Pattern Recognition and Machine Learning
Cristopher Bishop

Exercise Solutions

Stefan Stefanache

February 24, 2022

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 $\mathbf{w}$:

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

$$= \sum_{n=1}^{N} y(x_n, \mathbf{w}) \frac{\mathrm{d}}{\mathrm{d}\mathbf{w}} y(x_n, \mathbf{w}) - \sum_{n=1}^{N} t_n \frac{\mathrm{d}}{\mathrm{d}\mathbf{w}} y(x_n, \mathbf{w}) \tag{1.1.1}$$

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 $\text{var}[f(x)]$ satisfies (1.39).

Proof. The variance is defined by

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

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

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

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

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

□

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

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

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:

$$\frac{\partial}{\partial\sigma^2} \left(\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} \mathrm{d}x \right) = 0$$

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

$$-\frac{1}{2\sigma^2}+\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}\frac{(x-\mu)^2}{2\sigma^4}\exp\left\{-\frac{(x+\mu)^2}{2\sigma^2}\right\}\mathrm{d}x=0$$

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:

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

□

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

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

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

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:

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

□

Exercise 1.11 $\star$

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:

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

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 $\star\star$

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:

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

We compute each expectation separately and get:

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

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

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 $w_{ij} = w_{ij}^S + w_{ij}^A$ where w_{ij}^S and w_{ij}^A are symmetric and anti-symmetric matrices, respectively, satisfying $w_{ij}^S = w_{ji}^S$ 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

$$\lim_{D\to\infty} \frac{\pi^{D/2}}{D2^{D-1}\Gamma(D/2)} = \lim_{D\to\infty} \frac{\pi^{D/2}}{D2^{D-1}(2\pi)^{1/2} e^{1-D/2}(D/2-1)^{D/2-1/2}}$$

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

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:

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

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

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

□

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) $I[x,y]$

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) $H[x,y] = \ln 3$

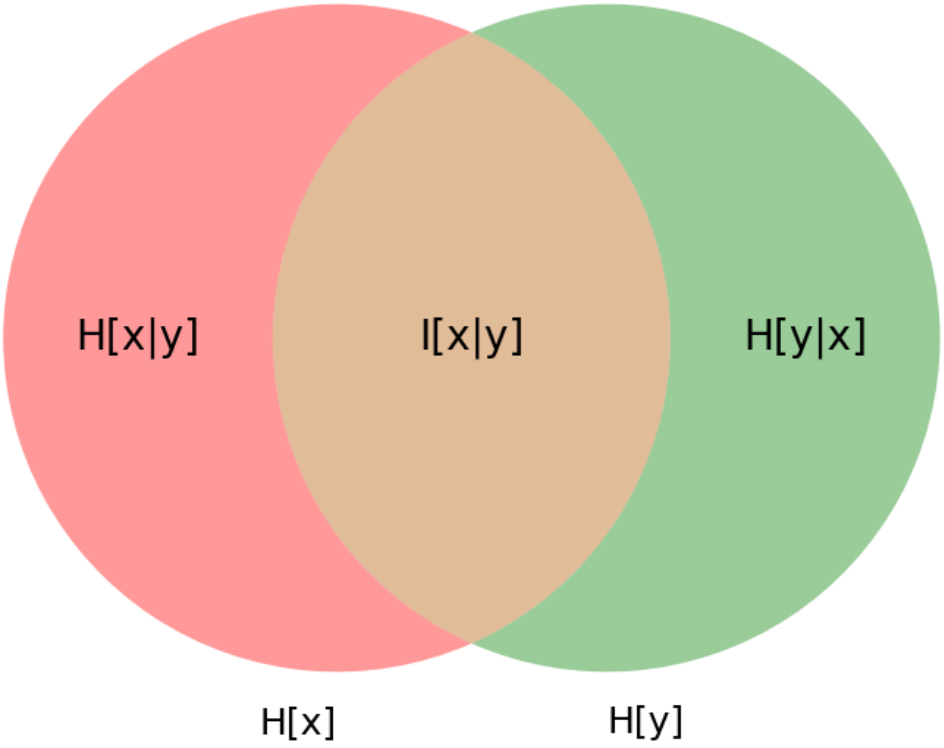
(b) $H[y] = -2/3\ln 2 + \ln 3$ (d) $H[y|x] = 2/3\ln 2$ (f) $I[x,y] = -4/3\ln 2 + \ln 3$

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.



Exercise 1.39 Diagram

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

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

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

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:

$$\mathrm{var}[\mu] = \int_0^1 \left(\mu - \frac{a}{a+b}\right)^2 \mathrm{Beta}(\mu|a,b)\,\mathrm{d}\mu$$

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

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:

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

$$\lambda = \frac{l+m}{a+m+b+l}$$

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:

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

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

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

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

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

$$= \frac{\alpha_j \alpha_l}{\alpha_0(\alpha_0+1)}$$

Finally, the covariance becomes

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

□

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:

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

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

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

Now, since

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

it follows that

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

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

Note that the results (3.50∗) and (3.51∗) derived in Exercise 3.12 seem to be different than (3.50) and (3.51) from the book. There doesn't seem to be any mention of them in the errata comments, but the results used in the web solution for Exercise 3.23 seems to be the ones we've got, and not the ones from the book.

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 $\mathbf{y} = \mathbf{\Phi}\mathbf{w}_{\text{ML}}$, where

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

is the least-squares solution. □

Exercise 3.3 $\star$

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

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

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

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

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

□

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

$$\mathbb{E}[y(\mathbf{x}_n, \mathbf{w})^2] = \mathbb{E}\left[y(\mathbf{x}^*, \mathbf{w})^2 + 2y(\mathbf{x}^*, \mathbf{w})\sum_{i=1}^{D} w_i\epsilon_i + \left(\sum_{i=1}^{D} w_i\epsilon_i\right)^2\right]$$

$$= y(\mathbf{x}^*, \mathbf{w})^2 + \sum_{i=1}^{D} w_i^2 \mathbb{E}[\epsilon_i^2] + 2 \sum_{i=1}^{D} \sum_{j=i+1}^{D} w_i w_j \mathbb{E}[\epsilon_i \epsilon_j]$$

$$= y(x^*, \mathbf{w})^2 + \sigma \sum_{n=1}^{D} w_i^2$$

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 ⋆

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 ⋆

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 $\boldsymbol{\Sigma}$. Show that the maximum likelihood solution for $\boldsymbol{\Sigma}$ is given by

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

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

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

Our goal is to maximise this function with respect to $\mathbf{W}$. One could prove that for a symmetric matrix $\mathbf{B}$,

$$\frac{\partial}{\partial \mathbf{A}} (\mathbf{a} - \mathbf{A}\mathbf{b})^T \mathbf{B} (\mathbf{a} - \mathbf{A}\mathbf{b}) = -2\mathbf{B}(\mathbf{a} - \mathbf{A}\mathbf{b})\mathbf{b}^T$$

Therefore, we take the derivative of the likelihood and use the fact that $\boldsymbol{\Sigma}^{-1}$ is symmetric to obtain:

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

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

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

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

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 ⋆

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:

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

$$= \beta \mathbf{w}^T \mathbf{\Phi}^T \mathbf{t} - \frac{\beta}{2} \mathbf{w}^T \mathbf{\Phi}^T \mathbf{\Phi} \mathbf{w} + \text{const}$$

By replacing back into (3.7.1),

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

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

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

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

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

which gives

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

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:

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

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 integral now becomes:

$$\begin{aligned} p(t|\mathbf{x}, \mathbf{t}, \alpha, \beta) &= \int p(t|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\mathbf{t}, \mathbf{x}, \alpha, \beta)\, \mathrm{d}\mathbf{w} \\ &= \int \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)\mathcal{N}(t|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}), \beta^{-1})\, \mathrm{d}\mathbf{w} \end{aligned} \tag{3.57}$$

Our goal is to find the parameters of this distribution. Since the integral involves the convolution of two Gaussians, by following 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:

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

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

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

Therefore,

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

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

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

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

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

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

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

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 also keeping each term depending on β. The log likelihood is derived like in Exercise 3.7, that is:

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

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:

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

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

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

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

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

so

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

From the constant terms with respect to $\mathbf{w}$ we'll obtain the parameters of the Gamma distribution. b_N is obtained by using the linear terms containing β. Since we already know the covariance and the mean, we can deduce the linear terms of the posterior distribution, so we'll have that:

$$-\beta b_N - \frac{\beta}{2}\mathbf{m}_N^T\mathbf{S}_N^{-1}\mathbf{m}_N = -\frac{\beta}{2}\mathbf{t}^T\mathbf{t} - \frac{\beta}{2}\mathbf{m}_0^T\mathbf{S}_0^{-1}\mathbf{m}_0 - \beta b_0$$

which gives

$$b_N = b_0 + \frac{1}{2}\big(\mathbf{t}^T\mathbf{t} + \mathbf{m}_0^T\mathbf{S}_0^{-1}\mathbf{m}_0 - \mathbf{m}_N^T\mathbf{S}_N^{-1}\mathbf{m}_N\big) \tag{3.12.1}$$

Finally, a_N is given by the terms containing $\ln \beta$. By knowing the $\ln \beta$ terms that will be used in the expansion of the log posterior, we have that

$$(a_N - 1) \ln \beta = \frac{N}{2} \ln \beta + (a_0 - 1) \ln \beta$$

Hence, it is straightforward to obtain the result

$$a_N = a_0 + \frac{N}{2} \tag{2.150}$$

□

Exercise 3.13 ⋆⋆

Show that the predictive distribution $p(t|\mathbf{x}, \mathbf{t})$ for the model discussed in Exercise 3.12 is given by a Student's t-distribution of the form

$$p(t|\mathbf{x}, \mathbf{t}) = \mathrm{St}(t|\mu, \lambda, \nu) \tag{3.114}$$

and obtain expressions for μ, λ, ν.

Proof. The Student's t-distribution is given by

$$\mathrm{St}(x|\mu, \lambda, \nu) = \frac{\Gamma(\nu/2 + 1/2)}{\Gamma(\nu/2)} \left(\frac{\lambda}{\pi\nu}\right)^{1/2} \left[1 + \frac{\lambda(x - \mu)^2}{\nu}\right]^{-\nu/2 - 1/2} \tag{2.159}$$

However, our goal is to obtain it in the form

$$\mathrm{St}(x|\mu, \lambda, \nu) = \int_0^\infty \mathcal{N}\big(x|\mu, (\eta\lambda)^{-1}\big) \mathrm{Gam}(\eta|\nu/2, \nu/2)\, \mathrm{d}\eta \tag{2.158}$$

which for $\nu = 2a$ and $\lambda = a/b$ is equivalent to (2.159).

We have that the predictive distribution is given by

$$p(t|\mathbf{x}, \mathbf{t}) = \iint p(t|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}, \beta|\mathbf{x}, \mathbf{t})\, \mathrm{d}\mathbf{w}\, \mathrm{d}\beta$$

The factors under the integral are already known from (3.8) and (3.113), so

$$\begin{aligned} p(t|\mathbf{x}, \mathbf{t}) &= \iint \mathcal{N}(t|\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}), \beta^{-1}) \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \beta^{-1}\mathbf{S}_N) \mathrm{Gam}(\beta|a_N, b_N)\, \mathrm{d}\mathbf{w}\, \mathrm{d}\beta \\ &= \int \left(\int \mathcal{N}(t|\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}), \beta^{-1}) \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \beta^{-1}\mathbf{S}_N)\, \mathrm{d}\mathbf{w} \right) \mathrm{Gam}(\beta|a_N, b_N)\, \mathrm{d}\beta \end{aligned}$$

The integral with respect to $\mathbf{w}$ is actually (3.57), so we know that it's equal to (3.58). Knowing this, we have that

$$p(t|\mathbf{x}, \mathbf{t}) = \int \mathcal{N}\big(t|\boldsymbol{\phi}(\mathbf{x})^T \mathbf{m}_N, \beta^{-1}\big[1 + \boldsymbol{\phi}(\mathbf{x})^T \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x})\big]\big) \mathrm{Gam}(\beta|a_N, b_N)\, \mathrm{d}\beta$$

□

Exercise 3.14 ⋆⋆

In this exercise, we explore in more detail the properties of the equivalent kernel defined by (3.62), where $\mathbf{S}_N$ is defined by (3.54). Suppose that the basis functions $\phi_j(\mathbf{x})$ are linearly independent and that the number N of data points is greater than the number M of basis functions. Furthermore, let one of the basis functions be constant, say $\phi_0(\mathbf{x}) = 1$. By taking suitable linear combinations of these basis functions, we can construct a new basis set $\psi_j(\mathbf{x})$ spanning the same space but orthonormal, so that

$$\sum_{n=1}^{N} \psi_j(\mathbf{x}_n)\psi_k(\mathbf{x}_n) = \mathbf{I}_{jk} \tag{3.115}$$

where $\mathbf{I}_{jk}$ is defined to be 1 if $j = k$ and 0 otherwise, and we take $\psi_0(\mathbf{x}) = 1$. Show that for $\alpha = 0$, the equivalent kernel can be written as $k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\psi}(\mathbf{x})^T\boldsymbol{\psi}(\mathbf{x}')$ where $\boldsymbol{\psi} = (\psi_0, \ldots, \psi_{M-1})^T$. Use this result to show that the kernel satisfies the summation constraint

$$\sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) = 1 \tag{3.116}$$

Proof. The equivalent kernel is defined by

$$k(\mathbf{x}, \mathbf{x}') = \beta\boldsymbol{\phi}(\mathbf{x})^T\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}') \tag{3.62}$$

where

$$\mathbf{S}_N^{-1} = \alpha\mathbf{I} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi} \tag{3.54}$$

We'll use the newly defined basis set and construct the coresponding *design matrix*, whose elements are given by $\boldsymbol{\Psi}_{nj} = \psi_j(\mathbf{x}_n)$, so that

$$\boldsymbol{\Psi} = \begin{pmatrix} \psi_0(\mathbf{x}_1) & \psi_1(\mathbf{x}_1) & \cdots & \psi_{M-1}(\mathbf{x}_1) \\ \psi_0(\mathbf{x}_2) & \psi_1(\mathbf{x}_2) & \cdots & \psi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_0(\mathbf{x}_N) & \psi_1(\mathbf{x}_N) & \cdots & \psi_{M-1}(\mathbf{x}_N) \end{pmatrix}$$

Since the basis set is orthonormal, we have that

$$\boldsymbol{\Psi}^T\boldsymbol{\Psi} = \sum_{n=1}^{N} \boldsymbol{\psi}(\mathbf{x}_n)\boldsymbol{\psi}(\mathbf{x}_n)^T = \mathbf{I}$$

Now, for $\alpha = 0$, $\mathbf{S}_N$ becomes

$$\mathbf{S}_N = \left(\beta\boldsymbol{\Psi}^T\boldsymbol{\Psi}\right)^{-1} = \frac{1}{\beta}$$

Therefore, by following (3.62) the equivalent kernel can be written as

$$k(\mathbf{x}, \mathbf{x}') = \beta\boldsymbol{\psi}(\mathbf{x})^T\mathbf{S}_N\boldsymbol{\psi}(\mathbf{x}') = \boldsymbol{\psi}(\mathbf{x})^T\boldsymbol{\psi}(\mathbf{x}')$$

Finally, the summation constraint (3.116) obviously holds, since from $\psi_0(\mathbf{x}) = 1$, we have that

$$\sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) = \sum_{n=1}^{N} \boldsymbol{\psi}(\mathbf{x})^T\boldsymbol{\psi}(\mathbf{x}_n) = \sum_{n=1}^{N}\sum_{j=0}^{M-1} \psi_j(\mathbf{x})\psi_j(\mathbf{x}_n) = \sum_{j=0}^{M-1} \psi_j(\mathbf{x}) \sum_{n=1}^{N} \psi_j(\mathbf{x}_n)\psi_0(\mathbf{x}_n)$$

$$= \sum_{j=0}^{M-1} \psi_j(\mathbf{x}) \mathbf{I}_{j+1,1} = 1$$

□

Exercise 3.15 ⋆

Consider a linear basis function model for regression in which the parameters α and β are set using the evidence framework. Show that the function $E(\mathbf{m}_N)$ defined by (3.82) satisfies the relation $2E(\mathbf{m}_N) = N$.

Proof. Our function is given by

$$E(\mathbf{m}_N) = \frac{\beta}{2}||\mathbf{t} - \mathbf{\Phi}\mathbf{m}_N||^2 + \frac{\alpha}{2}\mathbf{m}_N^T\mathbf{m}_N \tag{3.82}$$

We will be using the quantity γ defined in Section 3.5.2 to derive our result. From (3.92) we get that

$$\mathbf{m}_N^T\mathbf{m}_N = \frac{\gamma}{\alpha}$$

and (3.95) gives

$$||\mathbf{t} - \mathbf{\Phi}\mathbf{m}_N||^2 = \frac{N - \gamma}{\beta}$$

Therefore,

$$2E(\mathbf{m}_N) = \beta||\mathbf{t} - \mathbf{\Phi}\mathbf{m}_N||^2 + \alpha\mathbf{m}_N^T\mathbf{m}_N = N - \gamma + \gamma = N$$

□

Exercise 3.16 ⋆⋆

Derive the result (3.86) for the log evidence function $p(\mathbf{t}|\alpha, \beta)$ of the linear regression model by making use of (2.115) to evaluate the integral (3.77) directly.

Proof. The marginal likelihood function is given by the integral

$$p(\mathbf{t}|\alpha, \beta) = \int p(\mathbf{t}|\mathbf{w}, \beta)p(\mathbf{w}|\alpha)\,\mathrm{d}\mathbf{w} \tag{3.77}$$

The first factor under the integral is the likelihood (3.10), while the second factor is given by (3.52). Therefore, the evidence function becomes

$$p(\mathbf{t}|\alpha, \beta) = \int \prod_{n=1}^{N} p(t_n|\mathbf{w}\boldsymbol{\phi}(\mathbf{x}_N), \beta^{-1})\mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})\,\mathrm{d}\mathbf{w}$$

Our aim is to find a proportional Gaussian form for the likelihood term and then use (2.115) to evaluate the integral directly. We've seen in Exercise 3.12 that

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

This can be rewritten as a quadratic form which corresponds to a Gaussian:

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

Therefore,

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

so the evidence function is now given by

$$p(\mathbf{t}|\alpha, \beta) \propto \int \mathcal{N}(\mathbf{t}|\boldsymbol{\Phi}\mathbf{w}, \beta^{-1}\mathbf{I})\mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})\, \mathrm{d}\mathbf{w}$$

Since the integral involves the convolution of two Gaussians, by following the notation used in (2.113) and (2.114), we'd have that

$$\mathbf{x} = \mathbf{w} \qquad \mathbf{y} = \mathbf{t} \qquad \boldsymbol{\mu} = \mathbf{0} \qquad \boldsymbol{\Lambda}^{-1} = \alpha\mathbf{I} \qquad \mathbf{A} = \boldsymbol{\Phi} \qquad \mathbf{b} = 0 \qquad \mathbf{L}^{-1} = \beta\mathbf{I}$$

Applying (2.115) yields the Gaussian form of the evidence function:

$$p(\mathbf{t}|\alpha, \beta) \propto \mathcal{N}(\mathbf{t}|\mathbf{0}, \beta^{-1}\mathbf{I} + \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^T)$$

By applying the Woodbury identity (C.7) with

$$\mathbf{A} = \beta^{-1}\mathbf{I} \qquad \mathbf{B} = \boldsymbol{\Phi} \qquad \mathbf{C} = \alpha^{-1}\mathbf{I} \qquad \mathbf{D} = \boldsymbol{\Phi}^T$$

the precision matrix of this Gaussian will be given by

$$\begin{aligned}(\beta^{-1}\mathbf{I} + \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^T)^{-1} &= \beta\mathbf{I} - \beta^2\boldsymbol{\Phi}(\alpha\mathbf{I} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T \\ &= \beta\mathbf{I} - \beta^2\boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^T\end{aligned}$$

where $\mathbf{A}$ is given by (3.81). Hence, by using (3.81) and (3.84), we obtain that the quadratic term in the exponential of the Gaussian has the form:

$$\begin{aligned}-\frac{1}{2}\mathbf{t}^T(\beta^{-1}\mathbf{I} + \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^T)^{-1}\mathbf{t} &= -\frac{1}{2}\mathbf{t}^T(\beta\mathbf{I} - \beta^2\boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^T)^{-1}\mathbf{t} \\ &= -\frac{\beta}{2}\mathbf{t}^T\mathbf{t} + \frac{\beta}{2}\mathbf{t}^T\boldsymbol{\Phi}\mathbf{m}_N \\ &= -\frac{\beta}{2}\mathbf{t}^T\mathbf{t} + \frac{\beta}{2}\mathbf{m}_N^T\mathbf{A}\mathbf{m}_N \\ &= -\frac{\beta}{2}\mathbf{t}^T\mathbf{t} + \beta\mathbf{m}_N^T\mathbf{A}\mathbf{m}_N - \frac{\beta}{2}\mathbf{m}_N^T\mathbf{A}\mathbf{m}_N \\ &= -\frac{\beta}{2}\mathbf{t}^T\mathbf{t} + \beta\mathbf{m}_N^T\mathbf{A}\mathbf{m}_N - \frac{\alpha}{2}\mathbf{m}_N^T\mathbf{m}_N - \frac{\beta}{2}\mathbf{m}_N\boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{m}_N\end{aligned}$$

As seen in Exercise 3.18, this is actually equal to $-E(\mathbf{m}_N)$, so

$$-\frac{1}{2}\mathbf{t}^T(\beta^{-1}\mathbf{I} + \alpha^{-1}\mathbf{\Phi}\mathbf{\Phi}^T)^{-1}\mathbf{t} = -E(\mathbf{m}_N)$$

Now, since $\mathbf{\Phi}$ is a $N \times M$ matrix, we apply (C.14) and have that:

$$\begin{aligned}
|\beta^{-1}\mathbf{I}_N + \alpha^{-1}\mathbf{\Phi}\mathbf{\Phi}^T| &= \beta^{-N}\left|\mathbf{I}_N + \frac{\beta}{\alpha}\mathbf{\Phi}\mathbf{\Phi}^T\right| \\
&= \beta^{-N}\left|\mathbf{I}_M + \frac{\beta}{\alpha}\mathbf{\Phi}^T\mathbf{\Phi}\right| \\
&= \alpha^{-M}\beta^{-N}|\alpha\mathbf{I}_M + \beta\mathbf{\Phi}^T\mathbf{\Phi}| \\
&= \alpha^{-M}\beta^{-N}|\mathbf{A}|
\end{aligned}$$

Threfore, we finally expand the Gaussian form of the evidence function to obtain:

$$\begin{aligned}
p(\mathbf{t}|\alpha,\beta) &\propto \mathcal{N}(\mathbf{t}|\mathbf{0}, \beta^{-1}\mathbf{I} + \alpha^{-1}\mathbf{\Phi}\mathbf{\Phi}^T) \\
&\propto \frac{1}{(2\pi)^{N/2}}\frac{1}{|\beta^{-1}\mathbf{I}_N + \alpha^{-1}\mathbf{\Phi}\mathbf{\Phi}^T|^{1/2}}\exp\left\{-\frac{1}{2}\mathbf{t}^T(\beta^{-1}\mathbf{I} + \alpha^{-1}\mathbf{\Phi}\mathbf{\Phi}^T)^{-1}\mathbf{t}\right\} \\
&\propto \frac{1}{(2\pi)^{N/2}}\alpha^{-M/2}\beta^{-N/2}|A|^{-1/2}\exp\{-E(\mathbf{m}_N)\}
\end{aligned}$$

Hence, we easily derive the log marginal likelihood as

$$\ln p(\mathbf{t}|\alpha,\beta) = -\frac{N}{2}\ln(2\pi) + \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - \frac{1}{2}\ln|\mathbf{A}| - E(\mathbf{m}_N) + \text{const} \tag{3.86}$$

□

Exercise 3.17 ⋆

Show that the evidence function for the Bayesian linear regression model can be written in the form (3.78) in which $E(\mathbf{w})$ is defined by (3.79).

