mathbf{0} \\
\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{x} &= \mathbf{\Phi}^T\mathbf{v} \\
\mathbf{x} &= (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{v}
\end{aligned}$$

and since $\mathbf{p} = \mathbf{\Phi}\mathbf{x}$, this proves our hypothesis, i.e.

$$\mathbf{p} = \mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{v}$$

This translates directly to the least-squares geometry described in Section 3.1.3, where the manifold $\mathcal{S}$ is the space spanned by the columns of $\mathbf{\Phi}$. From what we proved above, the projection of $\mathbf{t}$ onto the manifold $\mathcal{S}$ is given by $\mathbf{y} = \mathbf{\Phi}\mathbf{w}_{\mathrm{ML}}$, where

$$\mathbf{w}_{\mathrm{ML}} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{t} \tag{3.15}$$

is the least-squares solution. □

Exercise 3.3 ⋆

Consider a data set in which each data point t_n is associated with a weighting factor $r_n > 0$, so that the sum of squares error function becomes

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} r_n\{t_n - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)\}^2 \tag{3.104}$$

Find an expression for the solution $\mathbf{w}^\star$ that minimizes this error function. Give two alternative interpretations of the weighted sum-of-squares error function in terms of (i) data dependent noise variance and (ii) replicated data points.

Method 1.

Proof. Since the least-squares error function is convex, the function is minimized in its only critical point. Similarly to (3.13), the derivative is given by:

$$\begin{aligned}\frac{\partial}{\partial \mathbf{w}} E_D(\mathbf{w}) &= \frac{1}{2}\sum_{n=1}^{N} r_n \left(\frac{\partial}{\partial \mathbf{w}} \{t_n - \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_n)\}^2 \right) \\ &= \sum_{n=1}^{N} r_n \{\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_n) - t_n\} \boldsymbol{\phi}(\mathbf{x}_n)^T \\ &= \mathbf{w}^T \left(\sum_{i=1}^{N} r_n \boldsymbol{\phi}(\mathbf{x}_n) \boldsymbol{\phi}(\mathbf{x}_n)^T \right) - \sum_{n=1}^{N} r_n t_n \boldsymbol{\phi}(\mathbf{x}_n)^T\end{aligned}$$

By defining the matrix $R = \text{diag}(r_1, r_2, \ldots, r_n)$ and then setting the derivative to 0, we obtain the equality

$$\mathbf{w}^T \boldsymbol{\Phi} R \boldsymbol{\Phi}^T = \mathbf{t}^T R \boldsymbol{\Phi}$$

which gives the weighted least-squares solution (we get the column vector form):

$$\mathbf{w}^\star = (\boldsymbol{\Phi}^T R \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T R \mathbf{t}$$

□

Method 2.

Proof. We define the diagonal matrices $R = \text{diag}(r_1, r_2, \ldots, r_n)$ and $R^{1/2} = \text{diag}(\sqrt{r_1}, \sqrt{r_2}, \ldots, \sqrt{r_n})$ such that $R^{1/2} R^{1/2} = R$. We notice that we can rewrite (3.104) as:

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} \left(\sqrt{r_n} \{t_n - \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_n)\} \right)^2$$

which we can translate into matrix notation as:

$$E_D(\mathbf{w}) = \frac{1}{2} \big(R^{1/2}(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}) \big)^T \big(R^{1/2}(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}) \big)$$

Since the least-squares error function is convex, the function is minimized in its only critical point. The derivative is given by

$$\begin{aligned}\frac{\partial}{\partial \mathbf{w}} E_D(\mathbf{w}) &= -\boldsymbol{\Phi}^T (R^{1/2})^T (R^{1/2}\mathbf{t} - R^{1/2}\boldsymbol{\Phi}\mathbf{w}) \\ &= \boldsymbol{\Phi}^T R \boldsymbol{\Phi}\mathbf{w} - \boldsymbol{\Phi}^T R \mathbf{t}\end{aligned}$$

By setting it to 0, we obtain the solution that minimizes the weighted least-squares error function:

$$\mathbf{w}^\star = (\boldsymbol{\Phi}^T R \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T R \mathbf{t}$$

□

Exercise 3.4 ⋆

Consider a linear model of the form

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i \tag{3.105}$$

together with a sum-of-squares error function of the form

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(\mathbf{x}_n, \mathbf{w}) - t_n\}^2 \tag{3.106}$$

Now suppose that Gaussian noise ϵ_i with zero mean and variance σ^2 is added independently to each of the input variables x_i. By making use of $\mathbb{E}[\epsilon_i] = 0$ and $\mathbb{E}[\epsilon_i \epsilon_j] = \delta_{ij}\sigma^2$, show that minimizing E_D averaged over the noise distribution is equivalent to minimizing the sum-of-squares error for noise-free input variables with the addition of a weight-decay regularization term, in which the bias parameter w_0 is omitted from the regularizer.

