Pattern Recognition and Machine Learning
Cristopher Bishop

Exercise Solutions

Stefan Stefanache

August 29, 2021

Chapter 1

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

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{e^{2x} - 1}{e^{2x} + 1} \tag{1.1}$$

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

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

which we can translate into matrix notation as:

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

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

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

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

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

□

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

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

Therefore, we have that

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

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

Exercise 3.5 ⋆

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

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

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

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

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

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

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

□

Exercise 3.6 $\star$

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

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

where

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

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

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

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

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

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

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

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

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

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

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

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 $\mathbf{W}_{\mathrm{ML}}$, we'd find that they are of the form (3.15), i.e. the n^{th} column of $\mathbf{W}_{\mathrm{ML}}$ is given by

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

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:

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

By replacing back into (3.7.1),

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

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

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

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

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

which gives

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

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

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

□

Exercise 3.8 ⋆⋆

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

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

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

, so

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

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

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

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

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

Substituting these results back into (3.8.1) yields:

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

which is equivalent to

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

for

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

and

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

□

Exercise 3.9 ⋆⋆

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

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

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

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

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

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

where

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

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

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

and the likelihood is

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

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

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

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

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

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

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

□

Exercise 3.10

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

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

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

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

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

so the 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)

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

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

$$\sigma^2_{N+1}(\mathbf{x})\leq\sigma^2_N(\mathbf{x}) \tag{3.111}$$

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

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

so

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

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:

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

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

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

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

so

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

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

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

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

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

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

$$\begin{aligned}\sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) = \sum_{n=1}^{N} \boldsymbol{\psi}(\mathbf{x})^T\boldsymbol{\psi}(\mathbf{x}_n) = \sum_{n=1}^{N}\sum_{j=0}^{M-1} \psi_j(\mathbf{x})\psi_j(\mathbf{x}_n) &= \sum_{j=0}^{M-1} \psi_j(\mathbf{x}) \sum_{n=1}^{N} \psi_j(\mathbf{x}_n)\psi_0(\mathbf{x}_n) \\ &= \sum_{j=0}^{M-1} \psi_j(\mathbf{x})\mathbf{I}_{j+1,1} = 1\end{aligned}$$

□

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

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

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

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

the precision matrix of this Gaussian will be given by

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

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:

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

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 $E(\mathbf{w})$:

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

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

$$\begin{aligned}&\int \exp\{-E(\mathbf{w})\}\,\mathrm{d}\mathbf{w} \\ &= \exp\{-E(\mathbf{m}_N)\}(2\pi)^{M/2}|\mathbf{A}|^{-1/2}\int \frac{1}{(2\pi)^{M/2}}\frac{1}{|\mathbf{A}|^{-1/2}}\exp\left\{\frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^T\mathbf{A}(\mathbf{w}-\mathbf{m}_N)\right\}\mathrm{d}\mathbf{w} \\ &= \exp\{-E(\mathbf{m}_N)\}(2\pi)^{M/2}|\mathbf{A}|^{-1/2}\int \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{A}^{-1})\,\mathrm{d}\mathbf{w} \\ &= \exp\{-E(\mathbf{m}_N)\}(2\pi)^{M/2}|\mathbf{A}|^{-1/2} \qquad (3.85)\end{aligned}$$

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

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

□

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

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

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

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

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 3.23

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 $\mathcal{N}(\mathbf{t}|\mathbf{\Phi w},\beta^{-1}\mathbf{I})$, so the marginal probability becomes

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

By expanding the three distributions, we have that

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

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

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

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,

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

and since both the Gaussian and Gamma distributions are normalized, the marginal probability

finally becomes what we wanted:

$$
\begin{aligned}
p(\mathbf{t}) &= \frac{1}{(2\pi)^{\frac{N+M}{2}}}\frac{b_0^{a_0}}{\Gamma(a_0)}\frac{1}{|\mathbf{S}_0|^{1/2}}\int \beta^{a_0-1+N/2+M/2}\int \exp\{\beta b_N\}\exp\left\{-\frac{\beta}{2}(\mathbf{w}-\mathbf{m}_N)^T\mathbf{S}_N^{-1}(\mathbf{w}-\mathbf{m}_N)\right\}\mathrm{d}\mathbf{w}\,\mathrm{d}\beta \\
&= \frac{1}{(2\pi)^{\frac{N+M}{2}}}\frac{b_0^{a_0}}{\Gamma(a_0)}\frac{1}{|\mathbf{S}_0|^{1/2}}\int \beta^{a_0-1+N/2+M/2}\int \left(\frac{2\pi}{\beta}\right)^{M/2}|\mathbf{S}_N|^{1/2}\exp\{\beta b_N\}\mathcal{N}(\mathbf{w}|\mathbf{m}_N,\beta^{-1}\mathbf{S}_N)\,\mathrm{d}\mathbf{w}\,\mathrm{d}\beta \\
&= \frac{1}{(2\pi)^{\frac{N}{2}}}\frac{b_0^{a_0}}{\Gamma(a_0)}\frac{|\mathbf{S}_N|^{1/2}}{|\mathbf{S}_0|^{1/2}}\int \beta^{a_0-1+N/2}\exp\{\beta b_N\}\int \mathcal{N}(\mathbf{w}|\mathbf{m}_N,\beta^{-1}\mathbf{S}_N)\,\mathrm{d}\mathbf{w}\,\mathrm{d}\beta \\
&= \frac{1}{(2\pi)^{\frac{N}{2}}}\frac{b_0^{a_0}}{\Gamma(a_0)}\frac{|\mathbf{S}_N|^{1/2}}{|\mathbf{S}_0|^{1/2}}\int \beta^{a_0-1+N/2}\exp\{\beta b_N\}\,\mathrm{d}\beta \\
&= \frac{1}{(2\pi)^{\frac{N}{2}}}\frac{b_0^{a_0}}{\Gamma(a_0)}\frac{|\mathbf{S}_N|^{1/2}}{|\mathbf{S}_0|^{1/2}}\int \frac{\Gamma(a_N)}{b_N^{a_N}}\mathrm{Gam}(\beta|a_N,b_N)\,\mathrm{d}\beta \\
&= \frac{1}{(2\pi)^{N/2}}\frac{b_0^{a_0}}{b_N^{a_N}}\frac{\Gamma(a_N)}{\Gamma(a_0)}\frac{|\mathbf{S}_N|^{1/2}}{|\mathbf{S}_0|^{1/2}} \qquad (3.118)
\end{aligned}
$$

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

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

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:

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

We've seen in the previous exercise that

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

so the numerator can be written as

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

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:

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

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

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

□