Proof. The log likelihood is given by

$$\ln p(\mathbf{t}|\mathbf{w},\beta) = \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta E_D(\mathbf{w}) \tag{3.11}$$

By applying the exponential function on both sides of the expression, we obtain that

$$p(\mathbf{t}|\mathbf{w},\beta) = \left(\frac{\beta}{2\pi}\right)^{N/2}\exp\{-\beta E_D(\mathbf{w})\}$$

We continue by expanding the Gaussian

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) \tag{3.52}$$

to get that

$$p(\mathbf{w}|\alpha) = \left(\frac{\alpha}{2\pi}\right)^{M/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\right\} = \left(\frac{\alpha}{2\pi}\right)^{M/2} \exp\{-\alpha E_W(\mathbf{w})\}$$

Therefore, by replacing into (3.77), we obtain

$$\begin{aligned} p(\mathbf{t}|\alpha,\beta) = \int p(\mathbf{t}|\mathbf{w},\beta)p(\mathbf{w}|\alpha)\,\mathrm{d}\mathbf{w} &= \left(\frac{\beta}{2\pi}\right)^{N/2}\left(\frac{\alpha}{2\pi}\right)^{M/2}\int \exp\{-\alpha E_W(\mathbf{w}) - \beta E_D(\mathbf{w})\}\,\mathrm{d}w \\ &= \left(\frac{\beta}{2\pi}\right)^{N/2}\left(\frac{\alpha}{2\pi}\right)^{M/2}\int \exp\{-E(\mathbf{w})\}\,\mathrm{d}\mathbf{w} \end{aligned} \tag{3.78}$$

where

$$E(\mathbf{w}) = \beta E_D(\mathbf{w}) + \alpha E_W(\mathbf{w}) = \frac{\beta}{2}||\mathbf{t} - \mathbf{\Phi}\mathbf{w}||^2 + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} \tag{3.79}$$

□

Exercise 3.18 ⋆⋆

By completing the square over $\mathbf{w}$, show that the error function (3.79) in Bayesian linear regression can be written in the form (3.80).

Proof. Our first step is expanding $E(\mathbf{w})$:

$$\begin{aligned} E(\mathbf{w}) &= \frac{\beta}{2}||\mathbf{t} - \mathbf{\Phi}\mathbf{w}||^2 + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} \\ &= \frac{\beta}{2}\mathbf{t}^T\mathbf{t} - \frac{\beta}{2}\mathbf{t}^T\mathbf{\Phi}\mathbf{w} - \frac{\beta}{2}\mathbf{w}^T\mathbf{\Phi}^T\mathbf{t} + \frac{\beta}{2}\mathbf{w}^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{w} + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} \end{aligned} \tag{3.79}$$

We continue by doing the same for $E(\mathbf{m}_N)$ and obtain that:

$$\begin{aligned} E(\mathbf{m}_N) &= \frac{\beta}{2}||\mathbf{t} - \mathbf{\Phi}\mathbf{m}_N||^2 + \frac{\alpha}{2}\mathbf{m}_N^T\mathbf{m}_N \\ &= \frac{\beta}{2}\mathbf{t}^T\mathbf{t} - \frac{\beta}{2}\mathbf{t}^T\mathbf{\Phi}\mathbf{m}_N - \frac{\beta}{2}\mathbf{m}_N^T\mathbf{\Phi}^T\mathbf{t} + \frac{\beta}{2}\mathbf{m}_N^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{m}_N + \frac{\alpha}{2}\mathbf{m}_N^T\mathbf{m}_N \end{aligned}$$

$\mathbf{A}$ is a Hessian matrix, so it's symmetric. By using this and the expressions

$$\mathbf{A} = \alpha\mathbf{I} + \beta\mathbf{\Phi}^T\mathbf{\Phi} \tag{3.81}$$

$$\mathbf{m}_N = \beta\mathbf{A}^{-1}\mathbf{\Phi}^T\mathbf{t} \tag{3.84}$$

We notice that the negative terms in the expansions can be written as

$$-\frac{\beta}{2}\mathbf{t}^T\mathbf{\Phi}\mathbf{w} - \frac{\beta}{2}\mathbf{w}^T\mathbf{\Phi}\mathbf{t} = -\frac{1}{2}\mathbf{m}_N^T\mathbf{A}\mathbf{w} - \frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{m}_N = -\mathbf{m}_N^T\mathbf{A}\mathbf{w}$$

$$-\frac{\beta}{2}\mathbf{t}^T\mathbf{\Phi}\mathbf{m}_N - \frac{\beta}{2}\mathbf{m}_N^T\mathbf{\Phi}\mathbf{t} = -\frac{1}{2}\mathbf{m}_N^T\mathbf{A}\mathbf{m}_N - \frac{1}{2}\mathbf{m}_N^T\mathbf{A}\mathbf{m}_N = -\mathbf{m}_N^T\mathbf{A}\mathbf{m}_N$$

Hence,

$$E(\mathbf{w}) = \frac{\beta}{2}\mathbf{t}^T\mathbf{t} - \mathbf{m}_N^T\mathbf{A}\mathbf{w} + \frac{\beta}{2}\mathbf{w}^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{w} + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}$$

$$E(\mathbf{m}_N) = \frac{\beta}{2}\mathbf{t}^T\mathbf{t} - \mathbf{m}_N^T\mathbf{A}\mathbf{m}_N + \frac{\beta}{2}\mathbf{m}_N^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{m}_N + \frac{\alpha}{2}\mathbf{m}_N^T\mathbf{m}_N$$

By taking the difference of the error functions and then repeatedly making use of (3.81), we reach a point when we can complete the square:

$$\begin{aligned}
E(\mathbf{w}) - E(\mathbf{m}_N) &= -\mathbf{m}_N^T\mathbf{A}\mathbf{w} + \mathbf{m}_N^T\mathbf{A}\mathbf{m}_N + \frac{\beta}{2}\mathbf{w}^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{w} - \frac{\beta}{2}\mathbf{m}_N^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{m}_N + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} - \frac{\alpha}{2}\mathbf{m}_N^T\mathbf{m}_N \\
&= -\mathbf{m}_N^T\mathbf{A}\mathbf{w} + \mathbf{m}_N^T\mathbf{A}\mathbf{m}_N + \frac{1}{2}\mathbf{w}^T\big(\alpha\mathbf{I} + \beta\mathbf{\Phi}^T\mathbf{\Phi}\big)\mathbf{w} - \mathbf{m}_N^T\big(\alpha\mathbf{I} + \beta\mathbf{\Phi}^T\mathbf{\Phi}\big)\mathbf{m}_N \\
&= -\mathbf{m}_N^T\mathbf{A}\mathbf{w} + \mathbf{m}_N^T\mathbf{A}\mathbf{m}_N + \frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w} - \frac{1}{2}\mathbf{m}_N^T\mathbf{A}\mathbf{m}_N \\
&= \frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w} - \frac{1}{2}\mathbf{m}_N^T\mathbf{A}\mathbf{w} - \frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{m}_N + \frac{1}{2}\mathbf{m}_N^T\mathbf{A}\mathbf{m}_N \\
&= \frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^T\mathbf{A}(\mathbf{w} - \mathbf{m}_N)
\end{aligned}$$

which directly proves that (3.79) can be written as

$$E(\mathbf{w}) = E(\mathbf{m}_N) + \frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^T\mathbf{A}(\mathbf{w} - \mathbf{m}_N) \tag{3.80}$$

□

Exercise 3.19 ⋆⋆

Show that the integration over $\mathbf{w}$ in the Bayesian linear regression model gives the result (3.85). Hence show that the log marginal likelihood is given by (3.86).

Proof. We start by rewriting $E(\mathbf{w})$ like in (3.80) and obtain that

$$\begin{aligned}
\int \exp\{-E(\mathbf{w})\}\,\mathrm{d}\mathbf{w} &= \int \exp\left\{ -E(\mathbf{m}_N) - \frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^T\mathbf{A}(\mathbf{w} - \mathbf{m}_N)\right\}\mathrm{d}\mathbf{w} \\
&= \int \exp\{-E(\mathbf{m}_N)\}\exp\left\{ -\frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^T\mathbf{A}(\mathbf{w} - \mathbf{m}_N)\right\}\mathrm{d}\mathbf{w} \\
&= \exp\{-E(\mathbf{m}_N)\}\int \exp\left\{ -\frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^T\mathbf{A}(\mathbf{w} - \mathbf{m}_N)\right\}\mathrm{d}\mathbf{w}
\end{aligned}$$

The integral is easily solved by noticing that the quadratic term under the exponential term corresponds to a Gaussian of the form $\mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{A}^{-1})$. Because the Gaussian distribution is normalized, we then have that

$$\begin{aligned}
&\int \exp\{-E(\mathbf{w})\}\,\mathrm{d}\mathbf{w} \\
&\quad= \exp\{-E(\mathbf{m}_N)\}(2\pi)^{M/2}|\mathbf{A}|^{-1/2}\int \frac{1}{(2\pi)^{M/2}}\frac{1}{|\mathbf{A}|^{-1/2}}\exp\left\{\frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^T\mathbf{A}(\mathbf{w} - \mathbf{m}_N)\right\}\mathrm{d}\mathbf{w}
\end{aligned}$$

$$= \exp\{-E(\mathbf{m}_N)\}(2\pi)^{M/2}|\mathbf{A}|^{-1/2} \int \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{A}^{-1})\,\mathrm{d}\mathbf{w}$$
$$= \exp\{-E(\mathbf{m}_N)\}(2\pi)^{M/2}|\mathbf{A}|^{-1/2} \tag{3.85}$$

By substituting this into (3.78), the evidence function becomes

$$p(\mathbf{t}|\alpha,\beta) = \left(\frac{\beta}{2\pi}\right)^{N/2} \alpha^{M/2}|\mathbf{A}|^{-1/2} \exp\{-E(\mathbf{m}_N)\}$$

Hence, the log marginal likelihood is given by

$$\ln p(\mathbf{t}|\alpha,\beta) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - E(\mathbf{m}_N) - \frac{1}{2}\ln|\mathbf{A}| - \frac{N}{2}\ln(2\pi) \tag{3.86}$$

□

Exercise 3.20 ⋆⋆

Verify all of the steps needed to show that maximization of the log marginal likelihood function (3.86) with respect to α leads to the re-estimation equation (3.92).

Proof. The steps taken in the maximization of the (3.86) are quite straightforward, so this proof will be very similar to what's in the book. By defining the eigenvector equation

$$(\beta\mathbf{\Phi}^T\mathbf{\Phi})\mathbf{u}_i = \lambda_i\mathbf{u}_i \tag{3.87}$$

we'd have that

$$\mathbf{A}\mathbf{u}_i = (\alpha\mathbf{I} + \beta\mathbf{\Phi}^T\mathbf{\Phi})\mathbf{u}_i = \alpha\mathbf{u}_i + (\beta\mathbf{\Phi}^T\mathbf{\Phi})\mathbf{u}_i = (\alpha+\lambda_i)\mathbf{u}_i$$

which shows that the eigenvalues of $\mathbf{A}$ are $\alpha + \lambda_i$, where $\mathbf{A}$ is given by (3.81). Now, since the determinant of a matrix is the product of its eigenvalues, we have that

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\ln|\mathbf{A}| = \frac{\mathrm{d}}{\mathrm{d}\alpha}\ln\prod_i(\lambda_i+\alpha) = \frac{\mathrm{d}}{\mathrm{d}\alpha}\sum_i\ln(\lambda_i+\alpha) = \sum_i\frac{1}{\lambda_i+\alpha} \tag{3.88}$$

The derivative of (3.86) with respect to α is given by

$$\frac{\partial}{\partial\alpha}\ln p(\mathbf{t}|\alpha,\beta) = \frac{M}{2\alpha} - \frac{1}{2}\mathbf{m}_N^T\mathbf{m}_N - \frac{1}{2}\sum_i\frac{1}{\lambda_i+\alpha}$$

so the stationary points will satisfy

$$0 = \frac{M}{2\alpha} - \frac{1}{2}\mathbf{m}_N^T\mathbf{m}_N - \frac{1}{2}\sum_i\frac{1}{\lambda_i+\alpha} \tag{3.89}$$

Multiplying by 2α and then rearranging, we have that

$$\alpha\mathbf{m}^T\mathbf{m}_N = M - \alpha\sum_i\frac{1}{\lambda_i+\alpha} = \sum_i\left(1 - \frac{\alpha}{\lambda_i+\alpha}\right) = \sum_i\frac{\lambda_i}{\alpha+\lambda_i} = \gamma$$

Therefore, the value of α that maximizes the marginal likelihood is given by

$$\alpha = \frac{\gamma}{\mathbf{m}_N^T \mathbf{m}_N} \tag{3.92}$$

where γ is defined by

$$\gamma = \sum_i \frac{\lambda_i}{\alpha + \lambda_i} \tag{3.91}$$

□

Exercise 3.21 ⋆⋆

An alternative way to derive the result (3.92) for the optimal value of α in the evidence framework is to make use of the identity (note that we changed the variables $\mathbf{A}$ and α for $\mathbf{D}$ and δ so that there is no confusion with the variables used in our framework)

$$\frac{\mathrm{d}}{\mathrm{d}\delta} \ln |\mathbf{D}| = \mathrm{Tr}\left(\mathbf{D}^{-1} \frac{\mathrm{d}}{\mathrm{d}\delta} \mathbf{D} \right) \tag{3.117}$$

Prove this identity by considering the eigenvalue expansion of a real symmetric matrix $\mathbf{D}$, and making use of the standard results for the determinant and trace of $\mathbf{D}$ expressed in terms of its eigenvalues (Appendix C). Then make use of (3.117) to derive (3.92) starting from (3.86).

Proof. Since $\mathbf{D}$ is symmetric, we can consider the N eigenvector equations

$$\mathbf{D}\mathbf{u}_i = \omega_i \mathbf{u}_i$$

where the eigenvectors $\mathbf{u}_i$ were chosen to be orthonormal, as seen in Appendix C. By using (C.47) to rewrite the determinant, the left-hand side of (3.117) becomes

$$\frac{\mathrm{d}}{\mathrm{d}\delta} \ln |\mathbf{D}| = \frac{\mathrm{d}}{\mathrm{d}\delta} \ln \prod_{i=1}^{N} \omega_i = \frac{\mathrm{d}}{\mathrm{d}\delta} \sum_{i=1}^{N} \ln \omega_i = \sum_{i=1}^{N} \frac{\mathrm{d}}{\mathrm{d}\delta} \ln \omega_i = \sum_{i=1}^{N} \frac{1}{\omega_i} \frac{\mathrm{d}}{\mathrm{d}\delta} \omega_i$$

Now, by considering the eigenvalue expansions given by (C.45) and (C.46),

$$\mathbf{D} = \sum_{i=1}^{N} \omega_i \mathbf{u}_i \mathbf{u}_i^T \qquad\qquad \mathbf{D}^{-1} = \sum_{i=1}^{N} \frac{1}{\omega_i} \mathbf{u}_i \mathbf{u}_i^T$$

we can rewrite the term inside the trace operator as:

$$\begin{aligned}
\mathbf{D}^{-1} \frac{\mathrm{d}}{\mathrm{d}\delta} \mathbf{D} &= \left(\sum_{i=1}^{N} \frac{1}{\omega_i} \mathbf{u}_i \mathbf{u}_i^T \right) \left(\frac{\mathrm{d}}{\mathrm{d}\delta} \sum_{j=1}^{N} \omega_i \mathbf{u}_i \mathbf{u}_i^T \right) \\
&= \left(\sum_{i=1}^{N} \frac{1}{\omega_i} \mathbf{u}_i \mathbf{u}_i^T \right) \sum_{j=1}^{N} \frac{\mathrm{d}}{\mathrm{d}\delta} (\omega_j \mathbf{u}_j \mathbf{u}_j^T) \\
&= \left(\sum_{i=1}^{N} \frac{1}{\omega_i} \mathbf{u}_i \mathbf{u}_i^T \right) \sum_{j=1}^{N} \left[\left(\frac{\mathrm{d}}{\mathrm{d}\delta} \omega_j \right) \mathbf{u}_j \mathbf{u}_j^T + \omega_j \left(\frac{\mathrm{d}}{\mathrm{d}\delta} \mathbf{u}_j \right) \mathbf{u}_j^T + \omega_i \mathbf{u}_i \left(\frac{\mathrm{d}}{\mathrm{d}\delta} \mathbf{u}_j^T \right) \right]
\end{aligned}$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{\omega_i}\left(\frac{\mathrm{d}}{\mathrm{d}\delta}\omega_j\right)\mathbf{u}_i\mathbf{u}_i^T\mathbf{u}_j\mathbf{u}_j^T + \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\omega_j}{\omega_i}\mathbf{u}_i\mathbf{u}_i^T\left[\left(\frac{\mathrm{d}}{\mathrm{d}\delta}\mathbf{u}_j\right)\mathbf{u}_j^T + \mathbf{u}_j\left(\frac{\mathrm{d}}{\mathrm{d}\delta}\mathbf{u}_j^T\right)\right]$$

We'll analyze the two sum terms separately. Since the eigenvectors were chosen to be orthonormal, we have that

$$\mathbf{u}_i^T\mathbf{u}_j = \mathbf{I}_{ij} \tag{C.33}$$

Therefore, we'll only have to keep the terms for which $i = j$ in the first sum, so

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{\omega_i}\left(\frac{\mathrm{d}}{\mathrm{d}\delta}\omega_j\right)\mathbf{u}_i\mathbf{u}_i^T\mathbf{u}_j\mathbf{u}_j^T = \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{\omega_i}\left(\frac{\mathrm{d}}{\mathrm{d}\delta}\omega_j\right)\mathbf{u}_i\mathbf{I}_{ij}\mathbf{u}_j^T = \sum_{i=1}^{N}\frac{1}{\omega_i}\left(\frac{\mathrm{d}}{\mathrm{d}\delta}\omega_j\right)\mathbf{u}_i\mathbf{u_i}^T$$