Proof. Let the noise-free input variables be denoted by $\mathbf{x}^*$, such that $x_i = x_i^* + \epsilon_i$. (3.105) will then be equivalent to

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i^* + \sum_{i=1}^{D} w_i \epsilon_i = y(\mathbf{x}^*, \mathbf{w}) + \sum_{i=1}^{D} w_i \epsilon_i$$

Now, we aim to find the expression of E_D averaged over the noise distribution, that is:

$$\mathbb{E}[E_D(\mathbf{w})] = \frac{1}{2} \sum_{n=1}^{N} \{\mathbb{E}[y(\mathbf{x}_n, \mathbf{w})^2] - 2t_n \mathbb{E}[y(\mathbf{x}_n, \mathbf{w})] + t_n^2\}$$

The individual expectations are straightforward to compute. Since $\mathbb{E}[\epsilon_i] = 0$, we have that

$$\mathbb{E}[y(\mathbf{x}_n, \mathbf{w})] = \mathbb{E}[y(\mathbf{x}^*, \mathbf{w})] + \sum_{i=1}^{D} w_i \mathbb{E}[\epsilon_i] = y(\mathbf{x}^*, \mathbf{w})$$

Also, $\mathbb{E}[\epsilon_i \epsilon_j] = \delta_{ij}\sigma^2$, so

$$\begin{aligned}
\mathbb{E}[y(\mathbf{x}_n, \mathbf{w})^2] &= \mathbb{E}\left[y(\mathbf{x}^*, \mathbf{w})^2 + 2y(\mathbf{x}^*, \mathbf{w}) \sum_{i=1}^{D} w_i \epsilon_i + \left(\sum_{i=1}^{D} w_i \epsilon_i \right)^2 \right] \\
&= y(\mathbf{x}^*, \mathbf{w})^2 + \sum_{i=1}^{D} w_i^2 \mathbb{E}[\epsilon_i^2] + 2 \sum_{i=1}^{D} \sum_{j=i+1}^{D} w_i w_j \mathbb{E}[\epsilon_i \epsilon_j] \\
&= y(x^*, \mathbf{w})^2 + \sigma \sum_{n=1}^{D} w_i^2
\end{aligned}$$

Therefore, we have that

$$\mathbb{E}[E_D(\mathbf{w})] = \frac{1}{2} \sum_{n=1}^{D} \{y(\mathbf{x}_n^*, \mathbf{w}) - t_n\}^2 + \frac{N\sigma}{2} \sum_{n=1}^{D} w_i^2$$

which shows that E_D averaged over the noise distribution is equivalent to the regularized least-squares error function with $\lambda = N\sigma$. Hence, since the expressions are equivalent, minimizing them is also equivalent, proving our hypothesis. □

Exercise 3.5 $\star$

Using the technique of Lagrange multipliers, discussed in Appendix E, show that minimization of the regularized error function (3.29) is equivalent to minimizing the unregularized sum-of-squares error (3.12) subject to the constraint (3.30). Discuss the relationship between the parameters η and λ.

Proof. To minimize the unregularized sum-of-squares error (3.12) subject to the constraint (3.30), is equivalent to minimizing the Lagrangian

$$L(\mathbf{x}, \lambda) = \frac{1}{2}\sum_{n=1}^{N}\{y(\mathbf{x}_n, \mathbf{w}) - t_n\}^2 - \lambda\left(\eta - \sum_{j=1}^{M}|w_j|^q\right)$$

subject to the KKT conditions (see E.9, E.10, E.11 in Appendix E). Our Lagrangian and the regularized sum-of-squares error have the same dependency over $\mathbf{w}$, so their minimization is equivalent. By following (E.11), we have that

$$\lambda\left(\eta - \sum_{j=1}^{M}|w_j|^q\right) = 0$$

which means that if $\mathbf{w}^\star(\lambda)$ is the solution of minimization for a fixed $\lambda > 0$, we then have that

$$\eta = \sum_{j=1}^{M}|w^\star(\lambda)_j|^q$$

□

Exercise 3.6 $\star$

Consider a linear basis function regression model for a multivariate target variable $\mathbf{t}$ having a Gaussian distribution of the form

$$p(\mathbf{t}|\mathbf{W}, \mathbf{\Sigma}) = \mathcal{N}(\mathbf{t}|\mathbf{y}(\mathbf{x}, \mathbf{W}), \mathbf{\Sigma}) \tag{3.107}$$

where

$$\mathbf{y}(\mathbf{x}, \mathbf{W}) = \mathbf{W}^T\boldsymbol{\phi}(\mathbf{x}) \tag{3.108}$$

together with a training data set compromising input basis vectors $\boldsymbol{\phi}(\mathbf{x}_n)$ and corresponding target vectors $\mathbf{t}_n$, with $n = 1, \ldots, N$. Show that the maximum likelihood solution $\mathbf{W}_{\text{ML}}$ for the parameter matrix $\mathbf{W}$ has the property that each column is given by an expression of the form (3.15), which was the solution for an isotropic noise distribution. Note that this is independent of the covariance matrix $\mathbf{\Sigma}$. Show that the maximum likelihood solution for $\mathbf{\Sigma}$ is given by

$$\mathbf{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{t}_n - \mathbf{W}_{\text{ML}}^T\boldsymbol{\phi}(\mathbf{x}_n)\right)\left(\mathbf{t}_n - \mathbf{W}_{\text{ML}}^T\boldsymbol{\phi}(\mathbf{x}_n)\right)^T \tag{3.109}$$

Proof. Similarly to what we did in Section 3.1.5, we combine the set of target vectors into a matrix $\mathbf{T}$ of size $N \times K$ such that the n^{th} row is given by $\mathbf{t}_n^T$. We do the same for $\mathbf{X}$. The log likelihood function is then given by

$$\begin{aligned}
\ln p(\mathbf{T}|\mathbf{X}, \mathbf{W}, \boldsymbol{\Sigma}) &= \ln \prod_{n=1}^{N} \mathcal{N}(\mathbf{t}_n|\mathbf{W}^T\boldsymbol{\phi}(\mathbf{x}_n), \boldsymbol{\Sigma}) \\
&= \sum_{n=1}^{N} \ln \mathcal{N}(\mathbf{t}_n|\mathbf{W}^T\boldsymbol{\phi}(\mathbf{x}_n), \boldsymbol{\Sigma}) \\
&= \sum_{n=1}^{N} \ln \left[\frac{1}{(2\pi)^{K/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{ \big(\mathbf{t}_n - \mathbf{W}^T\boldsymbol{\phi}(\mathbf{x}_n)\big)^T \boldsymbol{\Sigma}^{-1} \big(\mathbf{t}_n - \mathbf{W}^T\boldsymbol{\phi}(\mathbf{x}_n)\big) \right\} \right] \\
&= -\frac{NK}{2} \ln(2\pi) - \frac{1}{2} \ln|\boldsymbol{\Sigma}| + \sum_{n=1}^{N} \big(\mathbf{t}_n - \mathbf{W}^T\boldsymbol{\phi}(\mathbf{x}_n)\big)^T \boldsymbol{\Sigma}^{-1} \big(\mathbf{t}_n - \mathbf{W}^T\boldsymbol{\phi}(\mathbf{x}_n)\big)
\end{aligned}$$

Our goal is to maximise this function with respect to $\mathbf{W}$. We take the derivative of the likelihood and use the fact that $\boldsymbol{\Sigma}^{-1}$ is symmetric and (88) from the matrix cookbook to obtain:

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{W}} \ln p(\mathbf{T}|\mathbf{X}, \mathbf{W}, \boldsymbol{\Sigma}) &= \sum_{n=1}^{N} \frac{\partial}{\partial \mathbf{W}} \big(\mathbf{t}_n - \mathbf{W}^T\boldsymbol{\phi}(\mathbf{x}_n)\big)^T \boldsymbol{\Sigma}^{-1} \big(\mathbf{t}_n - \mathbf{W}^T\boldsymbol{\phi}(\mathbf{x}_n)\big) \\
&= -2\boldsymbol{\Sigma}^{-1} \sum_{n=1}^{N} \big(\mathbf{t}_n - \mathbf{W}^T\boldsymbol{\phi}(\mathbf{x}_n)\big)\boldsymbol{\phi}(\mathbf{x}_n)^T
\end{aligned}$$