Notice that the trace of this term is actually the left-hand side of (3.117), so we'll continue by proving that the second sum term has a trace of 0. The second sum term can then be expanded as

$$\begin{aligned}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\omega_j}{\omega_i}\mathbf{u}_i\mathbf{u}_i^T\left[\left(\frac{\mathrm{d}}{\mathrm{d}\delta}\mathbf{u}_j\right)\mathbf{u}_j^T + \mathbf{u}_j\left(\frac{\mathrm{d}}{\mathrm{d}\delta}\mathbf{u}_j^T\right)\right] &= \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{2\omega_j}{\omega_i}\mathbf{u}_i\mathbf{u}_i^T\mathbf{u}_j\left(\frac{\mathrm{d}}{\mathrm{d}\delta}\mathbf{u}_j^T\right) = \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{2\omega_j}{\omega_i}\mathbf{u}_i\mathbf{I}_{ij}\left(\frac{\mathrm{d}}{\mathrm{d}\delta}\mathbf{u}_j^T\right)\\ &= \sum_{i=1}^{N}2\mathbf{u}_i\left(\frac{\mathrm{d}}{\mathrm{d}\delta}\mathbf{u}_i^T\right) = \sum_{i=1}^{N}\left[\left(\frac{\mathrm{d}}{\mathrm{d}\delta}\mathbf{u}_i\right)\mathbf{u}_i^T + \mathbf{u}_i\left(\frac{\mathrm{d}}{\mathrm{d}\delta}\mathbf{u}_i^T\right)\right]\\ &= \sum_{i=1}^{N}\frac{\mathrm{d}}{\mathrm{d}\delta}\left(\mathbf{u}_i\mathbf{u}_i^T\right) = \frac{\mathrm{d}}{\mathrm{d}\delta}\sum_{i=1}^{N}\mathbf{u}_i\mathbf{u}_i^T = \frac{\mathrm{d}}{\mathrm{d}\delta}\mathbf{I} = \mathbf{0}_N\end{aligned}$$

Finally, we have that

$$\begin{aligned}\mathrm{Tr}\left(\mathbf{D}^{-1}\frac{\mathrm{d}}{\mathrm{d}\delta}\mathbf{D}\right) &= \mathrm{Tr}\left\{\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{\omega_i}\left(\frac{\mathrm{d}}{\mathrm{d}\delta}\omega_j\right)\mathbf{u}_i\mathbf{u}_i^T\mathbf{u}_j\mathbf{u}_j^T + \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\omega_j}{\omega_i}\mathbf{u}_i\mathbf{u}_i^T\left[\left(\frac{\mathrm{d}}{\mathrm{d}\delta}\mathbf{u}_j\right)\mathbf{u}_j^T + \mathbf{u}_j\left(\frac{\mathrm{d}}{\mathrm{d}\delta}\mathbf{u}_j^T\right)\right]\right\}\\ &= \mathrm{Tr}\left\{\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{\omega_i}\left(\frac{\mathrm{d}}{\mathrm{d}\delta}\omega_j\right)\mathbf{u}_i\mathbf{u}_i^T\mathbf{u}_j\mathbf{u}_j^T\right\} + \mathrm{Tr}\left\{\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\omega_j}{\omega_i}\mathbf{u}_i\mathbf{u}_i^T\left[\left(\frac{\mathrm{d}}{\mathrm{d}\delta}\mathbf{u}_j\right)\mathbf{u}_j^T + \mathbf{u}_j\left(\frac{\mathrm{d}}{\mathrm{d}\delta}\mathbf{u}_j^T\right)\right]\right\}\\ &= \mathrm{Tr}\left[\sum_{i=1}^{N}\frac{1}{\omega_i}\left(\frac{\mathrm{d}}{\mathrm{d}\delta}\omega_i\right)\mathbf{u}_i\mathbf{u}_i^T\right] + \mathrm{Tr}(\mathbf{0}_N)\\ &= \sum_{i=1}^{N}\frac{1}{\omega_i}\left(\frac{\mathrm{d}}{\mathrm{d}\delta}\omega_i\right)\\ &= \frac{\mathrm{d}}{\mathrm{d}\delta}\ln|\mathbf{D}|\end{aligned} \tag{3.117}$$

which proves the needed identity. We can derive (3.92) by following exactly the same steps as in Exercise 3.20 or in the book, except that now we compute (3.88) by using the previously proven identity and (C.47). Note that $\mathbf{A}$ is given by (3.81) and it has the eigenvalues $\lambda_i + \alpha$, so

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\ln|\mathbf{A}| = \mathrm{Tr}\left(\mathbf{A}^{-1}\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathbf{A}\right) = \mathrm{Tr}\left[\mathbf{A}^{-1}\frac{\mathrm{d}}{\mathrm{d}\alpha}\left(\alpha\mathbf{I} + \beta\mathbf{\Phi}^T\mathbf{\Phi}\right)\right] = \mathrm{Tr}(\mathbf{A}^{-1}) = \sum_i\frac{1}{\lambda_i + \alpha}$$

□

Exercise 3.22 ⋆⋆

Starting from (3.86) verify all of the steps needed to show that maximization of the log marginal likelihood function (3.86) with respect to β leads to the re-estimation equation (3.95).

Proof. As in Exercise 3.20, one could prove that $\mathbf{A}$ has the eigenvalues $\lambda_i + \alpha$. Also, from (3.87) the eigenvalues are proportional to β, so $d\lambda_i/d\beta = \lambda_i/\beta$, so

$$\frac{\mathrm{d}}{\mathrm{d}\beta} \ln |\mathbf{A}| = \frac{\mathrm{d}}{\mathrm{d}\beta} \sum_i \ln(\lambda_i + \alpha) = \frac{1}{\beta} \sum_i \frac{\lambda_i}{\lambda_i + \alpha} = \frac{\gamma}{\beta} \tag{3.93}$$

where γ is given by (3.91). Therefore, the derivative of (3.87) with respect to β is given by

$$\frac{\partial}{\partial \beta} p(\mathbf{t}|\alpha, \beta) = \frac{N - \gamma}{2\beta} - \frac{1}{2} \sum_{n=1}^{N} \{t_n - \mathbf{m}_N^T \boldsymbol{\phi}(\mathbf{x}_n)\}^2$$

and the critical points satisfy

$$\frac{N - \gamma}{2\beta} = \frac{1}{2} \sum_{n=1}^{N} \{t_n - \mathbf{m}_N^T \boldsymbol{\phi}(\mathbf{x}_n)\}^2$$

which after multiplying both sides by $2/(N - \gamma)$ gives us the re-estimation equation

$$\frac{1}{\beta} = \frac{1}{N - \gamma} \sum_{n=1}^{N} \{t_n - \mathbf{m}_N^T \boldsymbol{\phi}(\mathbf{x}_n)\}^2 \tag{3.95}$$

□

Exercise 3.23 ⋆⋆

Show that the marginal probability of the data, in other words the model evidence, for the model described in Exercise 3.12 is given by

$$p(\mathbf{t}) = \frac{1}{(2\pi)^{N/2}} \frac{b_0^{a_0}}{b_N^{a_N}} \frac{\Gamma(a_N)}{\Gamma(a_0)} \frac{|\mathbf{S}_N|^{1/2}}{|\mathbf{S}_0|^{1/2}} \tag{3.118}$$

by first marginalizing with respect to $\mathbf{w}$ and then with respect to β.

Proof. By marginalizing with respect to β and then with respect to β, the model evidence will be given by

$$p(\mathbf{t}) = \iint p(\mathbf{w}, \beta) p(\mathbf{t}|\mathbf{w}, \beta) \, \mathrm{d}\mathbf{w} \, \mathrm{d}\beta$$

The first factor unde the integral is the prior given by (3.112), while the second factor is the likelihood (3.11). We proved in Exercise 3.16 that the likelihood is proportional to $\mathcal{N}(\mathbf{t}|\boldsymbol{\Phi}\mathbf{w}, \beta^{-1}\mathbf{I})$, so the marginal probability becomes

$$p(\mathbf{t}) = \iint \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \beta^{-1}\mathbf{S}_0) \mathrm{Gam}(\beta|a_0, b_0) \mathcal{N}(\mathbf{t}|\boldsymbol{\Phi}\mathbf{w}, \beta^{-1}\mathbf{I}) \, \mathrm{d}\mathbf{w} \, \mathrm{d}\beta$$

$$= \iint \text{Gam}(\beta|a_0, b_0)\mathcal{N}(\mathbf{w}|\mathbf{m}_0, \beta^{-1}\mathbf{S}_0)\mathcal{N}(\mathbf{t}|\mathbf{\Phi w}, \beta^{-1}\mathbf{I})\,\mathrm{d}\mathbf{w}\,\mathrm{d}\beta$$

By expanding the three distributions, we have that

$$p(\mathbf{t}) = \frac{1}{(2\pi)^{\frac{N+M}{2}}}\frac{b_0^{a_0}}{\Gamma(a_0)}\frac{1}{|\mathbf{S}_0|^{1/2}} \iint \beta^{a_0-1+N/2+M/2}\exp\{\beta b_0\}\exp\Big\{-\frac{\beta}{2}||\mathbf{t}-\mathbf{\Phi w}||^2\Big\}\exp\Big\{-\frac{\beta}{2}(\mathbf{w}-\mathbf{m}_0)^T\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0)\Big\}\,\mathrm{d}\mathbf{w}\,\mathrm{d}\beta$$

Let's expand the term under the $\mathbf{w}$ integral, and then use (3.50) and (3.51) to complete the square:

$$\begin{aligned}
&\exp\Big\{-\frac{\beta}{2}||\mathbf{t}-\mathbf{\Phi w}||^2\Big\}\exp\Big\{-\frac{\beta}{2}(\mathbf{w}-\mathbf{m}_0)^T\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0)\Big\} \\
&\quad= \exp\Big\{-\frac{\beta}{2}||\mathbf{t}-\mathbf{\Phi w}||^2 - \frac{\beta}{2}(\mathbf{w}-\mathbf{m}_0)^T\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0)\Big\} \\
&\quad= \exp\Big\{-\frac{\beta}{2}\big(\mathbf{t}^T\mathbf{t} - 2\mathbf{w}^T\mathbf{\Phi}^T\mathbf{t} + \mathbf{w}^T\mathbf{\Phi}^T\mathbf{\Phi w} + \mathbf{w}^T\mathbf{S}_0^{-1}\mathbf{w} - 2\mathbf{w}^T\mathbf{S}_0^{-1}\mathbf{m}_0 + \mathbf{m}_0^T\mathbf{S}_0^{-1}\mathbf{m}_0\big)\Big\} \\
&\quad= \exp\Big\{-\frac{\beta}{2}\big[\mathbf{w}^T\big(\mathbf{S}_0^{-1} + \mathbf{\Phi}^T\mathbf{\Phi}\big)\mathbf{w} - 2\mathbf{w}^T\big(\mathbf{S}_0^{-1}\mathbf{m}_0 + \mathbf{\Phi}^T\mathbf{t}\big) + \mathbf{t}^T\mathbf{t} + \mathbf{m}_0^T\mathbf{S}_0^{-1}\mathbf{m}_0\big]\Big\} \\
&\quad= \exp\Big\{-\frac{\beta}{2}\big(\mathbf{w}^T\mathbf{S}_N^{-1}\mathbf{w} - 2\mathbf{w}^T\mathbf{S}_N^{-1}\mathbf{m}_N + \mathbf{t}^T\mathbf{t} + \mathbf{m}_0^T\mathbf{S}_0^{-1}\mathbf{m}_0\big)\Big\} \\
&\quad= \exp\Big\{-\frac{\beta}{2}\big[(\mathbf{w}-\mathbf{m}_N)^T\mathbf{S}_N^{-1}(\mathbf{w}-\mathbf{m}_N) - \mathbf{m}_N^T\mathbf{S}_N^{-1}\mathbf{m}_N + \mathbf{t}^T\mathbf{t} + \mathbf{m}_0^T\mathbf{S}_0^{-1}\mathbf{m}_0\big]\Big\}
\end{aligned}$$

We can rewrite (3.12.1) as

$$b_0 = b_N - \frac{1}{2}\mathbf{t}^T\mathbf{t} - \frac{1}{2}\mathbf{m}_0^T\mathbf{S}_0^{-1}\mathbf{m}_0 + \frac{1}{2}\mathbf{m}_N^T\mathbf{S}_N^{-1}\mathbf{m}_N$$

Therefore,

$$\begin{aligned}
&\exp\{\beta b_0\}\exp\Big\{-\frac{\beta}{2}||\mathbf{t}-\mathbf{\Phi w}||^2\Big\}\exp\Big\{-\frac{\beta}{2}(\mathbf{w}-\mathbf{m}_0)^T\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0)\Big\} \\
&\qquad\qquad= \exp\{\beta b_N\}\exp\Big\{-\frac{\beta}{2}(\mathbf{w}-\mathbf{m}_N)^T\mathbf{S}_N^{-1}(\mathbf{w}-\mathbf{m}_N)\Big\}
\end{aligned}$$

and since both the Gaussian and Gamma distributions are normalized, the marginal probability finally becomes what we wanted:

$$\begin{aligned}
p(\mathbf{t}) &= \frac{1}{(2\pi)^{\frac{N+M}{2}}}\frac{b_0^{a_0}}{\Gamma(a_0)}\frac{1}{|\mathbf{S}_0|^{1/2}}\int \beta^{a_0-1+N/2+M/2}\int \exp\{\beta b_N\}\exp\Big\{-\frac{\beta}{2}(\mathbf{w}-\mathbf{m}_N)^T\mathbf{S}_N^{-1}(\mathbf{w}-\mathbf{m}_N)\Big\}\,\mathrm{d}\mathbf{w}\,\mathrm{d}\beta \\
&= \frac{1}{(2\pi)^{\frac{N+M}{2}}}\frac{b_0^{a_0}}{\Gamma(a_0)}\frac{1}{|\mathbf{S}_0|^{1/2}}\int \beta^{a_0-1+N/2+M/2}\int \left(\frac{2\pi}{\beta}\right)^{M/2}|\mathbf{S}_N|^{1/2}\exp\{\beta b_N\}\mathcal{N}(\mathbf{w}|\mathbf{m}_N, \beta^{-1}\mathbf{S}_N)\,\mathrm{d}\mathbf{w}\,\mathrm{d}\beta \\
&= \frac{1}{(2\pi)^{\frac{N}{2}}}\frac{b_0^{a_0}}{\Gamma(a_0)}\frac{|\mathbf{S}_N|^{1/2}}{|\mathbf{S}_0|^{1/2}}\int \beta^{a_0-1+N/2}\exp\{\beta b_N\}\int \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \beta^{-1}\mathbf{S}_N)\,\mathrm{d}\mathbf{w}\,\mathrm{d}\beta
\end{aligned}$$

$$= \frac{1}{(2\pi)^{\frac{N}{2}}} \frac{b_0^{a_0}}{\Gamma(a_0)} \frac{|\mathbf{S}_N|^{1/2}}{|\mathbf{S}_0|^{1/2}} \int \beta^{a_0-1+N/2} \exp\{\beta b_N\} \, \mathrm{d}\beta$$
$$= \frac{1}{(2\pi)^{\frac{N}{2}}} \frac{b_0^{a_0}}{\Gamma(a_0)} \frac{|\mathbf{S}_N|^{1/2}}{|\mathbf{S}_0|^{1/2}} \int \frac{\Gamma(a_N)}{b_N^{a_N}} \mathrm{Gam}(\beta|a_N, b_N) \, \mathrm{d}\beta$$
$$= \frac{1}{(2\pi)^{N/2}} \frac{b_0^{a_0}}{b_N^{a_N}} \frac{\Gamma(a_N)}{\Gamma(a_0)} \frac{|\mathbf{S}_N|^{1/2}}{|\mathbf{S}_0|^{1/2}} \tag{3.118}$$

where a_N and b_N were derived in Exercise 3.12 and are given by (2.150), respectively (3.12.1). □

Exercise 3.24 ⋆⋆

Repeat the previous exercise but now use Bayes' theorem in the form

$$p(\mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{w}, \beta) p(\mathbf{w}, \beta)}{p(\mathbf{w}, \beta|\mathbf{t})} \tag{3.119}$$

and then substitute for the prior and posterior distributions and the likelihood function in order to derive the result (3.118).

Proof. We start by substituting the prior (3.112), the posterior (3.113) and the likelihood (3.10). The evidence function becomes

$$p(\mathbf{t}) = \frac{\mathcal{N}(\mathbf{t}|\mathbf{\Phi}\mathbf{w}, \beta^{-1}\mathbf{I}) \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \beta^{-1}\mathbf{S}_0) \mathrm{Gam}(\beta|a_0, b_0)}{\mathcal{N}(\mathbf{w}|\mathbf{m}_N, \beta^{-1}\mathbf{S}_N) \mathrm{Gam}(\beta|a_N, b_N)}$$

Let's expand the numerator:

$$\mathcal{N}(\mathbf{t}|\mathbf{\Phi}\mathbf{w}, \beta^{-1}\mathbf{I}) \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \beta^{-1}\mathbf{S}_0) \mathrm{Gam}(\beta|a_0, b_0) =$$
$$= \left(\frac{\beta}{2\pi}\right)^{M/2} \left(\frac{\beta}{2\pi}\right)^{N/2} \frac{b_0^{a_0} \beta^{a_0-1}}{\Gamma(a_0)|\mathbf{S}_0|^{1/2}} \exp\{\beta b_0\} \exp\left\{ -\frac{\beta}{2} ||\mathbf{t} - \mathbf{\Phi}\mathbf{w}||^2 \right\} \exp\left\{ -\frac{\beta}{2} (\mathbf{w} - \mathbf{m}_0)^T \mathbf{S}_0^{-1} (\mathbf{w} - \mathbf{m}_0) \right\}$$

We've seen in the previous exercise that

$$\exp\{\beta b_0\} \exp\left\{ -\frac{\beta}{2} ||\mathbf{t} - \mathbf{\Phi}\mathbf{w}||^2 \right\} \exp\left\{ -\frac{\beta}{2} (\mathbf{w} - \mathbf{m}_0)^T \mathbf{S}_0^{-1} (\mathbf{w} - \mathbf{m}_0) \right\}$$
$$= \exp\{\beta b_N\} \exp\left\{ -\frac{\beta}{2} (\mathbf{w} - \mathbf{m}_N)^T \mathbf{S}_N^{-1} (\mathbf{w} - \mathbf{m}_N) \right\}$$

so the numerator can be written as

$$\mathcal{N}(\mathbf{t}|\mathbf{\Phi}\mathbf{w}, \beta^{-1}\mathbf{I}) \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \beta^{-1}\mathbf{S}_0) \mathrm{Gam}(\beta|a_0, b_0) =$$
$$= \left(\frac{\beta}{2\pi}\right)^{M/2} \left(\frac{\beta}{2\pi}\right)^{N/2} \frac{b_0^{a_0} \beta^{a_0-1}}{\Gamma(a_0)|\mathbf{S}_0|^{1/2}} \exp\{\beta b_N\} \exp\left\{ -\frac{\beta}{2} (\mathbf{w} - \mathbf{m}_N)^T \mathbf{S}_N^{-1} (\mathbf{w} - \mathbf{m}_N) \right\}$$

Since we already have the exponential terms of the Gamma and Gaussian distributions, we can obtain the distributions by dividing by their normalization constants, so:

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

$$
\begin{aligned}
&= \left(\frac{\beta}{2\pi}\right)^{M/2} \left(\frac{\beta}{2\pi}\right)^{N/2} \frac{b_0^{a_0}\beta^{a_0-1}}{\Gamma(a_0)|\mathbf{S}_0|^{1/2}} \left[\frac{\Gamma(a_N)}{b_N^{a_N}\beta^{a_N-1}}\mathrm{Gam}(\beta|a_N,b_N)\right]\left[\left(\frac{2\pi}{\beta}\right)^{M/2}|\mathbf{S}_N|^{1/2}\mathcal{N}(\mathbf{w}|\mathbf{m}_N,\beta^{-1}\mathbf{S}_N)\right] \\
&= \frac{1}{(2\pi)^{N/2}}\frac{b_0^{a_0}}{b_N^{a_N}}\frac{\Gamma(a_N)}{\Gamma(a_0)}\frac{|\mathbf{S}_N|^{1/2}}{|\mathbf{S}_0|^{1/2}}\mathrm{Gam}(\beta|a_N,b_N)\mathcal{N}(\mathbf{w}|\mathbf{m}_N,\beta^{-1}\mathbf{S}_N)
\end{aligned}
$$

Finally, we substitute the numerator back into the evidence function and since the distribution forms factor out, we prove our hypothesis, that:

$$
p(\mathbf{t}) = \frac{1}{(2\pi)^{N/2}}\frac{b_0^{a_0}}{b_N^{a_N}}\frac{\Gamma(a_N)}{\Gamma(a_0)}\frac{|\mathbf{S}_N|^{1/2}}{|\mathbf{S}_0|^{1/2}} \tag{3.118}
$$

□

Chapter 4

Linear Models for Classification

Exercise 4.1 ⋆⋆

Given a set of data points $\{\mathbf{x}_n\}$, we can define the *convex hull* to be the data set of all points $\mathbf{x}$ given by

$$\mathbf{x} = \sum_n \alpha_n \mathbf{x}_n \tag{4.156}$$

where $\alpha_n \geq 0$ and $\sum_n \alpha_n = 1$. Consider a second set of points $\{\mathbf{y}_n\}$ together with their corresponding convex hull. By definition, the two set of points will be linearly separable if there exists a vector $\widehat{\mathbf{w}}$ and a scalar w_0 such that $\widehat{\mathbf{w}}^T\mathbf{x}_n + w_0 > 0$ for all $\mathbf{x}_n$, and $\widehat{\mathbf{w}}^T\mathbf{y}_n + w_0 < 0$ for all $\mathbf{y}_n$. Show that if their convex hulls intersect, the two sets of points cannot be linearly separable, and conversely that if they are linearly separable, their covex hulls do not intersect.

Proof. The vertices of the convex hulls are the data points $\{\mathbf{x}_n\}$ and $\{\mathbf{y}_n\}$. Therefore, the edges of the hulls will be represented by some segments beteween the data points. As a result, any point situated on the boundary of the hull can be written as a convex combination of the end-points of the segment it's contained by. Also, one can easily see that if two hulls intersect, they intersect in at least one point that is contained by the boundaries of both hulls.