By setting the derivative equal to 0, we find the maximum likelihood solution for $\mathbf{W}$:

$$\begin{aligned}
-2\boldsymbol{\Sigma}^{-1} \sum_{n=1}^{N} \big(\mathbf{t}_n - \mathbf{W}_{\text{ML}}^T\boldsymbol{\phi}(\mathbf{x})\big)\boldsymbol{\phi}(\mathbf{x})^T &= 0 \\
\boldsymbol{\Sigma}^{-1} \sum_{n=1}^{N} \mathbf{t}_n\boldsymbol{\phi}(\mathbf{x}_n)^T &= \boldsymbol{\Sigma}^{-1}\mathbf{W}_{\text{ML}}^T \sum_{n=1}^{N} \boldsymbol{\phi}(\mathbf{x}_n)\boldsymbol{\phi}(\mathbf{x}_n)^T \\
\boldsymbol{\Sigma}^{-1}\mathbf{T}^T\boldsymbol{\Phi} &= \boldsymbol{\Sigma}^{-1}\mathbf{W}_{\text{ML}}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi} \\
\boldsymbol{\Phi}^T\mathbf{T}\boldsymbol{\Sigma}^{-1} &= \boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{W}_{\text{ML}}\boldsymbol{\Sigma}^{-1}
\end{aligned}$$

Note that $\boldsymbol{\Sigma}^{-1}$ cancels out and we finally get that:

$$\mathbf{W}_{\text{ML}} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\mathbf{T}$$

Now, let A, B be two matrices of size $N \times M$ and let $b_1, b_2, \ldots, b_N$ be the column vectors of B. One could easily prove that

$$AB = A\big(b_1\ b_2\ \ldots\ b_N\big) = \big(Ab_1\ Ab_2\ \ldots\ Ab_N\big)$$

By using this for our case, that is to find the columns of $\mathbf{W}_{\text{ML}}$, we'd find that they are of the form (3.15), i.e. the n^{th} column of $\mathbf{W}_{\text{ML}}$ is given by

$$\mathbf{W}_{\text{ML}}^{(n)} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\mathbf{T}^{(n)}$$

where $\mathbf{T}^{(n)}$ is the n^{th} column of $\mathbf{T}$. □

Exercise 3.7 $\star$

By using the technique of completing the square, verify the result (3.49) for the posterior distribution of the parameters $\mathbf{w}$ in the linear basis function model in which $\mathbf{m}_N$ and $\mathbf{S}_N$ are defined by (3.50) and (3.51) respectively.

Proof. Since

$$\begin{aligned} p(\mathbf{w}|\mathbf{t}) &\propto p(\mathbf{w})p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta^{-1}) \\ &\propto \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0) \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}) \end{aligned}$$

we have that

$$\ln p(\mathbf{w}|\mathbf{t}) = \ln \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) + \ln \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}) + \text{const} \tag{3.7.1}$$

We compute the first logarithm, expand the square and keep only the terms that depend on $\mathbf{w}$ to obtain:

$$\ln \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0) = -\frac{1}{2}\mathbf{w}^T\mathbf{S}_0^{-1}\mathbf{w} + \mathbf{w}^T\mathbf{S}_0^{-1}\mathbf{m}_0 + \text{const}$$

By doing the same for the second term, we'll have that:

$$\begin{aligned} \ln \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}) &= \sum_{n=1}^{N} \ln \mathcal{N}(t_n|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}) \\ &= -\beta\mathbf{w}^T \sum_{n=1}^{N} t_n\boldsymbol{\phi}(\mathbf{x}_n) - \frac{\beta}{2}\sum_{n=1}^{N} \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)\boldsymbol{\phi}(\mathbf{x}_n)^T\mathbf{w} + \text{const} \\ &= -\beta\mathbf{w}^T\boldsymbol{\Phi}^T\mathbf{t} - \frac{\beta}{2}\mathbf{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{w} + \text{const} \end{aligned}$$

By replacing back into (3.7.1),

$$\begin{aligned} \ln p(\mathbf{w}|\mathbf{t}) &= -\frac{1}{2}\mathbf{w}^T\mathbf{S}_0^{-1}\mathbf{w} + \mathbf{w}^T\mathbf{S}_0^{-1}\mathbf{m}_0 - \beta\mathbf{w}^T\boldsymbol{\Phi}^T\mathbf{t} - \frac{\beta}{2}\mathbf{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{w} + \text{const} \\ &= -\frac{1}{2}\mathbf{w}(\mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi})\mathbf{w}^T + \mathbf{w}^T(\mathbf{S}_0^{-1}\mathbf{m}_0 - \beta\boldsymbol{\Phi}^T\mathbf{t}) + \text{const} \end{aligned}$$

The quadratic term corresponds to a Gaussian with the covariance matrix $\mathbf{S}_N$, where

$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi} \tag{3.51}$$

Now, since the mean is found in the linear term, we'd have that

$$\mathbf{w}^T(\mathbf{S}_0^{-1}\mathbf{m}_0 - \beta\boldsymbol{\Phi}^T\mathbf{t}) = \mathbf{w}^T\mathbf{S}_N^{-1}\mathbf{m}_N$$

which gives

$$\mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m_0} + \beta\boldsymbol{\Phi}^T\mathbf{t}) \tag{3.50}$$

Since we proved both (3.50) and (3.51), we showed that

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) \tag{3.49}$$

□

Exercise 3.8 ⋆⋆

Consider the linear basis function model in Section 3.1, and suppose that we already have observed N data points, so that the posterior distribution over $\mathbf{w}$ is given by (3.49). This posterior can be regarded as the prior for the next observation. By considering an additional data point $(\mathbf{x}_{N+1}, t_{N+1})$, and by completing the square in the exponential, show that the resulting posterior distribution is again given by (3.49) but with $\mathbf{S}_N$ replaced by $\mathbf{S}_{N+1}$ and $\mathbf{m}_N$ replaced by $\mathbf{m}_{N+1}$.

Proof. Our approach will be very similar to the previous exercise. The posterior distribution is given by the proportionality relation

$$\begin{aligned} p(\mathbf{w}|\mathbf{t}) &\propto p(\mathbf{w})p(t_{N+1}|\mathbf{x}_{N+1}, \mathbf{w}, \beta^{-1}) \\ &\propto \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)\mathcal{N}(t_{N+1}|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_{N+1}), \beta^{-1}) \end{aligned}$$

, so

$$\ln p(\mathbf{w}|\mathbf{t}) = \ln \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) + \ln \mathcal{N}(t_{N+1}|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_{N+1}), \beta^{-1}) + \text{const} \tag{3.8.1}$$

We now compute the log likelihood and keep only the terms depending on $\mathbf{w}$ to obtain:

$$\ln \mathcal{N}(t_{N+1}|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_{N+1}), \beta^{-1}) = -\frac{\beta}{2}\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_{N+1})\boldsymbol{\phi}(\mathbf{x}_{N+1})^T\mathbf{w} - \beta t_{N+1}\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_{N+1}) + \text{const}$$

By expanding the square and then doing the same with the prior, we have that:

$$\ln \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) = -\frac{1}{2}\mathbf{w}^T\mathbf{S}_N^{-1}\mathbf{w} + \mathbf{w}^T\mathbf{S}_N^{-1}\mathbf{m}_N + \text{const}$$

Substituting these results back into (3.8.1) yields:

$$\ln p(\mathbf{w}|\mathbf{t}) = -\frac{1}{2}\mathbf{w}^T\big(\mathbf{S}_N^{-1} - \beta\boldsymbol{\phi}(\mathbf{x}_{N+1})\boldsymbol{\phi}(\mathbf{x}_{N+1})^T\big)\mathbf{w} + \mathbf{w}^T\big(\mathbf{S}_N^{-1}\mathbf{m}_N - \beta t_{N+1}\boldsymbol{\phi}(\mathbf{x}_{N+1})\big) + \text{const}$$

which is equivalent to

$$\ln p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_{N+1}, \mathbf{S}_{N+1})$$

for

$$\mathbf{S}_{N+1}^{-1} = \mathbf{S}_N^{-1} + \beta\boldsymbol{\phi}(\mathbf{x}_{N+1})\boldsymbol{\phi}(\mathbf{x}_{N+1})^T \tag{3.8.2}$$

and

$$\mathbf{m}_{N+1} = \mathbf{S}_{N+1}\big(\mathbf{S}_N^{-1}\mathbf{m}_N - \beta t_{N+1}\boldsymbol{\phi}(\mathbf{x}_{N+1})\big)$$

□

Exercise 3.9 ⋆⋆

Repeat the previous exercise but instead of completing the square by hand, make use of the general result for linear-Gaussian models given by (2.116).