1st Hypothesis: If the hulls intersect, the two sets of points are not linearly separable

Assume that the two hulls intersect in the point $\mathbf{z}$ situated on both hulls boundaries. From what we've seen above, the point $\mathbf{z}$ can be expressed as a convex combination between two data points of each set of data points. Therefore, there exist $\mathbf{x}_A, \mathbf{x}_B$ from $\{\mathbf{x}_n\}, \mathbf{y}_A, \mathbf{y}_B$ from $\{\mathbf{y}_n\}$ and $\lambda_\mathbf{x}, \lambda_\mathbf{y} \in [0, 1]$ such that we can express $\mathbf{z}$ as

$$\lambda_\mathbf{x}\mathbf{x}_A + (1 - \lambda_\mathbf{x})\mathbf{x}_B = \lambda_\mathbf{y}\mathbf{y}_A + (1 - \lambda_\mathbf{y})\mathbf{y}_B$$

Suppose that the sets $\{\mathbf{x}_n\}$ and $\{\mathbf{y}_n\}$ are linearly separable. Then there exists a discriminant function

$$\theta(\mathbf{a}) = \widehat{\mathbf{w}}^T\mathbf{a} + w_0$$

such that $\theta(\mathbf{x}_n) > 0$ for all $\mathbf{x}_n$ and $\theta(\mathbf{y}_n) < 0$ for all $\mathbf{y}_n$. From the linearity of the discriminant function, and rewritting $\theta(\mathbf{z})$ using the convex combinations forms, we have that

$$\lambda_\mathbf{x}\theta(\mathbf{x}_A) + (1 - \lambda_\mathbf{x})\theta(\mathbf{x}_B) = \lambda_\mathbf{y}\theta(\mathbf{y}_A) + (1 - \lambda_\mathbf{y})\theta(\mathbf{y}_B)$$

Since $\theta(\mathbf{x}_A), \theta(\mathbf{x}_B) > 0$ and $\theta(\mathbf{y}_A), \theta(\mathbf{y}_B) < 0$, this expression is obviously false, since the left-hand side of the equality is positive and the right-hand one is negative. Therefore, our supposition that the data sets are linearly separable is false and our main hypothesis is true.

2nd Hypothesis: If the two sets of points are linearly separable, then the hulls don't intersect.

This hypothesis is the counterpositive of the 1st hypothesis. Therefore, it's valid too. □

Exercise 4.2 ⋆ TODO

Consider the minimization of a sum-of-squares error function (**??**), and suppose that all of the target vectors in the training set satisfy a linear constraint

$$\mathbf{a}^T\mathbf{t}_n + b = 0 \tag{4.157}$$

where $\mathbf{t}_n$ corresponds to the n^{th} row of the matrix $\mathbf{T}$ in (**??**). Show that as a consequence of this constraint, the elements of the model prediction $\mathbf{y}(\mathbf{x})$ given by the least-squares solution (**??**) also satisfy this constraint, so that

$$\mathbf{a}^T\mathbf{y}(\mathbf{x}) + b = 0 \tag{4.158}$$

To do so, assume that one of the basis functions $\phi_0(\mathbf{x}) = 1$, so that the corresponding parameter w_0 plays the role of a bias.

Exercise 4.4 ⋆

Show that maximization of the class separation criterion given by (3.24) with respect to $\mathbf{w}$, using a Lagrange multiplier to enforce the constraint $\mathbf{w}^T\mathbf{w} = 1$, leads to the result that $\mathbf{w} \propto (\mathbf{m}_2 - \mathbf{m}_1)$.

Proof. Our goal is to maximize

$$m_2 - m_1 = \mathbf{w}^T(\mathbf{m}_2 - \mathbf{m}_1) \tag{3.24}$$

with the constraint that $\mathbf{w}^T\mathbf{w} = 1$. The corresponding Lagrangian is given by

$$\mathcal{L}(\mathbf{w}, \lambda) = \mathbf{w}^T(\mathbf{m}_2 - \mathbf{m}_1) + \lambda(\mathbf{w}^T\mathbf{w} - 1)$$

By taking the gradient of this with respect to $\mathbf{w}$ and λ, we have that

$$\nabla_{\mathbf{w},\lambda}\mathcal{L}(\mathbf{w}, \lambda) = \begin{pmatrix} \mathbf{m}_2 - \mathbf{m}_1 + 2\lambda\mathbf{w} \\ \mathbf{w}^T\mathbf{w} - 1 \end{pmatrix}$$

Setting to 0 the derivative with respect to $\mathbf{w}$ gives the initial result, that is

$$\mathbf{w} = \frac{\mathbf{m}_1 - \mathbf{m}_2}{2\lambda}$$

By replacing into the λ derivative and setting it to 0, we'd obtain that

$$\lambda = \frac{1}{4}||\mathbf{m}_1 - \mathbf{m}_2||^2$$

which gives

$$\mathbf{w} = \frac{2(\mathbf{m}_1 - \mathbf{m}_2)}{||\mathbf{m}_1 - \mathbf{m}_2||^2} \propto (\mathbf{m}_2 - \mathbf{m}_1)$$

□

Exercise 4.5 ⋆

By making use of (4.20), (4.23), and (4.24), show that the Fischer criterion (4.25) can be written in the form (4.26).

Proof. The Fisher criterion is defined to be the ratio of the between-class variance to the within-class variance and is given by

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} \tag{4.25}$$

where

$$m_k = \mathbf{w}^T \mathbf{m}_k \tag{4.23}$$

and

$$s_k^2 = \sum_{n \in \mathcal{C}_k} (y_n - m_k)^2 \tag{4.24}$$

By substituting (4.23) into the numerator of the Fischer expression,

$$\begin{aligned}(m_2 - m_1)^2 = (\mathbf{w}^T\mathbf{m}_2 - \mathbf{w}^T\mathbf{m}_1)^2 = \mathbf{w}^T(\mathbf{m}_2 - \mathbf{m}_1)\mathbf{w}^T(\mathbf{m}_2 - \mathbf{m}_1) &= \mathbf{w}^T(\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T\mathbf{w} \\ &= \mathbf{w}^T\mathbf{S}_B\mathbf{w}\end{aligned}$$

where $\mathbf{S}_B$ is the *between-class* covariance matrix and is given by

$$\mathbf{S}_B = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T \tag{4.27}$$

Similarily, we use the fact that the projection of the D-dimensional input vector $\mathbf{w}$ to one dimension is given by

$$y = \mathbf{w}^T\mathbf{x} \tag{4.20}$$

along with (4.23) and (4.24) to rewrite the denominator as

$$\begin{aligned}s_1^2 + s_2^2 &= \sum_{n \in \mathcal{C}_1} (y_n - m_1)^2 + \sum_{n \in \mathcal{C}_2} (y_n - m_2)^2 \\ &= \sum_{n \in \mathcal{C}_1} (\mathbf{w}^T\mathbf{x}_n - \mathbf{w}^T\mathbf{m}_1)^2 + \sum_{n \in \mathcal{C}_2} (\mathbf{w}^T\mathbf{x}_n - \mathbf{w}^T\mathbf{m}_2)^2 \\ &= \sum_{n \in \mathcal{C}_1} \mathbf{w}^T(\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^T\mathbf{w} + \sum_{n \in \mathcal{C}_2} \mathbf{w}^T(\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^T\mathbf{w} \\ &= \mathbf{w}^T\left[\sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^T + \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^T\right]\mathbf{w} \\ &= \mathbf{w}^T\mathbf{S}_W\mathbf{w}\end{aligned}$$

where $\mathbf{S}_W$ is the *within-class* covariance matrix and is given by

$$\mathbf{S}_W = \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^T + \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^T \tag{4.28}$$

Finally, by substituting the new expressions into (4.25), we can rewrite the Fischer criterion as

$$J(\mathbf{w}) = \frac{\mathbf{w}^T\mathbf{S}_B\mathbf{w}}{\mathbf{w}^T\mathbf{S}_W\mathbf{w}} \tag{4.26}$$

□

Exercise 4.7 ⋆

Show that the logistic sigmoid function (4.59) satisfies the property $\sigma(-a) = 1 - \sigma(a)$ and that its inverse is given by $\sigma^{-1}(y) = \ln\left(\frac{y}{1-y}\right)$.

Proof. The sigmoid function is given by

$$\sigma(a) = \frac{1}{1+e^{-a}} \tag{4.59}$$

The symmetry property is easily satisfied, as

$$\sigma(-a) = \frac{1}{1+e^{a}} = \frac{1+e^{a}+1+e^{-a}}{(1+e^{-a})(1+e^{a})} - \frac{1+e^{a}}{(1+e^{-a})(1+e^{a})} = 1 - \frac{1}{1+e^{-a}} = 1 - \sigma(a) \tag{4.60}$$

The sigmoid function is bijective, so inversable. Therefore, let $\sigma(x) = y$. Then,

$$y = \frac{1}{1+e^{-x}} \iff (1+e^{-x})y = 1 \iff e^{-x} = \frac{1-y}{y} \iff x = \ln\left(\frac{y}{1-y}\right)$$

so the inverse of the sigmoid function is given by

$$\sigma^{-1}(y) = \ln\left(\frac{y}{1-y}\right)$$

□

Exercise 4.8 ⋆

Using (4.57) and (4.58), derive the result (4.65) for the posterior class probability in the two-class generative model with Gaussian densities, and verify the results (4.66) and (4.67) for the parameters $\mathbf{w}$ and w_0.

Proof. It is known that the posterior probability for class $\mathcal{C}_1$ can be written as

$$p(\mathcal{C}_1|\mathbf{x}) = \sigma(a) \tag{4.57}$$

where we have defined

$$a = \ln \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \tag{4.58}$$

and σ is the logistic sigmoid function defined by (4.59). We start by expanding a and rewritting it as

$$a = \ln p(\mathbf{x}|\mathcal{C}_1) - \ln p(\mathbf{x}|\mathcal{C}_2) + \ln \frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}$$

Since the class-conditional densities are Gaussian, i.e. the density for a class $\mathcal{C}_k$ is given by

$$p(\mathbf{x}|\mathcal{C}_k) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) \right\} \tag{4.64}$$

one can easily obtain that

$$\begin{aligned} a &= \ln p(\mathbf{x}|\mathcal{C}_1) - \ln p(\mathbf{x}|\mathcal{C}_2) + \ln \frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)} \\ &= \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \mathbf{x} - \boldsymbol{\mu}_2^T \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 + \frac{1}{2} \boldsymbol{\mu}_2^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 + \ln \frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)} \\ &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 + \frac{1}{2} \boldsymbol{\mu}_2^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 + \ln \frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)} \\ &= \mathbf{w}^T \mathbf{x} + w_0 \end{aligned}$$

where we have defined

$$\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{4.66}$$

$$w_0 = -\frac{1}{2} \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 + \frac{1}{2} \boldsymbol{\mu}_2^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 + \ln \frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)} \tag{4.67}$$

Therefore, the posterior probability for class $\mathcal{C}_1$ is given by

$$p(\mathcal{C}_1|\mathbf{x}) = \sigma(\mathbf{w}^T \mathbf{x} + w_0) \tag{4.65}$$

□

Exercise 4.9 ⋆

Consider a generative classification model for K classes defined by prior class probabilities $p(\mathcal{C}_k) = \pi_k$ and general class-conditional densities $p(\boldsymbol{\phi}|\mathcal{C}_k)$ where ϕ is the input feature vector. Suppose we are given a training set $\{\boldsymbol{\phi}_n, \mathbf{t}_n\}$ where $n = 1, \ldots, N$, and $\mathbf{t}_n$ is a binary target vector of length K that use the 1-of-K coding scheme, so that it has components $t_{nj} = \mathbf{I}_{jk}$ if pattern n is from class $\mathcal{C}_k$. Assuming that the data points are drawn independently from this model, show that the maximum-likelihood solution for the prior probabilites is given by

$$\pi_k = \frac{N_k}{N} \tag{4.159}$$

where N_k is the number of data points assigned to class $\mathcal{C}_k$.

Proof. Let $\mathbf{T}$ be the $N \times K$ matrix with the rows $\mathbf{t}_n^T$ and $\boldsymbol{\Phi}$ the $N \times M$ matrix with the rows $\boldsymbol{\phi}_n^T$. Also, let's define the column vector $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K)^T$. We have that

$$p(\boldsymbol{\phi}_n, \mathcal{C}_k) = p(\mathcal{C}_k) p(\boldsymbol{\phi}_n|\mathcal{C}_k) = \pi_k p(\boldsymbol{\phi}_n|\mathcal{C}_k)$$

so the likelihood function is given by

$$p(\mathbf{T}|\boldsymbol{\Phi}, \boldsymbol{\pi}) = \prod_{n=1}^{N} p(\mathbf{t}_n|\boldsymbol{\Phi}, \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{j=1}^{K} \Big[\pi_j p(\boldsymbol{\phi}_n|\mathcal{C}_j) \Big]^{t_{nj}}$$

The log likelihood is then easily derived as

$$\ln p(\mathbf{T}|\boldsymbol{\Phi}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{j=1}^{K} \Big(t_{nj} \ln \pi_j + t_{nj} \ln p(\boldsymbol{\phi}_n|\mathcal{C}_j) \Big)$$

We aim to minimize this with respect to π_k, while still maintaining the constraint $\sum_{k=1}^{N} \pi_k = 1$ Therefore, by only keeping the terms depending on π_k, we obtain the Lagrangian

$$\mathcal{L}(\boldsymbol{\pi}, \lambda) = \sum_{n=1}^{N} \sum_{j=1}^{K} t_{nj} \ln \pi_j + \lambda \left(\sum_{j=1}^{N} \pi_j - 1 \right)$$

with the gradient

$$\nabla_{\pi_k, \lambda} \mathcal{L}(\boldsymbol{\pi}, \lambda) = \begin{pmatrix} \dfrac{1}{\pi_k} \displaystyle\sum_{n=1}^{N} t_{nk} + \lambda \\ \\ \displaystyle\sum_{k=1}^{N} \pi_k - 1 \end{pmatrix}$$

By setting this gradient to 0, from the first relation we have that

$$\pi_k \lambda = -\sum_{n=1}^{N} t_{nk} = -N_k$$

Summing this over k, one can see that

$$\lambda = -N$$

After substituting this into the derivative and then setting it to 0, we obtain the maximum-likelihood solution for π_k, that is

$$\pi_k = \frac{N_k}{N}$$

□

Exercise 4.10 ⋆⋆

Consider the classification model of Exercise 4.9 and now suppose that the class-conditional densities are given by Gaussian distributions with a shared covariance matrix, so that

$$p(\boldsymbol{\phi}|\mathcal{C}_k) = \mathcal{N}(\boldsymbol{\phi}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}) \tag{4.160}$$

Show that the maximum likelihood solution for the mean of the Gaussian distribution for class $\mathcal{C}_k$ is given by

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} t_{nk} \boldsymbol{\phi}_n \tag{4.161}$$

which represents the mean of those feature vectors assigned to class $\mathcal{C}_k$. Similarly, show that the maximum likelihood solution for the shared covariance matrix is given by

$$\boldsymbol{\Sigma} = \sum_{k=1}^{K} \frac{N_k}{N} \mathbf{S}_k \tag{4.162}$$

where

$$\mathbf{S}_k = \frac{1}{N_k} \sum_{n=1}^{N} t_{nk} (\boldsymbol{\phi}_n - \boldsymbol{\mu_k})(\boldsymbol{\phi}_k - \boldsymbol{\mu}_k)^T \tag{4.163}$$

Thus $\mathbf{\Sigma}$ is given by a weighted average of the covariances of the data associated with each class, in which the weighting coefficients are given by the prior probabilities of the classes.

Proof. Using the same notations as in the last exercise, we remember that the log likelihood is given by

$$\ln p(\mathbf{T}|\mathbf{\Phi}, \boldsymbol{\pi}) = \sum_{n=1}^{N}\sum_{j=1}^{K}\Big(t_{nj}\ln\pi_j + t_{nj}\ln p(\boldsymbol{\phi}_n|\mathcal{C}_j)\Big)$$

By keeping only the parts depending on $\boldsymbol{\mu}_k$,

$$\ln p(\mathbf{T}|\mathbf{\Phi}, \boldsymbol{\pi}) = -\frac{1}{2}\sum_{n=1}^{N}\sum_{j=1}^{K} t_{nj}(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)^T\mathbf{\Sigma}^{-1}(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k) + \text{const}$$

For a symmetric matrix $\mathbf{W}$, one could show that

$$\frac{\partial}{\partial s}(\mathbf{x}-\mathbf{s})^T\mathbf{W}(\mathbf{x}-\mathbf{s}) = -2\mathbf{W}(\mathbf{x}-\mathbf{s})$$

Therefore, the derivative with respect to μ_k of the log-likelihood is given by

$$\frac{\partial}{\partial\boldsymbol{\mu}_k}\ln p(\mathbf{T}|\mathbf{\Phi}, \boldsymbol{\pi}) = \sum_{n=1}^{N} t_{nk}\mathbf{\Sigma}^{-1}(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)$$

Since $\sum_{n=1}^{N} t_{nk} = N_k$, by setting the derivative to 0 and rearranging the terms, we have that the solution for maximum likelihood is

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{n=1}^{N} t_{nk}\boldsymbol{\phi}_n \tag{4.161}$$

Similarly, we do the same for the shared covariance matrix. By keeping only the terms depending on $\mathbf{\Sigma}$, the log likelihood is given by

$$\begin{aligned}\ln p(\mathbf{T}|\mathbf{\Phi}, \boldsymbol{\pi}) &= -\frac{N}{2}\ln|\mathbf{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)^T\mathbf{\Sigma}^{-1}(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k) + \text{const}\\ &= -\frac{N}{2}\ln|\mathbf{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\big(\boldsymbol{\phi}_n^T\mathbf{\Sigma}^{-1}\boldsymbol{\phi}_n - 2\boldsymbol{\phi}_n^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_k + \boldsymbol{\mu}_k^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_k\big)\end{aligned}$$

By using (C.28) and

$$\frac{\partial}{\partial\mathbf{X}}\mathbf{a}^T\mathbf{X}^{-1}\mathbf{b} = -\mathbf{X}^{-T}\mathbf{a}\mathbf{b}^T\mathbf{X}^{-T}$$

we take the derivative of the log likelihood with respect to $\mathbf{\Sigma}$ and obtain:

$$\begin{aligned}\frac{\partial}{\partial\mathbf{\Sigma}}\ln p(\mathbf{T}|\mathbf{\Phi}, \boldsymbol{\pi}) &= -\frac{N}{2}\mathbf{\Sigma}^{-1} + \frac{1}{2}\sum_{n=1}^{N}\sum_{j=k}^{K} t_{nk}\mathbf{\Sigma}^{-1}\big(\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T - 2\boldsymbol{\phi}_n\boldsymbol{\mu}_k^T + \boldsymbol{\mu}_k\boldsymbol{\mu}_k^T\big)\mathbf{\Sigma}^{-1}\\ &= -\frac{N}{2}\mathbf{\Sigma}^{-1} + \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\mathbf{\Sigma}^{-1}(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)^T\mathbf{\Sigma}^{-1}\end{aligned}$$

$$= -\frac{N}{2}\boldsymbol{\Sigma}^{-1} + \frac{1}{2}\boldsymbol{\Sigma}^{-1}\bigg(\sum_{k=1}^{K} N_k \mathbf{S}_k\bigg)\boldsymbol{\Sigma}^{-1}$$

where $\mathbf{S}_k$ is defined by (4.163). Therefore, by setting this derivative to 0 and rearranging the terms, we obtain the maximum-likelihood solution for the shared covariance matrix

$$\boldsymbol{\Sigma} = \sum_{k=1}^{K} \frac{N_k}{N}\mathbf{S}_k \tag{4.162}$$

□

Exercise 4.11 ⋆⋆

Consider a classification problem with K classses for which the feature vector ϕ has M components each of which can take L discrete states. Let the values of the components be represented by a 1-of-L binary coding scheme. Further suppose that, conditioned on the class $\mathcal{C}_k$, the M components of $\boldsymbol{\phi}$ are independent, so that the class-conditional density factorizes with respect to the feature vector components. Show that the quantities given by (4.63), which appear in the argument to the softmax function describing the posterior class probabilties, are linear functions of the components of $\boldsymbol{\phi}$. Note that this represents an example of the naive Bayes model which is discussed in Section 8.2.2.

Proof. We've seen in Section 4.2 that the posterior probabilities can be written as *normalized exponentials*:

$$p(\mathcal{C}_k|\boldsymbol{\phi}) = \frac{\exp(a_k)}{\sum_j \exp(a_j)} \tag{4.62}$$

where

$$a_k = \ln p(\boldsymbol{\phi}|\mathcal{C}_k)p(\mathcal{C}_k) \tag{4.63}$$

Considering the setup of our classification problem, our class-conditional distribution will be of the form

$$p(\boldsymbol{\phi}|\mathcal{C}_k) = \prod_{i=1}^{M}\prod_{j=1}^{L} \boldsymbol{\mu}_{kij}^{\boldsymbol{\phi}_{ij}}$$

where $\boldsymbol{\mu}_k$ is given by (4.161). Therefore, by replacing into (4.63), the arguments of the softmax function are given by

$$a_k = \ln p(\mathcal{C}_k) + \sum_{i=1}^{M}\sum_{j=1}^{L} \boldsymbol{\phi}_{ij} \ln \boldsymbol{\mu}_{kij}$$

and are obviously linear functions of the components of $\boldsymbol{\phi}$. □

Exercise 4.12 ⋆

Verify the relation (4.88) for the derivative of the logistic sigmoid function defined by (4.59).