Proof. As shown in Section 2.3.3, given a marginal Gaussian distribution for $\mathbf{x}$ and a conditional Gaussian distribution for $\mathbf{y}$ given $\mathbf{x}$ in the form

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \tag{2.113}$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) \tag{2.114}$$

the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ is given by

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^T\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}) \tag{2.116}$$

where

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A})^{-1} \tag{2.117}$$

Our goal is to match these results with our model. The prior is given by

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{m}_N, \mathbf{S}_N)$$

and the likelihood is

$$p(t_{N+1}|\mathbf{x}_{N+1}, \mathbf{w}, \beta^{-1}) = \mathcal{N}(t_{N+1}|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1})$$

By comparing those with (2.113) and (2.114), we'd have that the variables are related as follows:

$$\mathbf{x} = \mathbf{w} \qquad \mathbf{y} = t_{N+1} \qquad \boldsymbol{\mu} = \mathbf{m}_N \qquad \boldsymbol{\Lambda}^{-1} = \mathbf{S}_N \qquad \mathbf{A} = \boldsymbol{\phi}(\mathbf{x}_N)^T \qquad \mathbf{b} = 0 \qquad \mathbf{L}^{-1} = \beta^{-1}$$

Therefore, the covariance matrix $\boldsymbol{\Sigma}$ of the conditional (the $\mathbf{S}_{N+1}$ of our posterior) will be given by substituting our variables into (2.117), so

$$\mathbf{S}_{N+1}^{-1} = \mathbf{S}_N^{-1} + \beta\boldsymbol{\phi}(\mathbf{x}_N)\boldsymbol{\phi}(\mathbf{x}_N)^T$$

The mean can also be easily obtained from (2.116) as

$$\mathbf{m}_{N+1} = \mathbf{S}_{N+1}\big(\mathbf{S}_N^{-1}\mathbf{m}_N - \beta t_{N+1}\boldsymbol{\phi}(\mathbf{x}_{N+1})\big)$$

□

Exercise 3.10

By making use of the result (2.115) to evaluate the integral in (3.57), verify that the predictive distribution for the Bayesian linear regression model is given by (3.58) in which the input-dependent variance is given by (3.59).

Proof. We've seen in Section 2.3.3 that given a marginal Gaussian distribution for $\mathbf{x}$ and a conditional Gaussian distribution for $\mathbf{y}$ given $\mathbf{x}$ in the forms (2.113) and (2.114), we have that the marginal distribution of $\mathbf{y}$ is given by

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T) \tag{2.115}$$

Therefore, if we consider the terms under the integral in (3.57), we have that

$$p(\mathbf{w}|\mathbf{t}, \mathbf{x}, \alpha, \beta) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

$$p(t|\mathbf{w}, \mathbf{x}, \alpha, \beta) = \mathcal{N}(t|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}), \beta^{-1})$$

so the marginal distribution of t will be given by (2.115). Our goal is to find the parameters of this distribution. By considering the notation used in (2.113), (2.114), and (2.115), we'd have that

$$\boldsymbol{\mu} = \mathbf{m}_N \quad \mathbf{S}_N = \boldsymbol{\Lambda}^{-1} \quad \mathbf{A} = \boldsymbol{\phi}(\mathbf{x})^T \quad \mathbf{b} = 0 \quad \mathbf{L}^{-1} = \beta^{-1}$$