Proof. By taking the derivative of (4.59), we have that:

$$\frac{\partial}{\partial a}\sigma(a) = \frac{\partial}{\partial a}\left(\frac{1}{1+e^{-a}}\right) = \frac{e^{-a}}{(1+e^{-a})^2} = \frac{1+e^{-a}}{(1+e^{-a})^2} - \frac{1}{(1+e^{-a})^2} = \frac{1}{1+e^{-a}} - \left(\frac{1}{1+e^{-a}}\right)^2$$

We recognize the expression of the logistic sigmoid function, so

$$\frac{\partial}{\partial a}\sigma(a) = \sigma(a) - \sigma(a)^2 = \sigma(a)\big(1-\sigma(a)\big) \tag{4.88}$$

□

Exercise 4.13 ⋆

By making use of the result (4.88) for the derivative of the logistic sigmoid, show that the derivative of the error function (4.90) for the logistic regression model is given by (4.91).

Proof. The error function for the logistic regression is given by

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\} \tag{4.90}$$

where $y_n = \sigma(a_n)$ and $a_n = \mathbf{w}^T\boldsymbol{\phi}_n$. Taking the derivative of the log likelihood function with respect to $\mathbf{w}$ gives

$$\begin{aligned}
\nabla_{\mathbf{w}} \ln p(\mathbf{t}|\mathbf{w}) &= \nabla_{\mathbf{w}} \sum_{n=1}^{N} \big\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\big\} \\
&= \sum_{n=1}^{N} \Big[t_n \nabla_{\mathbf{w}} \ln y_n + (1-t_n)\nabla_{\mathbf{w}} \ln(1-y_n)\Big] \\
&= \sum_{n=1}^{N} \left[\frac{t_n}{y_n}\nabla_{\mathbf{w}} y_n + \frac{1-t_n}{1-y_n}\nabla_{\mathbf{w}}(1-y_n)\right] \\
&= \sum_{n=1}^{N} \frac{t_n(1-y_n) - y_n(1-t_n)}{y_n(1-y_n)}\nabla_{\mathbf{w}} y_n \\
&= \sum_{n=1}^{N} \frac{t_n - y_n}{y_n(1-y_n)}\nabla_{\mathbf{w}} y_n
\end{aligned} \tag{4.13.1}$$

Using (4.88), we can compute the gradient term:

$$\nabla_{\mathbf{w}} y_n = \nabla_{\mathbf{w}}\sigma(\mathbf{w}^T\boldsymbol{\phi}_n) = \frac{\partial\sigma}{\partial a_n}\frac{\partial a_n}{\partial \mathbf{w}}\big(\mathbf{w}^T\boldsymbol{\phi}_n\big) = y_n(1-y_n)\boldsymbol{\phi}_n$$

so the gradient of the log likelihood is

$$\nabla_{\mathbf{w}} \ln p(\mathbf{t}|\mathbf{w}) = \sum_{n=1}^{N}(t_n - y_n)\boldsymbol{\phi}_n$$

and the gradient of the error function is then given by

$$\nabla_{\mathbf{w}} E(\mathbf{w}) = -\nabla_{\mathbf{w}} \ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N} (t_n - y_n)\boldsymbol{\phi}_n \tag{4.91}$$

□

Exercise 4.14 ⋆

Show that for a linearly separable data set, the maximum likelihood solution for the logistic regression model is obtained by finding a vector $\mathbf{w}$ whose decision boundary $\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}) = 0$ separates the classes and then taking the magnitude of $\mathbf{w}$ to infinity.

Proof. Suppose there exists $\mathbf{w}$ such that the hyperplane $\mathbf{w}^T\boldsymbol{\phi} = 0$ separates the data set. Because the data set is linearly separable, $\mathbf{w}^T\boldsymbol{\phi}_a < 0$ and $\mathbf{w}^T\boldsymbol{\phi}_b > 0$ for all $\boldsymbol{\phi}_a \in \mathcal{C}_1$ and $\boldsymbol{\phi}_b \in \mathcal{C}_2$. One can observe that the maximum likelihood is obtained when $p(\mathcal{C}_1|\boldsymbol{\phi}_a) = 1$ and $p(\mathcal{C}_2|\boldsymbol{\phi}_b) = 1$ for all $\boldsymbol{\phi}_a \in \mathcal{C}_1, \boldsymbol{\phi}_b \in \mathcal{C}_2$. Since our hyperplane is already chosen, there is a fixed angle θ_n between each $\boldsymbol{\phi}_n$ and $\mathbf{w}$ such that $\cos\theta_n \neq 0$. Therefore, by using the geometric definition of the dot product

$$\mathbf{w}^T\boldsymbol{\phi}_n = \|\mathbf{w}\|\|\boldsymbol{\phi}_n\|\cos\theta_n$$

we see that our maximization is achieved by taking the magnitude of $\|\mathbf{w}\|$ to infinity, as

$$\lim_{\|\mathbf{w}\|\to\infty} p(\mathcal{C}_1|\boldsymbol{\phi}_a) = \lim_{\|\mathbf{w}\|\to\infty} \sigma(\|\mathbf{w}\|\|\boldsymbol{\phi}_a\|\cos\theta_a) = \lim_{\|\mathbf{w}\|\to\infty} \frac{1}{1+\exp\{-\|\mathbf{w}\|\|\boldsymbol{\phi}_a\|\cos\theta_a\}} = 1$$

and

$$\lim_{\|\mathbf{w}\|\to\infty} p(\mathcal{C}_2|\boldsymbol{\phi}_b) = \lim_{\|\mathbf{w}\|\to\infty} \sigma(\|\mathbf{w}\|\|\boldsymbol{\phi}_b\|\cos\theta_b) = \lim_{\|\mathbf{w}\|\to\infty} \frac{1}{1+\exp\{-\|\mathbf{w}\|\|\boldsymbol{\phi}_b\|\cos\theta_b\}} = 0$$

where $\boldsymbol{\phi} \in \mathcal{C}_1, \boldsymbol{\phi}_b \in \mathcal{C}_2$ and we've used the fact that $\mathbf{w}^T\boldsymbol{\phi}_a < 0$ and $\mathbf{w}^T\boldsymbol{\phi}_b > 0$. □

Exercise 4.15 ⋆⋆

Show that the Hessian matrix $\mathbf{H}$ for the logistic regression model, given by (4.97), is positive definite. Here $\mathbf{R}$ is a diagonal matrix with elements $y_n(1-y_n)$, and y_n is the output of the logistic regression model for input vector $\mathbf{x}_n$. Hence show that the error function is a convex function of $\mathbf{w}$ and it has an unique minimum.

Proof. The Hessian of the error function is given by

$$\mathbf{H} = \nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} y_n(1-y_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T = \boldsymbol{\Phi}^T\mathbf{R}\boldsymbol{\Phi} \tag{4.97}$$

Let $\mathbf{u}$ be a M-dimensional column vector. By using the sum formulation for the hessian matrix, we have that

$$\mathbf{u}^T\mathbf{H}\mathbf{u} = \sum_{n=1}^{N} y_n(1-y_n)\mathbf{u}^T\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T\mathbf{u} = \sum_{n=1}^{N} y_n(1-y_n)\big(\boldsymbol{\phi}_n^T\mathbf{u}\big)^T\boldsymbol{\phi}_n^T\mathbf{u} = \sum_{n=1}^{N} y_n(1-y_n)\|\boldsymbol{\phi}_n^T\mathbf{u}\|^2$$

which is > 0 since y_n is the output of the logistic sigmoid function, so $0 < y_n < 1$. Because $\mathbf{u}$ was chosen arbitrarily, we have that $\mathbf{H}$ is positive definite. As a result, the error function is convex and has an unique minimum. □

Exercise 4.16 ⋆

Consider a binary classification problem in which each observation $\mathbf{x}_n$ is known to belong to one of two classes, corresponding to $t = 0$ and $t = 1$, and suppose that the procedure for collecting training data is imperfect, so that training points are sometimes mislabelled. For every data point $\mathbf{x}_n$, instead of having a value t for the class label, we have instead a value π_n representing the probability that $t_n = 1$. Given a probabilistic model $p(t = 1|\boldsymbol{\phi})$, write down the log likelihood function appropiate for such a data set.

Proof. Straight away, we can see that $p(t = 0|\boldsymbol{\phi}) = 1 - p(t = 1|\boldsymbol{\phi})$. An fair approach would be to express $p(t_n|\boldsymbol{\phi})$ as a weighted average of $p(t_n = 0|\boldsymbol{\phi})$ and $p(t_n = 1|\boldsymbol{\phi}$ dictated by π_n. Therefore, the likelihood would be given by

$$p(\mathbf{t}|\boldsymbol{\phi}) = \prod_{n=1}^{N} p(t_n|\boldsymbol{\phi}) = \prod_{n=1}^{N} p(t_n = 1|\boldsymbol{\phi})^{\pi_n} p(t_n = 0|\boldsymbol{\phi})^{1-\pi_n} = \prod_{n=1}^{N} p(t_n = 1|\boldsymbol{\phi})^{\pi_n}\big\{1 - p(t_n = 1|\boldsymbol{\phi})\big\}^{1-\pi_n}$$

which has the log likelihood

$$\ln p(\mathbf{t}|\boldsymbol{\phi}) = \sum_{n=1}^{N} \pi_n p(t_n = 1|\boldsymbol{\phi}) + (1-\pi_n)\big\{1 - p(t_n = 1|\boldsymbol{\phi})\big\}$$

□

Exercise 4.17 ⋆

Show that the derivatives of the softmax activation function (4.104) where the a_k are defined by (4.105), are given by (4.106).

Proof. The softmax activation function is given by

$$y_k = \frac{\exp(a_k)}{\sum_j \exp(a_j)} \tag{4.104}$$

where

$$a_k = \mathbf{w}_k^T\boldsymbol{\phi} \tag{4.105}$$

Taking the derivative of (4.104) with respect to a_j and applying the quotient rule gives

$$\frac{\partial y_k}{\partial a_j} = \frac{\partial}{\partial a_j}\left(\frac{\exp(a_k)}{\sum_i \exp(a_i)}\right) = \frac{I_{kj}\exp(a_k)\sum_i \exp(a_i) - \exp(a_k)\exp(a_j)}{\left(\sum_i \exp(a_i)\right)^2}$$
$$= \frac{\exp(a_k)}{\sum_j \exp(a_j)}\left(I_{kj} - \frac{\exp(a_j)}{\sum_j \exp(a_j)}\right)$$

which is equivalent to

$$\frac{\partial}{\partial a_j} y_k = y_k(I_{kj} - y_j) \tag{4.106}$$

□

Exercise 4.18 ⋆

Using the result (4.106) for the derivatives of the softmax activation function, show that the gradients of the cross-entropy error (4.108) are given by (4.109).

Proof. The cross-entropy error is given by

$$E(\mathbf{w}_1, \ldots, \mathbf{w}_K) = -\ln p(\mathbf{T}|\mathbf{w}_1, \ldots, \mathbf{w}_K) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_{nk} \tag{4.108}$$

Taking its derivative with respect to $\mathbf{w}_j$ yields

$$\frac{\partial}{\partial \mathbf{w}_j} E(\mathbf{w}_1, \ldots, \mathbf{w}_K) = \frac{\partial}{\partial \mathbf{w}_j}\left(-\sum_{n=1}^{K}\sum_{k=1}^{K} t_{nk} \ln y_{nk}\right) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \frac{\partial}{\partial \mathbf{w}_j} \ln y_{nk}$$

By using (4.106) and the chain rule, we have that

$$\frac{\partial}{\partial \mathbf{w}_j} \ln y_{nk} = \frac{1}{y_{nk}} \frac{\partial y_{nk}}{\partial \mathbf{w}_j} = \frac{1}{y_{nk}} \frac{\partial y_{nk}}{\partial a_j} \frac{\partial a_j}{\partial \mathbf{w}_j} = \frac{1}{y_{nk}} y_{nk}(I_{kj} - y_{nj})\boldsymbol{\phi}_n = (I_{kj} - y_{nj})\boldsymbol{\phi}_n$$

Replacing back into the gradient,

$$\frac{\partial}{\partial \mathbf{w}_j} E(\mathbf{w}_1, \ldots, \mathbf{w}_K) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}(I_{kj} - y_{nj})\boldsymbol{\phi}_n = -\sum_{n=1}^{N} t_{nj}\boldsymbol{\phi}_n + \sum_{n=1}^{N}\left(\sum_{k=1}^{N} t_{nk}\right) y_{nj}\boldsymbol{\phi}_n$$

gives the desired result

$$\frac{\partial}{\partial \mathbf{w}_j} E(\mathbf{w}_1, \ldots, \mathbf{w}_K) = \sum_{n=1}^{N}(y_{nj} - t_{nj})\boldsymbol{\phi}_n \tag{4.109}$$

□

Exercise 4.19 ⋆

Write down expressions for the gradient of the log likelihood, as well as the corresponding Hessian matrix, for the probit regression model defined in Section 4.3.5. These are quantities that would be required to train such a model using IRLS.

Proof. The probit function is given by

$$\Phi(a) = \int_{-\infty}^{a} \mathcal{N}(\theta|0,1)\,\mathrm{d}\theta \tag{4.114}$$

Therefore, from the fundamental theorem of calculus, we have that

$$\frac{\partial}{\partial a}\Phi(a) = \frac{\partial}{\partial a}\int_{-\infty}^{a} \mathcal{N}(\theta|0,1)\,\mathrm{d}\theta = \mathcal{N}(a|0,1) = \frac{1}{\sqrt{2\pi}}e^{-\frac{a^2}{2}}$$

so

$$\nabla_{\mathbf{w}} y_n = \nabla_{\mathbf{w}}\Phi(a_n) = \frac{\partial\Phi}{\partial a_n}\frac{\partial a_n}{\partial \mathbf{w}} = \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{a_n^2}{2}\right\}\boldsymbol{\phi}_n$$

The probit regression model has a very similar form with what we've used for the logistic regression model in Section 4.3.2. More specifically, the log likelihood is still given by (4.90), but this time with $y_n = \Phi(a_n)$. Therefore, the general form for gradient of the log likelihood derived in Exercise 4.13, (4.13.1) can be used here too, so:

$$\nabla_{\mathbf{w}} \ln p(\mathbf{t}|\mathbf{w}) = \sum_{n=1}^{N}\frac{t_n - y_n}{y_n(1-y_n)}\nabla_{\mathbf{w}} y_n = \frac{1}{\sqrt{2\pi}}\sum_{n=1}^{N}\frac{t_n - y_n}{y_n(1-y_n)}\exp\left\{-\frac{a_n^2}{2}\right\}\boldsymbol{\phi}_n$$

By taking the gradient of this again, we find the Hessian matrix using the chain rule:

$$\begin{aligned}\mathbf{H} = \nabla_{\mathbf{w}}\nabla_{\mathbf{w}} \ln p(\mathbf{t}|\mathbf{w}) &= \frac{1}{\sqrt{2\pi}}\sum_{n=1}^{N}\nabla_{\mathbf{w}}\left[\frac{t_n - y_n}{y_n(1-y_n)}\exp\left\{-\frac{a_n^2}{2}\right\}\right]\boldsymbol{\phi}_n \\ &= \frac{1}{\sqrt{2\pi}}\sum_{n=1}^{N}\left[\left(\nabla_{\mathbf{w}}\frac{t_n - y_n}{y_n(1-y_n)}\right)\exp\left\{-\frac{a_n^2}{2}\right\} + \frac{t_n - y_n}{y_n(1-y_n)}\left(\nabla_{\mathbf{w}}\exp\left\{-\frac{a_n^2}{2}\right\}\right)\right]\boldsymbol{\phi}_n\end{aligned}$$

We compute each gradient term separately, so

$$\begin{aligned}\nabla_{\mathbf{w}}\frac{t_n - y_n}{y_n(1-y_n)} &= -\frac{y_n(1-y_n) + (t_n - y_n)(1-2y_n)}{y_n^2(1-y_n)^2}\nabla_{\mathbf{w}} y_n = \frac{y_n^2 - 2t_n y_n + t_n}{y_n^2(1-y_n)^2}\nabla_{\mathbf{w}} y_n \\ &= \frac{y_n^2 - 2t_n y_n + t_n}{y_n^2(1-y_n)^2\sqrt{2\pi}}\exp\left\{-\frac{a_n^2}{2}\right\}\boldsymbol{\phi}_n\end{aligned}$$

and

$$\nabla_{\mathbf{w}}\exp\left\{-\frac{a_n^2}{2}\right\} = -a_n\exp\left\{-\frac{a_n^2}{2}\right\}\nabla_{\mathbf{w}} a_n = -a_n\exp\left\{-\frac{a_n^2}{2}\right\}\boldsymbol{\phi}_n$$

Hence, the hessian matrix becomes

$$\mathbf{H} = \frac{1}{\sqrt{2\pi}}\sum_{n=1}^{N}\left[\frac{y_n^2 - 2t_n y_n + t_n}{y_n^2(1-y_n)^2\sqrt{2\pi}}\exp\{-a_n^2\}\boldsymbol{\phi}_n - \frac{a_n(t_n - y_n)}{y_n(1-y_n)}\exp\left\{-\frac{a_n^2}{2}\right\}\boldsymbol{\phi}_n\right]\boldsymbol{\phi}_n$$

□

Exercise 4.21 ⋆

Show that the probit function (4.114) and the erf function (4.115) are related by (4.116).

Proof. The error function is given by

$$\mathrm{erf}(a) = \frac{2}{\sqrt{\pi}} \int_0^a \exp\left\{-\frac{\theta^2}{2}\right\} \mathrm{d}\theta \tag{4.115}$$

By using the fact that the Gaussian is symmetric around the mean, the probit function can be rewritten as

$$\begin{aligned}\Phi(a) = \int_{-\infty}^a \mathcal{N}(\theta|0,1)\,\mathrm{d}\theta = \int_{-\infty}^0 \mathcal{N}(\theta|0,1)\,\mathrm{d}\theta + \int_0^a \mathcal{N}(\theta|0,1)\,\mathrm{d}\theta &= \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_0^a \exp\left\{-\frac{\theta^2}{2}\right\} \mathrm{d}\theta \\ &= \frac{1}{2}\left\{1 + \frac{1}{\sqrt{2}}\,\mathrm{erf}(a)\right\} \end{aligned} \tag{4.116}$$

□

Exercise 4.22 ⋆

Using the result (4.135), derive the expression (4.137) for the log model evidence under the Laplace approximation.

Proof. The proof of this is almost identical to the one in Section 4.4.1. From Bayes' theorem the model evidence is given by

$$p(\mathcal{D}) = \int p(\mathcal{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \tag{4.136}$$

Identifying $f(\boldsymbol{\theta}) = p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ and $Z = p(\mathcal{D})$, and applying the result (4.135), we obtain the model evidence under the Laplace approximation:

$$p(\mathcal{D}) = p(\mathcal{D}|\boldsymbol{\theta}_{MAP}) p(\boldsymbol{\theta}_{MAP}) \frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}}$$

where $\boldsymbol{\theta}_{MAP}$ is the value of $\boldsymbol{\theta}$ at the mode of the posterior distribution, and $\mathbf{A}$ is the *Hessian* matrix of second derivatives of the negative log posterior

$$\mathbf{A} = -\nabla\nabla \ln p(\mathcal{D}|\boldsymbol{\theta}_{MAP}) p(\boldsymbol{\theta}_{MAP}) = -\nabla\nabla \ln p(\boldsymbol{\theta}_{MAP}|\mathcal{D}) \tag{4.138}$$

Therefore, the log model evidence is given by

$$\ln p(\mathcal{D}) = \ln p(\mathcal{D}|\boldsymbol{\theta}_{MAP}) + \ln p(\boldsymbol{\theta}_{MAP}) + \frac{M}{2} \ln(2\pi) - \frac{1}{2} \ln |\mathbf{A}| \tag{4.137}$$

□

Exercise 4.25 ⋆⋆

Suppose we wish to approximate the logistic sigmoid $\sigma(a)$ defined by (4.59) by a scaled probit function $\Phi(\lambda a)$ where $\Phi(a)$ is defined by (4.114). Show that if λ is chosen so that the derivatives of the two functions are equal at $a = 0$, then $\lambda^2 = \pi/8$.

Proof. We start by evaluating both function's derivatives at $a = 0$. We've seen in Exercise 4.19 that the derivative of the probit function is given by

$$\frac{\partial}{\partial a}\Phi(a) = \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{a^2}{2}\right\}$$

so

$$\frac{\partial}{\partial a}\Phi(\lambda a)\bigg|_{a=0} = \frac{\lambda}{\sqrt{2\pi}}\exp\left\{-\frac{a^2}{2}\right\}\bigg|_{a=0} = \frac{\lambda}{\sqrt{2\pi}}$$

From (4.88) we also obtain the derivative of the sigmoid function:

$$\frac{\partial}{\partial a}\sigma(a)\bigg|_{a=0} = \sigma(a)(1-\sigma(a))\bigg|_{a=0} = \frac{1}{4}$$

Finally, by using the fact that the derivatives of the functions are equal at $a = 0$, we quickly reach the result from the hypothesis, i.e. $\lambda^2 = \pi/8$. □

Chapter 5

Neural Networks

Exercise 5.1 ⋆⋆

Consider a two-layer network function of the form (5.7) in which the hidden-unit nonlinear activation functions $h(\cdot)$ are given by logistic sigmoid functions of the form

$$\sigma(a) = \frac{1}{1 + \exp(-a)}. \tag{5.191}$$

Show that there exists an equivalent network, which computes exactly the same function, but with hidden activation functions given by $\tanh(a)$ where the tanh function is defined by (5.59). Hint: first find the relation between $\sigma(a)$ and $\tanh(a)$, and then show that the parameters of the two networks differ by linear transformations.

Proof. The considered two-layer network has the form

$$y_k(\mathbf{x}, \mathbf{w}) = \sigma\left(\sum_{j=1}^{M} w_{kj}^{(2)} h\left(\sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)} \right) + w_{k0}^{(2)} \right) \tag{5.7}$$

Now, we've proved in Exercise 3.1 that

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

Therefore, we can rewrite y_k as

$$\begin{aligned} y_k(\mathbf{x}, \mathbf{w}) &= \sigma\left(\frac{1}{2} \sum_{j=1}^{M} w_{kj}^{(2)} \tanh\left(\frac{1}{2} \sum_{i=1}^{D} w_{ji}^{(1)} x_i + \frac{1}{2} w_{j0}^{(1)} \right) + \frac{1}{2} \sum_{j=1}^{M} w_{kj}^{(2)} + w_{k0}^{(2)} \right) \\ &= \sigma\left(\sum_{j=1}^{M} \omega_{kj}^{(2)} h\left(\sum_{i=1}^{D} \omega_{ji}^{(1)} x_i + \omega_{j0}^{(1)} \right) + \omega_{k0}^{(2)} \right) \end{aligned}$$

where

$$\omega_{ji}^{(1)} = \frac{1}{2} w_{ji}^{(1)} \qquad \omega_{j0}^{(1)} = \frac{1}{2} w_{j0}^{(1)} \qquad \omega_{kj}^{(2)} = \frac{1}{2} w_{kj}^{(2)} \qquad \omega_{k0}^{(2)} = \frac{1}{2} \sum_{j=1}^{M} w_{kj}^{(2)} + w_{k0}^{(2)}$$

Both new parameter sets can be obtained as linear transformations of the old ones, so there exists an equivalent two-layer network using tanh hidden activation functions, but different parameters. □

Exercise 5.2 $\star$

Show that maximizing the likelihood function under the conditional distribution (5.16) for a multioutput network is equivalent to minimizing the sum-of-squares error function (5.11).

Proof. The likelihood function is given by

$$p(\mathbf{T}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} p(\mathbf{t}_n|\mathbf{x}_n, \mathbf{w}, \beta)$$

The target variables are assumed to be distributed normally

$$p(\mathbf{t}_n|\mathbf{x}_n, \mathbf{w}, \beta) = \mathcal{N}(\mathbf{t}_n|\mathbf{y}(\mathbf{x}_n, \mathbf{w}), \beta^{-1}\mathbf{I}) \tag{5.16}$$

and since

$$\ln \mathcal{N}(\mathbf{t}_n|y(\mathbf{x}_n, \mathbf{w}), \beta^{-1}\mathbf{I}) = -\frac{N}{2}\ln\beta - \frac{NK}{2}\ln(2\pi) - \frac{\beta}{2}||\mathbf{y}(\mathbf{x}_n, \mathbf{w}) - \mathbf{t}_n||^2$$

the negative log-likelihood is given by

$$-\ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \frac{\beta}{2}\sum_{n=1}^{N}||\mathbf{y}(\mathbf{x}_n, \mathbf{w}) - \mathbf{t}_n||^2 + \text{const}$$

where we grouped the terms that don't depend on $\mathbf{w}$ under the constant term. Maximization of the likelihood function is equivalent to minimizing the negative log-likelihood. Therefore, one can easily find that this is equivalent to minimizing the error function

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}||\mathbf{y}(\mathbf{x}_n, \mathbf{w}) - \mathbf{t}_n||^2 \tag{5.11}$$

□

Exercise 5.3 $\star\star$

Consider a regression problem involving multiple target variables in which it is assumed that the distribution of the targets, conditioned on the input vector $\mathbf{x}$, is a Gaussian of the form

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}) = \mathcal{N}(\mathbf{t}|\mathbf{y}(\mathbf{x}, \mathbf{w}), \mathbf{\Sigma}) \tag{5.192}$$

where $\mathbf{y}(\mathbf{x}, \mathbf{w})$ is the output of a neural network with input vector $\mathbf{x}$ and weight vector $\mathbf{w}$, and $\mathbf{\Sigma}$ is the covariance of the assumed Gaussian noise on the targets. Given a set of independent observations of $\mathbf{x}$ and $\mathbf{t}$, write down the error function that must be minimized in order to find the maximum likelihood solution for $\mathbf{w}$, if we assume that $\mathbf{\Sigma}$ is fixed and known. Now assume that $\mathbf{\Sigma}$ is also to be determined from the data, and write down an expression for the maximum likelihood solution for $\mathbf{\Sigma}$. Note that the optimizations of $\mathbf{w}$ and $\mathbf{\Sigma}$ are now coupled, in contrast to the case of independent target variables discussed in Section 5.2.

Proof. The negative log-likelihood is given by

$$\begin{aligned}
-\ln p(\mathbf{T}|\mathbf{X},\mathbf{w}) &= -\sum_{i=1}^{N} \ln \mathcal{N}(\mathbf{t}_n|\mathbf{y}(\mathbf{x}_n,\mathbf{w}),\boldsymbol{\Sigma}) \\
&= \frac{NK}{2}\ln(2\pi) + \frac{N}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}\sum_{n=1}^{N}\big(\mathbf{y}(\mathbf{x}_n,\mathbf{w})-\mathbf{t}_n\big)^T\boldsymbol{\Sigma}^{-1}\big(\mathbf{y}(\mathbf{x}_n,\mathbf{w})-\mathbf{t}_n\big)
\end{aligned}$$

Maximizing the likelihood is equivalent to minimizing the negative log-likelihood. Therefore, the error function that must be minimized to obtain maximum likelihood is given by

$$E(\mathbf{w},\boldsymbol{\Sigma}) = \frac{N}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}\sum_{n=1}^{N}\big(\mathbf{y}(\mathbf{x}_n,\mathbf{w})-\mathbf{t}_n\big)^T\boldsymbol{\Sigma}^{-1}\big(\mathbf{y}(\mathbf{x}_n,\mathbf{w})-\mathbf{t}_n\big)$$

In the case when $\boldsymbol{\Sigma}$ is known, we can simply treat the determinant term as a constant, so minimizing the error function

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\big(\mathbf{y}(\mathbf{x}_n,\mathbf{w})-\mathbf{t}_n\big)^T\boldsymbol{\Sigma}^{-1}\big(\mathbf{y}(\mathbf{x}_n,\mathbf{w})-\mathbf{t}_n\big)$$

would yield the maximum likelihood solution $\mathbf{w}_{\text{ML}}$. If $\boldsymbol{\Sigma}$ is unknown, we can't do that and the determination of $\mathbf{w}_{\text{ML}}$ would use $\boldsymbol{\Sigma}$, so that's why this time the optimizations of $\mathbf{w}$ and $\boldsymbol{\Sigma}$ are coupled. The MLE for the covariance matrix is obtained by taking the derivative of the negative log-likelihood wrt. $\boldsymbol{\Sigma}^{-1}$, equalizing it to 0 and then solving for $\boldsymbol{\Sigma}$. Taking the derivative of the negative log-likelihood yields

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

where we've used the cyclic property of the trace operator and the fact that

$$\frac{\partial}{\partial\mathbf{A}}\ln|A| = A^{-T}$$

Now, equalizing the derivative with 0 and solving for $\boldsymbol{\Sigma}$ gives the MLE for the covariance matrix:

$$\boldsymbol{\Sigma}_{\text{ML}} = \frac{1}{N}\sum_{n=1}^{N}\big(\mathbf{y}(\mathbf{x}_n,\mathbf{w})-\mathbf{t}_n\big)^T\big(\mathbf{y}(\mathbf{x}_n,\mathbf{w})-\mathbf{t}_n\big)$$

□

Exercise 5.4 ⋆⋆

Consider a binary classification problem where the target values are $t \in \{0, 1\}$, with a network output $y(\mathbf{x}, \mathbf{w})$ that represents $p(t = 1|\mathbf{x})$, and suppose that there is a probability that the class label on a trainining data point has been incorrectly set. Assuming independent and identically distributed data, write down the error function corresponding to the negative log likelihood. Verify that the error function (5.21) is obtained when $\epsilon = 0$. Note that this error function makes the model robust to incorrectly labelled data, in contrast to the usual error function.

Proof. We're going to model the problem similarly with what we've done in Section 4.2.2, but this time taking into account the mislabelled training data probability. As a result, let $r \in \{0, 1\}$ the real target values, considering mislabelling. Therefore, we can find the label probabilities by weighting in the error chance:

$$p(r = 1|\mathbf{x}, \mathbf{w}) = (1 - \epsilon)p(t = 1|\mathbf{x}, \mathbf{w}) + \epsilon p(t = 0|\mathbf{x}, \mathbf{w}) = (1 - \epsilon)y(\mathbf{x}_n, \mathbf{w}) + \epsilon\big(1 - y(\mathbf{x}_n, \mathbf{w})\big)$$

$$p(r = 0|\mathbf{x}, \mathbf{w}) = (1 - \epsilon)p(t = 0|\mathbf{x}, \mathbf{w}) + \epsilon p(t = 1|\mathbf{x}, \mathbf{w}) = (1 - \epsilon)\big(1 - y(\mathbf{x}_n, \mathbf{w})\big) + \epsilon y(\mathbf{x}_n, \mathbf{w})$$

We can combine both of these into

$$\begin{aligned} p(r|\mathbf{x}, \mathbf{w}) &= p(r = 1|\mathbf{x}, \mathbf{w})^r p(r = 0|\mathbf{x}, \mathbf{w})^{1-r} \\ &= \big[(1 - \epsilon)y(\mathbf{x}_n, \mathbf{w}) + \epsilon\big(1 - y(\mathbf{x}_n, \mathbf{w})\big)\big]^r \big[(1 - \epsilon)\big(1 - y(\mathbf{x}_n, \mathbf{w})\big) + \epsilon y(\mathbf{x}_n, \mathbf{w})\big]^{1-r} \end{aligned}$$

Therefore, the negative log-likelihood is given by

$$-\ln p(\mathbf{r}|\mathbf{X}, \mathbf{w}) = -\ln \prod_{i=1}^{N} p(r_n|\mathbf{x}_n, \mathbf{w}) = -\sum_{i=1}^{N} \{r_n \ln p(r_n = 1|\mathbf{x}_n, \mathbf{w}) + (1 - r_n) \ln p(r_n = 0|\mathbf{x}, \mathbf{w})\}$$

As a result, this is equivalent to minimizing the error function

$$E(\mathbf{w}) = -\sum_{i=1}^{N} \big[r_n \ln \big\{(1-\epsilon)y(\mathbf{x}_n, \mathbf{w})+\epsilon\big(1-y(\mathbf{x}_n, \mathbf{w})\big)\big\}+(1-r_n) \ln \big\{(1-\epsilon)\big(1-y(\mathbf{x}_n, \mathbf{w})\big)+\epsilon y(\mathbf{x}_n, \mathbf{w})\big]$$

which for $\epsilon = 0$ is equivalent to (5.21). □

Exercise 5.5 ⋆

Show that maximizing likelihood for a multiclass neural network model in which the network outputs have the interpretation $y_k(\mathbf{x}, \mathbf{w}) = p(t_k = 1|\mathbf{x})$ is equivalent to minimization of the cross-entropy function (5.24).

Proof. Let's consider the binary target variables $t_k \in \{0, 1\}$ have a 1-of-K coding scheme indicating the class. If we assume the class labels are independent, given the input vector, the conditional distribution of the targets is

$$p(t_k|\mathbf{x}) = \prod_{k=1}^{K} p(t_k = 1|\mathbf{x})^{t_k}$$

As a result, the corresponding negative log likelihood is given by

$$-\ln p(\mathbf{T}|\mathbf{X},\mathbf{w}) = -\ln \prod_{n=1}^{N}\prod_{k=1}^{K} p(t_{nk}=1|\mathbf{x}_n)^{t_{nk}} = -\ln \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\ln p(t_{nk}=1|\mathbf{x}_n)$$

Therefore, maximizing the likelihood of the model is equivalent to minimization of the cross entropy function

$$E(\mathbf{w}) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\ln p(t_{nk}=1|\mathbf{x}_n) \tag{5.24}$$

□

Exercise 5.6 ⋆

Show the derivative of the error function (5.21) with respect to the activation a_k for output units having a softmax activation function satisfies (5.18).

Proof. The general result for the derivative of the softmax function with respect to the activation a_k was proved in Exercise 4.17 and is given by (4.106). Therefore, we have that

$$\frac{\partial y_k}{\partial a_k} = y_k(1-y_k)$$

Taking the derivative of

$$E(\mathbf{w}) = -\sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\} \tag{5.21}$$

with respect to a_k yields

$$\frac{\partial}{\partial a_k}E(\mathbf{w}) = -t_k\frac{\partial}{\partial a_k}\ln y_k - (1-t_k)\frac{\partial}{\partial a_k}\ln(1-y_k) = -t_k(1-y_k) + y_k(1-t_k)y_k = y_k - t_k$$

As a result,

$$\frac{\partial E}{\partial a_k} = y_k - t_k \tag{5.18}$$

□

Exercise 5.7 ⋆

Show the derivative of the error function (5.21) with respect to the activation a_k for an output unit having a logistic sigmoid activation function satisfies (5.18).

Proof. We've seen in Exercise 4.12 that

$$\frac{\partial}{\partial a}\sigma(a) = \sigma(a)(1-\sigma(a)) \tag{4.88}$$

Since the output unit has a logistic sigmoid activation function, then

$$y_k = \sigma(a_k)$$

Therefore, using (4.88) gives

$$\frac{\partial y_k}{\partial a_k} = \sigma(a_k)\big(1 - \sigma(a_k)\big) = y_k(1 - y_k)$$

Analogously to Exercise 5.6, one can quickly reach that (5.18) holds. □

Exercise 5.8 ⋆

We saw in (4.88) that the derivative of the logistic sigmoid activation function can be expressed in terms of the function value itself. Derive the corresponding result for the 'tanh' activation function defined by (5.59).

Proof. Taking the derivative of the 'tanh' function is straightforward:

$$\frac{\partial}{\partial a}\tanh(a) = \frac{\partial}{\partial a}\left(\frac{e^a - e^{-a}}{e^a + e^{-a}}\right) = \frac{\big(e^a + e^{-a}\big)^2 - \big(e^a - e^{-a}\big)^2}{\big(e^a + e^{-a}\big)^2} = 1 - \left(\frac{e^a - e^{-a}}{e^a + e^{-a}}\right)^2 = 1 - \tanh(a)^2$$

Notice that the derivative of the 'tanh' function can also be expressed as a function of itself. □

Exercise 5.9 ⋆

The error function (5.21) for binary classification problems was derived for a network having a logistic-sigmoid output activation function, so that $0 \leq y(\mathbf{x}, \mathbf{w}) \leq 1$, and data having target values $t \in \{0, 1\}$. Derive the corresponding error function if we consider a network having an output $-1 \leq y(\mathbf{x}, \mathbf{w}) \leq 1$ and target values $t = 1$ for class $\mathcal{C}_1$ and $t = -1$ for class $\mathcal{C}_2$. What would be the appropiate choice of output unit activation function?

Proof. The hyperbolic tangent is the appropiate choice for the ouput unit activation function, because 'tanh' is a sigmoid function and its values range between -1 and 1. Let's consider the case of binary classification in which we interpret the network output $y(\mathbf{x}, \mathbf{w})$ as the conditional probability $p(\mathcal{C}_1|\mathbf{x})$, with $p(\mathcal{C}_2|\mathbf{x})$ given by $1 - y(\mathbf{x}, \mathbf{w})$. The conditional distribution of targets given inputs is then of the form

$$p(t|\mathbf{x}, \mathbf{w}) = y(\mathbf{x}, \mathbf{w})^{\frac{1+t}{2}}\big\{1 - y(\mathbf{x}, \mathbf{w})\big\}^{\frac{1-t}{2}}$$

Taking the negative log-likelihood then yields

$$-\ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}) = -\ln \prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w}) = -\sum_{n=1}^{N}\left\{\frac{1+t}{2}\ln y(\mathbf{x}, \mathbf{w}) + \frac{1-t}{2}\ln\big(1 - y(\mathbf{x}, \mathbf{w})\big)\right\}$$

As a result, maximizing the likelihood is equivalent to minimizing the error function

$$E(\mathbf{w}) = -\sum_{n=1}^{N}\left\{\frac{1+t}{2}\ln y_n + \frac{1-t}{2}\ln(1 - y_n)\right\}$$

where y_n denotes $y(\mathbf{x}_n, \mathbf{w})$. □

Exercise 5.10 ⋆

Consider a Hessian matrix $\mathbf{H}$ with eigenvector equation (5.33). By setting the vector $\mathbf{v}$ in (5.39) equal to each of the eigenvectors $\mathbf{u}_i$ in turn, show that $\mathbf{H}$ is positive definite if, and only if, all of its eigenvalues are positive.

Proof. Consider the eigenvector equation

$$\mathbf{H}\mathbf{u}_i = \lambda_i \mathbf{u}_i \tag{5.33}$$

→ Assume that $\mathbf{H}$ is positive definite. Then,

$$\mathbf{u}_i^T \mathbf{H}\mathbf{u}_i = \lambda_i ||\mathbf{u}_i||^2 > 0$$

which happens only if the eigenvalues λ_i are positive.

← Suppose that the eigenvalues λ_i are positive. Since the eigenvectors form an orthonormal basis, an arbitrary vector $\mathbf{v}$ can be written in the form

$$\mathbf{v} = \sum_i c_i \mathbf{u}_i \tag{5.38}$$

Therefore,

$$\begin{aligned}\mathbf{v}^T\mathbf{H}\mathbf{v} &= \Big(\sum_i c_i\mathbf{u}_i\Big)^T \mathbf{H}\Big(\sum_i c_i\mathbf{u}_i\Big) = \Big(\sum_i c_i\mathbf{u}_i\Big)^T\Big(\sum_i c_i\lambda_i\mathbf{u}_i\Big) \\ &= \sum_i\sum_j \lambda_j c_i c_j \mathbf{u}_i^T\mathbf{u}_j = \sum_i \lambda_i c_i^2\end{aligned}$$

Since the eigenvalues λ_i are positive,

$$\mathbf{v}^T\mathbf{H}\mathbf{v} = \sum_i \lambda_i c_i^2 > 0$$

for all $\mathbf{v}$, which proves that $\mathbf{H}$ is positive definite.

□

Exercise 5.11 ⋆⋆

Consider a quadratic error function defined by (5.32), in which the Hessian matrix $\mathbf{H}$ has an eigenvalue equation given by (5.33). Show that the contours of constant error are ellipses whose axes are aligned with eigenvectors $\mathbf{u}_i$ with lengths that are inversly proportional to the square root of the corresponding eigenvalues λ_i.

Proof. Analogously to what we've seen in Section 5.3.2, we're going to rewrite

$$E(\mathbf{w}) \simeq E(\mathbf{w}^\star) + \frac{1}{2}(\mathbf{w} - \mathbf{w}^\star)^T \mathbf{H}(\mathbf{w} - \mathbf{w}^\star) \tag{5.32}$$

as

$$E(\mathbf{w}) \simeq E(\mathbf{w}^\star) + \frac{1}{2}\sum_i \lambda_i \alpha_i^2 \tag{5.36}$$

where we've expanded $(\mathbf{w} - \mathbf{w}^\star)$ as a linear combination of $\mathbf{H}$'s eigenvectors:

$$\mathbf{w} - \mathbf{w}^\star = \sum_i \alpha_i \mathbf{u}_i \tag{5.35}$$

Now, since $\mathbf{w}$ and $\mathbf{w}^\star$ are fixed, let $\xi = 2E(\mathbf{w}) - 2E(\mathbf{w}^\star)$. Therefore, one can rewrite (5.36) as

$$\xi \simeq \sum_i \lambda_i \alpha_i^2 = \sum_i \left(\frac{\alpha_i}{\lambda_i^{-1/2}}\right)^2$$

This equation describes an N-dimensional ellipsoid. Since the coordinates α_i that define it are using the orthonormal basis formed by $\{\mathbf{u}_i\}$, its axis are aligned with the eigenvectors $\mathbf{u}_i$. The axis length of an ellipse can be obtained by taking $\alpha_i = 0$, for $i \neq j$ such that

$$\xi \simeq \left(\frac{\alpha_j}{\lambda_j^{-1/2}}\right)^2$$

and respectively

$$\alpha_j \simeq \left(\frac{\xi}{\lambda_j}\right)^{1/2}$$

Therefore, the lenghts of the ellipses are inversly proportional to the square root of the corresponding eigenvalues λ_i. □

Exercise 5.12 ⋆⋆

By considering the local Taylor expansion (5.32) of an error function about a stationary point $\mathbf{w}^\star$, show that the necessary and sufficient condition for the stationary point to be a local minimum of the error function is that the Hessian matrix $\mathbf{H}$, defined by (5.30) with $\widehat{\mathbf{w}} = \mathbf{w}^\star$, be positive definite.