Finally, by substituting our values into (2.115), it is straightforward to see that the predictive distribution for the Bayesian linear regression model is given by

$$p(t|\mathbf{t}, \mathbf{x}, \alpha, \beta) = \mathcal{N}\big(t|\boldsymbol{\phi}(\mathbf{x})^T\mathbf{m}_N, \sigma_N^2(\mathbf{x})\big) \tag{3.58}$$

where the input-dependent variance is given by

$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})^T \tag{3.59}$$

□

Exercise 3.11

We have seen that, as the size of a data set increases, the uncertainty associated with the posterior distribution over model parameters decreases. Make use of the matrix identity (Appendix C)

$$\big(\mathbf{M} + \mathbf{v}\mathbf{v}^T\big)^{-1} = \mathbf{M}^{-1} - \frac{(\mathbf{M}^{-1}\mathbf{v})\big(\mathbf{v}^T\mathbf{M}^{-1}\big)}{1 + \mathbf{v}^T\mathbf{M}^{-1}\mathbf{v}} \tag{3.110}$$

to show that the uncertainty $\sigma_{N+1}^2(\mathbf{x})$ associated with the linear regression function given by (3.59) satisfies

$$\sigma_{N+1}^2(\mathbf{x}) \leq \sigma_N^2(\mathbf{x}) \tag{3.111}$$

Proof. By using (3.59) and then (3.8.2) we have that:

$$\sigma_{N+1}^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^T\mathbf{S}_{N+1}\boldsymbol{\phi}(\mathbf{x}) = \frac{1}{\beta}\boldsymbol{\phi}(\mathbf{x})^T\left[\mathbf{S}_N^{-1} + \beta\boldsymbol{\phi}(\mathbf{x}_N)\boldsymbol{\phi}(\mathbf{x}_N)^T\right]^{-1}\boldsymbol{\phi}(\mathbf{x})$$

We apply (3.110) with $\mathbf{M} = \mathbf{S}_N^{-1}$ and $\mathbf{v} = \beta^{1/2}\boldsymbol{\phi}(\mathbf{x})$ and get that

$$\begin{aligned}
\sigma_{N+1}^2(\mathbf{x}) &= \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^T\left[\mathbf{S}_N - \frac{\beta\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_N)\boldsymbol{\phi}(\mathbf{x}_N)^T\mathbf{S}_N}{1 + \beta\boldsymbol{\phi}(\mathbf{x}_N)^T\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_N)}\right]\boldsymbol{\phi}(\mathbf{x}) \\
&= \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^T\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\phi}(\mathbf{x})^T\frac{\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_N)\boldsymbol{\phi}(\mathbf{x}_N)^T\mathbf{S}_N}{\frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x}_N)^T\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_N)}\boldsymbol{\phi}(\mathbf{x}) \\
&= \sigma_N^2(\mathbf{x}) - \boldsymbol{\phi}(\mathbf{x})^T\frac{\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_N)\boldsymbol{\phi}(\mathbf{x}_N)^T\mathbf{S}_N}{\frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x}_N)^T\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_N)}\boldsymbol{\phi}(\mathbf{x})
\end{aligned}$$

Therefore,

$$\sigma_N^2(\mathbf{x}) - \sigma_{N+1}^2(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^T\frac{\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_N)\boldsymbol{\phi}(\mathbf{x}_N)^T\mathbf{S}_N}{\frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x}_N)^T\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_N)}\boldsymbol{\phi}(\mathbf{x}) \tag{3.11.1}$$

Since $\mathbf{S}_N$ is a precision matrix, it is symmetric, so:

$$\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_N)\boldsymbol{\phi}(\mathbf{x}_N)^T\mathbf{S}_N = \big(\boldsymbol{\phi}(\mathbf{x}_N)^T\mathbf{S}_N\big)^T\boldsymbol{\phi}(\mathbf{x}_N)^T\mathbf{S}_N = ||\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_N)||^2 \geq 0$$

Even more, because $\mathbf{S}_N$ is a precision matrix, it is positive semidefinite. By using this and the fact that the noise precision constant β is positive, we have that:

$$\frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x}_N)^T\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_N) \geq 0$$

Hence, we finally have that

$$\boldsymbol{\phi}(\mathbf{x})^T\frac{\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_N)\boldsymbol{\phi}(\mathbf{x}_N)^T\mathbf{S}_N}{\frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x}_N)^T\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_N)}\boldsymbol{\phi}(\mathbf{x}) = \frac{\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_N)\boldsymbol{\phi}(\mathbf{x}_N)^T\mathbf{S}_N}{\frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x}_N)^T\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}_N)}||\boldsymbol{\phi}(\mathbf{x})||^2 \geq 0$$

which, by (3.11.1), becomes equivalent to (3.111). □

Exercise 3.12

We saw in Section 2.3.6 that the conjugate prior for a Gaussian distribution with unknown mean and unknown precision (inverse variance) is a normal-gamma distributuion. This property also holds for the case of the conditional Gaussian distribution $p(t|\mathbf{x}, \mathbf{w}, \beta)$ of the linear regression model. If we consider the likelihood function (3.10), then the conjugate prior for $\mathbf{w}$ and β is given by

$$p(\mathbf{w}, \beta) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \beta^{-1}\mathbf{S}_0)\mathrm{Gam}(\beta|a_0, b_0) \tag{3.112}$$