Proof.

→ Suppose that $\mathbf{H}$ is positive definite. From (5.32) one could then find that

$$E(\mathbf{w}) - E(\mathbf{w}^\star) > 0$$

for $\mathbf{w} \neq \mathbf{w}^\star$. Therefore, $E(\mathbf{w}^\star)$ would be the minimum value of E.

$\leftarrow$ Assume that $\mathbf{w}^\star$ is a local minimum of E. Then,

$$E(\mathbf{w}) - E(\mathbf{w}^\star) > 0$$

which would mean that

$$\frac{1}{2}(\mathbf{w} - \mathbf{w}^\star)^T \mathbf{H}(\mathbf{w} - \mathbf{w}^\star) > 0$$

for $\mathbf{w} \neq \mathbf{w}^\star$, i.e. $\mathbf{H}$ is positive definite, since $\mathbf{w}$ respectively $\mathbf{w} - \mathbf{w}^\star$ can be chosen arbitrarily.

□

Exercise 5.13 ⋆

Show that as a consequence of the symmetry of the Hessian matrix $\mathbf{H}$, the number of independent elements in the quadratic error function (5.28) is given by $W(W+3)/2$.

Proof. The independent elements in the

$$E(\mathbf{w}) \simeq E(\widehat{\mathbf{w}}) + (\mathbf{w} - \widehat{\mathbf{w}})^T \mathbf{b} + \frac{1}{2}(\mathbf{w} - \widehat{\mathbf{w}})^T \mathbf{H}(\mathbf{w} - \widehat{\mathbf{w}}) \tag{5.28}$$

are given by the terms containing $\mathbf{b}$ and $\mathbf{H}$. Since $\mathbf{b}$ has W elements and $\mathbf{H}$ is a symmetric matrix with $W(W+1)/2$ independent elements (see Exercise 2.21), one has a total of

$$W + \frac{W(W+1)}{2} = \frac{W(W+3)}{2}$$

independent elements, where W is the dimensionality of $\mathbf{w}$. □

Exercise 5.14 ⋆

By making a Taylor expansion, verify that the terms that are $O(\epsilon)$ cancel on the right-hand side of (5.69).

Proof. Taking the Taylor expansion around w_{ji} of the terms on the right hand side of

$$\frac{\partial E_n}{\partial w_{ji}} = \frac{E_n(w_{ji} + \epsilon) - E_n(w_{ji} - \epsilon)}{2\epsilon} + O(\epsilon^2) \tag{5.69}$$

yields

$$E_n(w_{ji} + \epsilon) \simeq E_n(w_{ji}) + \epsilon E_n'(w_{ji}) + \frac{\epsilon^2}{2} E_n''(w_{ji}) + O(\epsilon^3)$$

$$E_n(w_{ji} - \epsilon) \simeq E_n(w_{ji}) - \epsilon E_n'(w_{ji}) + \frac{\epsilon^2}{2} E_n''(w_{ji}) + O(\epsilon^3)$$

Substituting these results into (5.69) cancels the $O(\epsilon)$ terms and gives

$$\frac{\partial E_n}{\partial w_{ji}} \simeq E_n'(w_{ji}) + O(\epsilon^2)$$

□

Exercise 5.15 ⋆⋆

In Section 5.3.4, we derived a procedure for evaluation the Jacobian matrix of a neural network using a backpropagation procedure. Derive an alternative formalism for finding the Jacobian based on *forward propagation* equations.

Proof. The Jacobian can be obtained by using the *forward propagation* technique. This is similar to what we've seen in Section 5.3.4, but this time the computations will start from the output end of the network. We have that

$$J_{ki} = \frac{\partial y_k}{\partial x_i} = \frac{\partial y_k}{\partial a_k}\frac{\partial a_k}{\partial x_i}$$

Summing over the j hidden units that have links to k units yields

$$\frac{\partial a_k}{\partial x_i} = \sum_j \frac{\partial a_k}{\partial a_j}\frac{\partial a_j}{\partial x_i}$$

From (5.48), it's obvious that

$$\frac{\partial a_j}{\partial x_i} = w_{ji}$$

As a result,

$$J_{ki} = \frac{\partial y_k}{\partial a_k}\sum_j \frac{\partial a_k}{\partial a_j}w_{ji} = \frac{\partial y_k}{\partial a_k}\sum_j \frac{\partial a_k}{\partial z_j}\frac{\partial z_j}{\partial a_j}w_{ji} = \frac{\partial y_k}{\partial a_k}\sum_j \frac{\partial z_j}{\partial a_j}w_{kj}w_{ji}$$

Suppose that h is the activation function for the ouput layer, respectively g for the hidden layer. Then,

$$J_{ki} = h'(a_k)\sum_j g'(a_j)w_{kj}w_{ji}$$

Since the main steps are computing a_j and a_k (in this order), the process of evaluating the Jacobian can be tought of as a forward propagation process. □

Exercise 5.16 ⋆

The outer product approximation to the Hessian matrix for a neural network using a sum-of-squares error function is given by (5.84). Extend this result to the case of multiple outputs.

Proof. The sum-of-square error function for multiple outputs is given by

$$E = \frac{1}{2}\sum_{n=1}^{n} ||\mathbf{y}_n - \mathbf{t}_n||^2$$

Similarly to Section 5.4.2, our goal is to obtain the outer product approximation for the Hessian matrix of the error. Hence, computing the Hessian yields:

$$\mathbf{H} = \nabla\nabla E = \nabla\left(\frac{1}{2}\sum_{n=1}^{N}\nabla||\mathbf{y}_n - \mathbf{t}_n||^2\right) = \nabla\left(\sum_{n=1}^{N}(\mathbf{y}_n - \mathbf{t}_n)^T\nabla\mathbf{y}_n\right)$$

$$= \sum_{n=1}^{N} \nabla \mathbf{y}_n \nabla \mathbf{y}_n^T + \sum_{n=1}^{N} (\mathbf{y}_n - \mathbf{t}_n)^T \nabla \nabla \mathbf{y}_n$$

Neglecting the second term yields the outer product approximation for the Hessian matrix:

$$\mathbf{H} \simeq \sum_{n=1}^{N} \nabla \mathbf{y}_n \nabla \mathbf{y}_n^T$$

which is analogous to (5.84) for $\mathbf{b}_n = \nabla \mathbf{y_n}$ in the multiple output case. Note that, for simplicity all the ∇ symbols refer to $\nabla_{\mathbf{w}}$ □

Exercise 5.17 ⋆

Consider a squared loss function of the form

$$E = \frac{1}{2} \iint \{y(\mathbf{x}, \mathbf{w}) - t\}^2 p(\mathbf{x}, t) \, \mathrm{d}\mathbf{x} \, \mathrm{d}t \tag{5.193}$$

where $y(\mathbf{x}, \mathbf{w})$ is a parametric function such as a neural network. The result (1.89) shows that the function $y(\mathbf{x}, \mathbf{w})$ that minimizes this error is given by the conditional expectation of t given $\mathbf{x}$. Use this result to show that the second derivative of E with respect to two elements w_r and w_s of the vector $\mathbf{w}$, is given by

$$\frac{\partial^2 E}{\partial w_r \partial w_s} = \int \frac{\partial y}{\partial w_r} \frac{\partial y}{\partial w_s} p(\mathbf{x}) \, \mathrm{d}\mathbf{x} \tag{5.194}$$

Note that, for a finite sample form $p(\mathbf{x})$, we obtain (5.84).

Proof. To simplify the notation, we'll denote $y(\mathbf{x}, \mathbf{w})$ as y. Taking the second derivative of E yields

$$\begin{aligned}
\frac{\partial^2 E}{\partial w_s \partial w_r} &= \frac{\partial^2}{\partial w_s \partial w_r} \left(\frac{1}{2} \iint \{y - t\}^2 p(\mathbf{x}, t) \, \mathrm{d}\mathbf{x} \, \mathrm{d}t \right) \\
&= \frac{1}{2} \frac{\partial}{\partial w_s} \iint 2(y - t) \frac{\partial y}{\partial w_r} p(\mathbf{x}, t) \, \mathrm{d}\mathbf{x} \, \mathrm{d}t \\
&= \iint \frac{\partial y}{\partial w_s} \frac{\partial y}{\partial w_r} p(\mathbf{x}, t) \, \mathrm{d}\mathbf{x} \, \mathrm{d}t + \iint (y - t) \frac{\partial^2 y}{\partial w_s \partial w_r} p(\mathbf{x}, t) \, \mathrm{d}\mathbf{x} \, \mathrm{d}t
\end{aligned}$$

Using (1.89), i.e. that $y = \mathbb{E}_t[t|\mathbf{x}]$ minimizes the error, proves that the second integral term is null:

$$\begin{aligned}
\iint (y - t) \frac{\partial^2 y}{\partial w_s \partial w_r} p(\mathbf{x}, t) \, \mathrm{d}\mathbf{x} \, \mathrm{d}t &= \int y p(\mathbf{x}) \frac{\partial^2 y}{\partial w_s \partial w_r} \, \mathrm{d}\mathbf{x} - \int \left(\int t p(t|x) \, \mathrm{d}t \right) p(\mathbf{x}) \frac{\partial^2 y}{\partial w_s \partial w_r} \, \mathrm{d}\mathbf{x} \\
&= \int (y - \mathbb{E}_t[t|\mathbf{x}]) p(\mathbf{x}) \frac{\partial^2 y}{\partial w_s \partial w_r} \, \mathrm{d}\mathbf{x} \\
&= 0
\end{aligned}$$

As a result, the second derivative can be written as

$$\frac{\partial^2 E}{\partial w_s \partial w_r} = \iint \frac{\partial y}{\partial w_s} \frac{\partial y}{\partial w_r} p(\mathbf{x}) p(t|\mathbf{x}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}t = \int \frac{\partial y}{\partial w_s} \frac{\partial y}{\partial w_r} p(\mathbf{x}) \, \mathrm{d}\mathbf{x} \tag{5.194}$$

□

Exercise 5.18 ⋆

Consider a two-layer network of the form shown in Figure 5.1 with the addition of extra parameters corresponding to skip-layer connections that go directly from inputs to the outputs. By extending the discussion of Section 5.3.2, write down the equations for the derivatives of the error function with respect to these additional parameters.

Proof. Let the weight corresponding to the skip-layer connections be denoted by $w_{ki}^{(s)}$. The outputs will gain an extra sum corresponding to those connections:

$$y_k = \sum_j w_{kj}^{(2)} z_j + \sum_i w_{ki}^{(s)} x_i$$

Since δ_k's functional form remains unchanged, the derivatives with respect to the first-layer and second-layer weights remain the same as before, i.e. (5.67). The derivative with respect to the skip-layer is now given by

$$\frac{\partial E_n}{\partial w_{ki}^{(s)}} = \frac{\partial E_n}{\partial a_k} \frac{\partial a_k}{\partial w_{ki}^{(s)}} = \delta_k x_i$$

the same as the one of the second-layer. □

Exercise 5.19 ⋆

Derive the expression (5.85) for the outer product approximation to the Hessian matrix for a network having a single output with a logistic sigmoid output-unit activation function and a cross-entropy error function, corresponding to the result (5.84) for the sum-of-squares error function.

Proof. For simplicity, let y_n denote $y(\mathbf{x}_n, \mathbf{w})$. Consider a network with the cross-entropy error function

$$E(\mathbf{w}) = -\sum_{n=1}^{N} \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\} \tag{5.21}$$

and a single output with activation

$$y_n = \sigma(a_n)$$

From (4.88) one has that

$$\nabla y_n = \sigma(a_n)[1 - \sigma(a_n)]\nabla a_n = y_n(1 - y_n)\nabla a_n$$

Using the chain rule of differential calculus,

$$\mathbf{H} = \nabla\nabla E(\mathbf{w}) = \nabla \sum_{n=1}^{N} \frac{\partial E}{\partial a_n} \frac{\partial a_n}{\partial \mathbf{w}} = \nabla \sum_{n=1}^{N} \frac{\partial E}{\partial a_n} \nabla a_n = \nabla \sum_{n=1}^{N} \frac{\partial E}{\partial a_n} \nabla a_n$$

Computing the derivative term is straightforward:

$$\frac{\partial E}{\partial a_n} = -\left\{ \frac{t_n}{y_n} \frac{\partial y_n}{\partial a_n} - \frac{1 - t_n}{1 - y_n} \frac{\partial y_n}{\partial a_n} \right\} = \frac{y_n(1 - t_n) - t_n(1 - y_n)}{y_n(1 - y_n)} \frac{\partial y_n}{\partial a_n} = y_n - t_n$$

Substituting this into the initial expression gives

$$\mathbf{H} = \nabla \sum_{n=1}^{N} (y_n - t_n) \nabla a_n = \sum_{n=1}^{N} \left\{ \nabla y_n \nabla a_n^T + (y_n - t_n) \nabla \nabla a_n \right\}$$

The second term inside the sum contains the term $(y_n - t_n)$, so we can neglect it as seen in Section 5.4.2 and arrive at the *outer approximation* for the Hessian matrix by expanding ∇y_n:

$$\mathbf{H} \simeq \sum_{n=1}^{N} y_n (1 - y_n) \nabla a_n \nabla a_n^T$$

which is equivalent to (5.85) for $\mathbf{b}_n = \nabla a_n$. Note that, for simplicity all the ∇ symbols refer to $\nabla_{\mathbf{w}}$ □

Exercise 5.20 ⋆

Derive an expression for the outer product approximation to the Hessian matrix for a network having K outputs with a softmax output-unit activation function and a cross-entropy error function, corresponding to the result (5.84) for the sum-of-squares error function.

Proof. We'll take a similar approach to the previous exercises. For simplicity, let y_{nk} denote $y_k(\mathbf{x}_n, \mathbf{w}_k)$. Consider a network with the cross-entropy error function

$$E(\mathbf{w}) = -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \ln y_{nk}$$

and K outputs with the softmax activation function

$$y_{nk} = \frac{\exp\{a_{nk}\}}{\sum_{j=1}^{K} \exp\{a_{nj}\}}$$

Note that both $\mathbf{y}_n$ and $\mathbf{a}_n$ are vectors of size $1 \times K$. Using the chain rule of calculus yields

$$\mathbf{H} = \nabla \nabla E(\mathbf{w}) = \nabla \sum_{n=1}^{N} \frac{\partial E}{\partial \mathbf{a}_n} \frac{\partial \mathbf{a}_n}{\partial \mathbf{w}} = \nabla \sum_{n=1}^{N} \frac{\partial E}{\partial \mathbf{a_n}} \nabla \mathbf{a}_n$$

Now, the derivative term will be of size $1 \times K$. The value of the i-th element will be given by

$$\left(\frac{\partial E}{\partial \mathbf{a}_n} \right)_i = \frac{\partial E}{\partial a_{ni}} = -\sum_{k=1}^{K} t_{nk} \frac{\partial \ln y_{nk}}{\partial a_{ni}} = -\sum_{k=1}^{K} \frac{t_{nk}}{y_{nk}} \frac{\partial y_{nk}}{\partial a_{ni}}$$

From Exercise 4.17, respectively (4.106), we have that

$$\frac{\partial y_{nk}}{\partial a_{ni}} = y_{nk} (\delta_{ik} - y_{ni})$$

Therefore,

$$\left(\frac{\partial E}{\partial \mathbf{a}_n}\right)_i = -\sum_{k=1}^{K} \frac{t_{nk}}{y_{nk}} y_{nk}(\delta_{ik} - y_{ni}) = -\sum_{k=1}^{K} t_{nk}(\delta_{ik} - y_{ni}) = y_{ni}\sum_{k=1}^{K} t_{nk} - t_{ni} = y_{ni} - t_{ni}$$

Hence, it's obvious that

$$\frac{\partial E}{\partial \mathbf{a}_n} = \mathbf{y}_n - \mathbf{t}_n$$

Substituting this back into the Hessian,

$$\mathbf{H} = \nabla \sum_{n=1}^{N} (\mathbf{y}_n - \mathbf{t}_n)\nabla \mathbf{a}_n = \sum_{n=1}^{N} \{\nabla \mathbf{y}_n \nabla \mathbf{a}_n + (\mathbf{y}_n - \mathbf{t}_n)\nabla\nabla \mathbf{a}_n\}$$

As before, the second term inside the sum contains the term $(\mathbf{y}_n - \mathbf{t}_n)$, so we can neglect it similarly to what we do in the previous exercises or Section 5.4.2. By ignoring the term, one arrives at the *outer approximation* of the Hessian matrix:

$$\mathbf{H} \simeq \sum_{n=1}^{N} \nabla \mathbf{y}_n \nabla \mathbf{a}_n = \sum_{n=1}^{N} \frac{\partial \mathbf{y}_n}{\partial \mathbf{a}_n} \nabla \mathbf{a}_n \nabla \mathbf{a}_n$$

where the derivative term is a $K \times K$ matrix with the elements

$$\left(\frac{\partial \mathbf{y}_n}{\partial \mathbf{a}_n}\right)_{ij} = \frac{\partial y_{ni}}{\partial a_{nj}} = y_{ni}(\delta_{ij} - y_{ni})$$

Note that for notation simplicity, all ∇ symbols refer to $\nabla_{\mathbf{w}}$. □

Exercise 5.21 ⋆⋆⋆ TODO

Extend the expression (5.86) for the outer product approximation of the Hessian matrix to the case of $K > 1$ output units. Hence, derive a recursive expression analogous to (5.87) for incrementing the number N of patterns and a similar expression for incrementing the number K of outputs. Use these results, together with the identity (5.88), to find sequential update expressions analogous to (5.89) for finding the inverse of the Hessian by incrementally including both extra patterns and extra outputs.

Proof. □

Exercise 5.22 ⋆⋆

Derive the results (5.93), (5.94), and (5.95) for the elements of the Hessian matrix of a two-layer feed-forward network by application of the chain rule of calculus.

Proof. As seen in Section 5.4.5, we consider the three separate blocks of the Hessian:

1. Both weights are in the second layer:

$$\begin{aligned}
\frac{\partial^2 E_n}{\partial w_{kj}^{(2)} \partial w_{k'j'}^{(2)}} &= \frac{\partial}{\partial w_{kj}^{(2)}} \left(\frac{\partial E_n}{\partial w_{k'j'}^{(2)}} \right) \\
&= \frac{\partial}{\partial w_{kj}^{(2)}} \left(\frac{\partial E_n}{\partial a_{k'}} \frac{\partial a_{k'}}{\partial w_{k'j'}^{(2)}} \right) \\
&= \frac{\partial}{\partial w_{kj}^{(2)}} \left(\frac{\partial E_n}{\partial a_{k'}} z_{j'} \right) \\
&= z_{j'} \frac{\partial}{\partial a_{k'}} \left(\frac{\partial E_n}{\partial a_k} \frac{\partial a_k}{\partial w_{kj}^{(2)}} \right) \\
&= z_j z_{j'} \frac{\partial^2 E_n}{\partial a_k \partial a_{k'}} \\
&= z_j z_{j'} M_{kk'} \qquad (5.93)
\end{aligned}$$

2. Both weights are in the first layer:

$$\begin{aligned}
\frac{\partial^2 E_n}{\partial w_{ji}^{(1)} \partial w_{j'i'}^{(1)}} &= \frac{\partial}{\partial w_{ji}^{(1)}} \left(\frac{\partial E_n}{\partial w_{j'i'}^{(1)}} \right) \\
&= \frac{\partial}{\partial w_{ji}^{(1)}} \left(\frac{\partial E_n}{\partial a_{j'}} \frac{\partial a_{j'}}{\partial w_{j'i'}^{(1)}} \right) \\
&= \frac{\partial}{\partial w_{ji}^{(1)}} \big(x_{i'} \delta_{j'} \big)
\end{aligned}$$

Using the backpropagation formula (5.56), we have that

$$\begin{aligned}
\frac{\partial^2 E_n}{\partial w_{ji}^{(1)} \partial w_{j'i'}^{(1)}} &= \frac{\partial}{\partial w_{ji}^{(1)}} \left(x_{i'} h'(a_{j'}) \sum_k w_{kj'}^{(2)} \delta_k \right) \\
&= x_{i'} \frac{\partial}{\partial w_{ji}^{(1)}} \big(h'(a_{j'}) \big) \sum_k \big(w_{kj'}^{(2)} \delta_k \big) + x_{i'} h'(a_{j'}) \sum_k w_{kj'}^{(2)} \frac{\partial \delta_k}{\partial w_{ji}^{(1)}}
\end{aligned}$$

For $j \neq j'$ the derivative in the first term is null. As a result, the first term can be written as

$$\begin{aligned}
x_{i'} \frac{\partial}{\partial w_{ji}^{(1)}} \big(h'(a_{j'}) \big) \sum_k w_{kj'}^{(2)} \delta &= \mathbf{I}_{jj'} x_{i'} \frac{\partial}{\partial w_{j'i}^{(1)}} \big(h'(a_{j'}) \big) \sum_k w_{kj'}^{(2)} \delta_k \\
&= \mathbf{I}_{jj'} x_i x_{i'} h''(a_{j'}) \sum_k w_{kj'}^{(2)} \delta_k
\end{aligned}$$

Now, let's compute the derivative in the second term:

$$\begin{aligned}
\frac{\partial \delta_k}{\partial w_{ji}^{(1)}} = \sum_{k'} \frac{\partial \delta_k}{\partial a_{k'}} \frac{\partial a_{k'}}{\partial w_{ji}^{(1)}} = \sum_{k'} \frac{\partial^2 E_n}{\partial a_k \partial a'_k} \frac{\partial a'_k}{\partial w_{ji}^{(1)}} &= \sum_{k'} M_{kk'} \frac{\partial}{\partial w_{ji}^{(1)}} \left(\sum_j w_{k'j}^{(2)} h(x_i w_{ji}^{(1)}) \right) \\
&= x_i h'(a_j) \sum_{k'} M_{kk'} w_{k'j}^{(2)}
\end{aligned}$$

Putting everything together yields the desired result:

$$\frac{\partial^2 E_n}{\partial w_{ji}^{(1)} \partial w_{j'i'}^{(1)}} = \mathbf{I}_{jj'} x_i x_{i'} h''(a_{j'}) \sum_k w_{kj'}^{(2)} \delta_k + x_i x_{i'} h'(a_j) h'(a_{j'}) \sum_k \sum_{k'} w_{kj'}^{(2)} w_{k'j}^{(2)} M_{kk'} \tag{5.94}$$

Note that this result is equivalent with the one in the book even if the k and k' are interchanged in the second term. This is because the sum ranges are chosen arbitrarily.