Show that the corresponding posterior distribution takes the same functional form, so that

$$p(\mathbf{w}, \beta|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \beta^{-1}\mathbf{S}_N)\mathrm{Gam}(\beta|a_N, b_N) \tag{3.113}$$

and find expressions for the posterior parameters $\mathbf{m}_N$, $\mathbf{S}_N$, a_N, and b_N.

Proof. We have that

$$\begin{aligned} p(\mathbf{w}, \beta|\mathbf{t}) &\propto p(\mathbf{w}, \beta)p(\mathbf{t}|\mathbf{w}, \beta) \\ &\propto \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \beta^{-1}\mathbf{S}_0)\mathrm{Gam}(\beta|a_0, b_0)\prod_{n=1}^{N}\mathcal{N}(t_n|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}) \end{aligned}$$

so

$$\ln p(\mathbf{w}, \beta|\mathbf{t}) = \ln\mathcal{N}(\mathbf{w}|\mathbf{m}_0, \beta^{-1}\mathbf{S}_0) + \ln\mathrm{Gam}(\beta|a_0, b_0) + \ln\prod_{n=1}^{N}\mathcal{N}(t_n|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}) + \text{const}$$

We decompose each logarithm, this time keeping each term. The log likelihood is derived like in Exercise 3.7, that is:

$$\ln\prod_{n=1}^{N}\mathcal{N}(t_n|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}) = -\frac{\beta}{2}\mathbf{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{w} - \beta\mathbf{w}^T\boldsymbol{\Phi}^T\mathbf{t} - \frac{\beta}{2}\mathbf{t}^T\mathbf{t}$$

The logarithms of factors in the prior are given by:

$$\ln \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \beta^{-1}\mathbf{S}_0) = -\frac{\beta}{2}\mathbf{w}^T\mathbf{S}_0^{-1}\mathbf{w} + \beta\mathbf{w}^T\mathbf{S}_0^{-1}\mathbf{m}_0 - \frac{\beta}{2}\mathbf{m}_0^T\mathbf{S}_0^{-1}\mathbf{m}_0$$

$$\ln \mathrm{Gam}(\beta|a_0, b_0) = -\ln\Gamma(a_0) + a_0 \ln b_0 + a_0 \ln\beta - \ln\beta - b_0\beta$$

Now, the log of the posterior is given by:

$$\begin{aligned}\ln p(\mathbf{w}, \beta|\mathbf{t}) = &-\frac{1}{2}\mathbf{w}^T(\beta\mathbf{S}_0^{-1} + \beta\mathbf{\Phi}^T\mathbf{\Phi})\mathbf{w} + \mathbf{w}^T(\beta\mathbf{S}_0^{-1}\mathbf{m}_0 - \beta\mathbf{\Phi}^T\mathbf{t}) - \frac{\beta}{2}\mathbf{t}^T\mathbf{t} - \frac{\beta}{2}\mathbf{m}_0^T\mathbf{S}_0^{-1}\mathbf{m}_0 \\ &+(a_0 - 1)\ln\beta - b_0\beta + \text{const}\end{aligned}$$

The covariance matrix of the posterior is easily found from the quadratic term, that is:

$$\mathbf{S}_N^{-1} = \beta\mathbf{S}_0^{-1} + \beta\mathbf{\Phi}^T\mathbf{\Phi}$$

The mean is obtained from the linear term by using the fact that

$$\mathbf{w}^T\mathbf{S}_N^{-1}\mathbf{m}_N = \mathbf{w}^T(\beta\mathbf{S}_0^{-1}\mathbf{m}_0 - \beta\mathbf{\Phi}^T\mathbf{t})$$

so

$$\mathbf{m}_N = \mathbf{S}_N(\beta\mathbf{S}_0^{-1}\mathbf{m}_0 - \beta\mathbf{\Phi}^T\mathbf{t})$$

Finally, the constant term with respect to $\mathbf{w}$ will give us the parameters of the Gamma distribution. □