3. One weight in each layer:

$$\begin{aligned}
\frac{\partial^2 E_n}{\partial w_{ji}^{(1)} \partial w_{kj'}^{(2)}} &= \frac{\partial}{\partial w_{ji}^{(1)}} \left(\frac{\partial E_n}{\partial w_{kj'}^{(2)}} \right) \\
&= \frac{\partial}{\partial w_{ji}^{(1)}} \left(\frac{\partial E_n}{\partial a_k} \frac{\partial a_k}{\partial w_{kj'}^{(2)}} \right) \\
&= \frac{\partial}{\partial w_{ji}^{(1)}} \left(\delta_k z_{j'} \right) \\
&= z_{j'} \frac{\partial \delta_k}{\partial w_{ji}^{(1)}} + \delta_k \frac{\partial z_{j'}}{\partial w_{ji}^{(1)}}
\end{aligned}$$

We found the value of the first term in the previous case. Also, the derivative in the second term is null for $j \neq j'$. Therefore, the above expression becomes

$$\begin{aligned}
\frac{\partial^2 E_n}{\partial w_{ji}^{(1)} \partial w_{kj'}^{(2)}} &= z_{j'} x_i h'(a_j) \sum_{k'} M_{kk'} w_{k'j}^{(2)} + \mathbf{I}_{jj'} \delta_k x_i h'(a_j) \\
&= x_i h'(a_j) \left\{ \delta_k \mathbf{I}_{jj'} + z_{j'} \sum_{k'} w_{k'j}^{(2)} M_{kk'} \right\}
\end{aligned} \tag{5.95}$$

□

Exercise 5.23 ⋆⋆

Extend the results of Section 5.4.5 for the exact Hessian of two-layer network to include skip-layer connections that go directly from input to outputs.

Proof. Let's denote the skip layer by the (s) superscript. One has 3 cases for the weight combinations that include skip weights:

1. The non-skip activation is in the second layer:

$$\begin{aligned}
\frac{\partial^2 E_n}{\partial w_{ki}^{(s)} \partial w_{k'j}^{(2)}} &= \frac{\partial}{\partial w_{ki}^{(s)}} \left(\frac{\partial E_n}{\partial w_{k'j}^{(2)}} \right) \\
&= \frac{\partial}{\partial w_{ki}^{(s)}} \left(\frac{\partial E_n}{\partial a_{k'}} \frac{\partial a_{k'}}{\partial w_{k'j}^{(2)}} \right) \\
&= \frac{\partial}{\partial w_{ki}^{(s)}} \left(\delta_{k'} z_j \right)
\end{aligned}$$

$$= z_j \frac{\partial \delta_{k'}}{\partial w_{ki}^{(s)}}$$

Computing the derivative term separately yields

$$\frac{\partial \delta_{k'}}{\partial w_{ki}^{(s)}} = \frac{\partial \delta_{k'}}{\partial a_k} \frac{\partial a_k}{\partial w_{ki}^{(s)}} = \frac{\partial^2 E_n}{\partial a_k \partial a_{k'}} x_i = M_{kk'} x_i$$

Hence,

$$\frac{\partial^2 E_n}{\partial w_{ki}^{(s)} \partial w_{k'j}^{(2)}} = x_i z_j M_{kk'}$$

2. The non-skip activation is in the first layer:

$$\begin{aligned}
\frac{\partial^2 E_n}{\partial w_{ki}^{(s)} \partial w_{ji'}^{(1)}} &= \frac{\partial}{\partial w_{ki}^{(s)}} \left(\frac{\partial E_n}{\partial w_{ji'}^{(1)}} \right) \\
&= \frac{\partial}{\partial w_{ki}^{(s)}} \left(\frac{\partial E_n}{\partial a_j} \frac{\partial a_j}{\partial w_{ji'}^{(1)}} \right) \\
&= \frac{\partial}{\partial w_{ki}^{(s)}} \left(x_{i'} \delta_j \right) \\
&= x_{i'} \frac{\partial \delta_j}{\partial w_{ki}^{(s)}}
\end{aligned}$$

Using the back-propagation formula (5.56) gives

$$\begin{aligned}
\frac{\partial^2 E_n}{\partial w_{ki}^{(s)} \partial w_{ji'}^{(1)}} &= x_{i'} \frac{\partial \delta_j}{\partial w_{ki}^{(s)}} \left(h'(a_j) \sum_{k'} w_{k'j}^{(2)} \delta_{k'} \right) \\
&= x_{i'} h'(a_j) \sum_{k'} w_{k'j}^{(2)} \frac{\partial \delta_{k'}}{\partial w_{ki}^{(s)}}
\end{aligned}$$

We've already computed the derivative term in the last case. Therefore,

$$\frac{\partial^2 E_n}{\partial w_{ki}^{(s)} \partial w_{ji'}^{(1)}} = x_i x_{i'} h'(a_j) \sum_{k'} w_{k'j}^{(2)} M_{kk'}$$

3. Both weights are skip weights:

$$\begin{aligned}
\frac{\partial^2 E_n}{\partial w_{ki}^{(s)} \partial w_{k'i'}^{(s)}} &= \frac{\partial}{\partial w_{ki}^{(s)}} \left(\frac{\partial E_n}{\partial w_{k'i'}^{(s)}} \right) \\
&= \frac{\partial}{\partial w_{ki}^{(s)}} \left(\frac{\partial E_n}{\partial a_{k'}} \frac{\partial a_{k'}}{\partial w_{k'i'}^{(s)}} \right) \\
&= \frac{\partial}{\partial w_{ki}^{(s)}} \left(\delta_{k'} x_{i'} \right) \\
&= x_{i'} \frac{\partial \delta_{k'}}{\partial w_{ki}^{(s)}}
\end{aligned}$$

We've already computed the derivative term in the first case. As a result,

$$\frac{\partial^2 E_n}{\partial w_{ki}^{(s)} \partial w_{k'i'}^{(s)}} = x_i x_{i'} M_{kk'}$$

□

Exercise 5.24 ⋆⋆

Verify that the network function defined by (5.113) and (5.114) is invariant under the transformation (5.115) applied to the inputs, provided the weights and biases are simultaneously transformed using (5.116) and (5.117). Similarly, show that the network outputs can be transformed according to (5.118) by applying the transformation (5.119) and (5.120) to the second-layer weights and biases.

Proof. Let's make the transformations (5.115), (5.116) and (5.117) and check the new value of $\widetilde{a_j}$.

$$\widetilde{a_j} = \sum_i \widetilde{w}_{ji}\widetilde{x}_i + \widetilde{w}_{j0} = \sum_i \frac{1}{a} w_{ji}\big(ax_i + b\big) + w_{j0} - \frac{b}{a}\sum_i w_{ji} = \sum_i w_{ji}x_i + w_{j0} = a_j$$

Since the activations of the hidden layer remain the same after the transformation, we can conclude that the outputs are invariant under the above transformations. Now, let's apply the transformations (5.119) and (5.120) to the second-layer weights and biases. The new outputs look like

$$\widetilde{y}_k = \sum_j \widetilde{w}_{kj} z_j + \widetilde{w}_{k0} = \sum_j c w_{kj} z_j + c w_{k0} + d = c\bigg(\sum_j w_{kj} z_j + w_{k0}\bigg) + d = c y_k + d$$

which proves that the network outputs can be transformed as in (5.118). □

Exercise 5.25 ⋆⋆⋆

Consider a quadratic error function of the form

$$E = E_0 + \frac{1}{2}(\mathbf{w} - \mathbf{w}^\star)^T \mathbf{H} (\mathbf{w} - \mathbf{w}^\star) \tag{5.195}$$

where $\mathbf{w}^*$ represents the minimum, and the Hessian matrix $\mathbf{H}$ is positive definite and constant. Suppose the initial weight vector $\mathbf{w}^{(0)}$ is chosen to be at the origin and is updated using simple gradient descent

$$\mathbf{w}^{(\tau)} = \mathbf{w}^{(\tau-1)} - \rho \nabla E \tag{5.196}$$

where τ denotes the step number, and ρ is the learning rate (which is assumed to be small). Show that, after τ steps, the components of the weight vector parallel to the eigenvectors of $\mathbf{H}$ can be written

$$w_j^{(\tau)} = \{1 - (1 - \rho\eta_j)^\tau\} w_j^\star \tag{5.197}$$

where $w_j = \mathbf{w}^T \mathbf{u}_j$, $\mathbf{u}_j$ and η_j are the eigenvectors and eigenvalues, respectively, of $\mathbf{H}$ so that

$$\mathbf{H}\mathbf{u}_j = \eta_j \mathbf{u}_j \tag{5.198}$$

Show that as $\tau \to \infty$, this gives $\mathbf{w}^{(\tau)} \to \mathbf{w}^\star$ as expected, provided $|1 - \rho\eta_j| < 1$. Now suppose that training is halted after a finite number τ of steps. Show that the components of the weight vector parallel to the eigenvectors of the Hessian satisfy

$$w_j^{(\tau)} \simeq w_j^\star \text{ when } \eta_j \gg (\rho\tau)^{-1} \tag{5.199}$$

$$|w_j^{(\tau)}| \ll |w_j^\star| \text{ when } \eta_j \ll (\rho\tau)^{-1} \tag{5.200}$$

Compare this result with the discussion in Section 3.5.3 of regularization with simple weight decay, and hence show that $(\rho\tau)^{-1}$ is analogous to the regularization parameter λ. The above results also show that the effective number of parameters in the network, as defined by (3.91), grows as the training progresses.

Proof. Taking the gradient of the error with respect to $\mathbf{w}$ yields

$$\nabla E = \nabla\left(E_0 + \frac{1}{2}(\mathbf{w} - \mathbf{w}^\star)^T\mathbf{H}(\mathbf{w} - \mathbf{w}^\star)\right) = \mathbf{H}(\mathbf{w} - \mathbf{w}^\star)$$

Substituting this into (5.196) gives

$$\mathbf{w}^{(\tau)} = \mathbf{w}^{(\tau-1)} - \rho\mathbf{H}(\mathbf{w}^{(\tau-1)} - \mathbf{w}^\star)$$

Now, we left-multiply by $\mathbf{u}_j^T$ both sides of the expression and use the fact that $w_j = \mathbf{w}^T\mathbf{u}_j$, along with (5.198) to obtain that

$$w_j^{(\tau)} = w_j^{(\tau-1)} - \rho\eta_j w_j^{(\tau-1)} + \rho\eta_j w_j^\star = (1 - \rho\eta_j)w_j^{(\tau-1)} + \rho\eta_j w_j^\star$$

Let's prove (5.197) by induction. The base case for $\tau = 1$ is obviously holding since $\mathbf{w}^{(0)}$ is the origin:

$$w_j^{(1)} = (1 - \rho\eta_j)w_j^{(0)} + \rho\eta_j w_j^\star = \rho\eta_j w_j^\star = \{1 - (1 - \rho\eta_j)\}w_j^\star$$

For the general case, let $\tau = t \in \mathbb{N}$ and assume that (5.197) holds:

$$w_j^{(t)} = \{1 - (1 - \rho\eta_j)^t\}w_j^\star$$

Substituting the value of $w_j^{(t)}$ into

$$w_j^{(t+1)} = (1 - \rho\eta_j)w_j^{(t)} + \rho\eta_j w_j^\star$$

gives

$$\begin{aligned} w_j^{(t+1)} &= (1 - \rho\eta_j)\{1 - (1 - \rho\eta_j)^t\}w_j^\star + \rho\eta_j w_j^\star \\ &= \left\{(1 - \rho\eta_j) - (1 - \rho\eta_j)^{t+1} + \rho\eta_j\right\}w_j^\star \\ &= \{1 - (1 - \rho\eta_j)^{(t+1)}\}w_j^\star \end{aligned}$$

Since the base case and general recursive implication holds, we proved by induction that (5.197) holds. Now, taking the number of steps τ to infinity yields

$$\lim_{\tau\to\infty} w_j^{(\tau)} = \lim_{\tau\to\infty}\{1 - (1 - \rho\eta_j)^\tau\}w_j^\star = w_j^\star$$

for $|1-\rho\eta_j| < 1$ since as $\tau \to \infty$, one has that $(1-\rho\eta_j)^\tau \to 0$. Since $\eta_j\rho\tau \gg 1$ and $|1-\rho\eta_j| < 1, \tau$ must be large. Therefore, as proved above, $w_j^{(\tau)} \simeq w_j^\star$. Now, since $\eta_j\rho\tau \ll 1$ and τ is finite, $\rho\tau$ must be very small. We use this fact by expanding the polynomial and ignoring the higher order terms:

$$\begin{aligned}
|w_j^{(\tau)}| &= |1-(1-\rho\eta_j)^\tau||w_j^\star| \\
&= |1-\big(1-\tau\rho\eta_j + O(\rho^2\eta_j^2)\big)||w_j^\star| \\
&\simeq |\tau\rho\eta_j||w_j^\star| \\
&\ll |w_j^\star|
\end{aligned}$$

□

Exercise 5.26 ⋆⋆

Consider a multilayer perceptron with arbitrary feed-forward topology, which is to be trained by minimzing the *tangent propagation* error function in which the regularizing function is given by (5.128). Show that the regularization term Ω can be written as a sum over patterns of terms of the form

$$\Omega_n = \frac{1}{2}\sum_k (\mathcal{G}y_{nk})^2 \tag{5.201}$$

where $\mathcal{G}$ is a differential operator defined by

$$\mathcal{G} \equiv \sum_i \tau_i \frac{\partial}{\partial x_i} \tag{5.202}$$

By acting on the forward propagation equations

$$z_j = h(a_j), \qquad a_j = \sum_i w_{ji} z_i \tag{5.203}$$

with the operator $\mathcal{G}$, show that Ω_n can be evaluated by forward propagation using the following equations:

$$\alpha_j = h'(a_j)\beta_j, \qquad \beta_j = \sum_i w_{ji}\alpha_i \tag{5.204}$$

where we have defined the new variables

$$\alpha_j \equiv \mathcal{G}z_j, \qquad \beta_j \equiv \mathcal{G}a_j \tag{5.205}$$

Now show that the derivatives of Ω_n with respect to a weight w_{rs} in the network can be written in the form

$$\frac{\partial\Omega_n}{\partial w_{rs}} = \sum_k \alpha_k\{\phi_{kr}z_s + \delta_{kr}\alpha_s\} \tag{5.206}$$

where we have defined

$$\delta_{kr} \equiv \frac{\partial y_k}{\partial a_r}, \qquad \phi_{kr} = \mathcal{G}\delta_{kr}. \tag{5.207}$$

Write down the backpropagation equations for δ_{kr}, and hence derive the set of backpropagation equations for the evaluation of ϕ_{kr}.

Proof. Simply evaluating (5.201) using (5.202) gives

$$\Omega_n = \frac{1}{2}\sum_k (\mathcal{G}y_{nk})^2 = \frac{1}{2}\sum_k \left(\sum_i \tau_{ni}\frac{\partial y_{nk}}{\partial x_{ni}}\right)^2 = \frac{1}{2}\sum_k \left(\sum_i \tau_{ni} J_{nki}\right)^2$$

where J_{nki} is the (k,i)-th element of the Jacobian for the n-th observation. Summing the above expression over n yields the regularization function:

$$\sum_n \Omega_n = \frac{1}{2}\sum_n \left(\sum_k \tau_{ni} J_{nki}\right)^2 = \Omega \tag{5.128}$$

By applying the differential operator $\mathcal{G}$ on the forward propagation equations (5.203), one obtains the same results as (5.204):

$$\alpha_j = \mathcal{G}z_j = \sum_i \tau_i \frac{\partial z_j}{\partial x_i} = \sum_i \tau_i \frac{\partial z_j}{\partial a_j}\frac{\partial a_j}{\partial x_i} = h'(a_j)\sum_i \tau_i \frac{\partial a_j}{\partial x_i} = h'(a_j)\mathcal{G}a_j = h'(a_j)\beta_j$$

$$\beta_j = \mathcal{G}a_j = \sum_i \tau_i \frac{\partial}{\partial x_i}\left(\sum_{i'} w_{ji'} z_{i'}\right) = \sum_i \tau_i \sum_{i'} w_{ji'}\frac{\partial z_{i'}}{\partial x_i} = \sum_{i'} w_{ji'}\mathcal{G}z_{i'} = \sum_{i'} w_{ji'}\alpha_{i'}$$

□

We notice that we can rewrite $\mathcal{G}y_k$ as

$$\begin{aligned}
\mathcal{G}y_k &= \sum_i \tau_i \frac{\partial y_k}{\partial x_i} \\
&= \sum_i \tau_i \frac{\partial y_k}{\partial a_k}\sum_j \frac{\partial a_k}{\partial z_j}\frac{\partial z_j}{\partial x_i} \\
&= h'(a_k)\sum_i \tau_i \sum_j w_{kj}\frac{\partial z_j}{\partial x_i} \\
&= h'(a_k)\sum_j w_{kj}\sum_i \tau_i \frac{\partial z_j}{\partial x_i} \\
&= h'(a_k)\sum_j w_{kj}\mathcal{G}z_j \\
&= h'(a_k)\beta_k \\
&= \alpha_k
\end{aligned}$$

Since α_j can be obtained from a_j and β_j (see (5.204)), this is proof that Ω_n can be evaluated by forward propagation using the equations (5.204) successively. We can see that the derivative of the differential operator can be written as

$$\frac{\partial \mathcal{G}f}{\partial \gamma} = \frac{\partial}{\partial \gamma}\sum_i \tau_i \frac{\partial f}{\partial x_i} = \sum_i \tau_i \frac{\partial}{\partial x_i}\frac{\partial f}{\partial \gamma} = \mathcal{G}\left(\frac{\partial f}{\partial \gamma}\right)$$

Therefore, the derivatives of the regularizers Ω_n with respect to a weight w_{rs} can be written as

$$\frac{\partial \Omega_n}{\partial w_{rs}} = \frac{\partial}{\partial w_{rs}}\left(\frac{1}{2}\sum_k \left(\mathcal{G}y_k\right)^2\right) = \sum_k \mathcal{G}y_k \frac{\partial \mathcal{G}y_k}{\partial w_{rs}} = \sum_k \alpha_k \mathcal{G}\left(\frac{\partial y_k}{\partial w_{rs}}\right) = \sum_k \alpha_k \mathcal{G}\left(\frac{\partial y_k}{\partial a_r}\frac{\partial a_r}{\partial w_{rs}}\right)$$

$$= \sum_k \alpha_k \mathcal{G}\left(\frac{\partial y_k}{\partial a_r} z_s\right)$$

Applying the product rule on the obtained expression and substituting the variables defined in (5.207) yields

$$\frac{\partial \Omega_n}{\partial w_{rs}} = \sum_k \alpha_k \{z_s \mathcal{G}\delta_{kr} + \delta_{kr}\mathcal{G}z_s\} = \sum_k \alpha_k \{\phi_{kr} z_s + \delta_{kr}\alpha_s\} \tag{5.206}$$

The backpropagation formula for δ_{kr} is obtained similarly to the one in Section 5.3.1. Suppose that the units l come after the units r. Then,

$$\delta_{kr} = \sum_l \frac{\partial y_k}{\partial a_l}\frac{\partial a_l}{\partial a_r} = \sum_l \frac{\partial y_k}{\partial a_l}\frac{\partial a_l}{\partial z_r}\frac{\partial z_r}{\partial a_r} = h'(a_r)\sum_l w_{lr}\delta_{kl}$$

As a result, the backpropagation equations for ϕ_{kr} can be obtained by applying the differential operator $\mathcal{G}$ on this result:

$$\begin{aligned}
\phi_{kr} &= \mathcal{G}\delta_{kr} \\
&= \mathcal{G}\left(h'(a_r)\sum_l w_{lr}\delta_{kl}\right) \\
&= \mathcal{G}\big(h'(a_r)\big)\sum_l w_{lr}\delta_{kl} + h'(a_r)\sum_l w_{lr}\mathcal{G}(\delta_{kl})
\end{aligned}$$

Exercise 5.27 ⋆⋆

Consider the framework for training with transformed data in the special case in which the transformation consists simply of the addition of random noise $\mathbf{x} \to \mathbf{x} + \boldsymbol{\xi}$ where $\boldsymbol{\xi}$ has a Gaussian distribution with zero mean and unit covariance. By following an argument analogous to that of Section 5.5.5, show that the resulting regularizer reduces to the Tikhonov from (5.1).

Proof. □