1.11 We use ℓ to denote $\ln p(\mathbf{X}|\mu, \sigma^2)$ from (1.54). By standard rules of differentiation we obtain

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2} \sum_{n=1}^{N} (x_n - \mu).$$

Setting this equal to zero and moving the terms involving μ to the other side of the equation we get

$$\frac{1}{\sigma^2} \sum_{n=1}^{N} x_n = \frac{1}{\sigma^2} N\mu$$

and by multiplying ing both sides by σ^2/N we get (1.55).

Similarly we have

$$\frac{\partial \ell}{\partial \sigma^2} = \frac{1}{2(\sigma^2)^2} \sum_{n=1}^{N} (x_n - \mu)^2 - \frac{N}{2} \frac{1}{\sigma^2}$$

and setting this to zero we obtain

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

Multiplying both sides by $2(\sigma^2)^2/N$ and substituting μ_{ML} for μ we get (1.56).

3.1 **NOTE**: In the 1st printing of PRML, there is a 2 missing in the denominator of the argument to the '$\tanh$' function in equation (3.102).

Using (3.6), we have

$$\begin{aligned}
2\sigma(2a) - 1 &= \frac{2}{1 + e^{-2a}} - 1 \\
&= \frac{2}{1 + e^{-2a}} - \frac{1 + e^{-2a}}{1 + e^{-2a}} \\
&= \frac{1 - e^{-2a}}{1 + e^{-2a}} \\
&= \frac{e^{a} - e^{-a}}{e^{a} + e^{-a}} \\
&= \tanh(a)
\end{aligned}$$

3.3 If we define $\mathbf{R} = \mathrm{diag}(r_1, \ldots, r_N)$ to be a diagonal matrix containing the weighting coefficients, then we can write the weighted sum-of-squares cost function in the form

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

Setting the derivative with respect to $\mathbf{w}$ to zero, and re-arranging, then gives

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

which reduces to the standard solution (3.15) for the case $\mathbf{R} = \mathbf{I}$.

If we compare (3.104) with (3.10)–(3.12), we see that r_n can be regarded as a precision (inverse variance) parameter, particular to the data point $(\mathbf{x}_n, t_n)$, that either replaces or scales β.

Alternatively, r_n can be regarded as an *effective* number of replicated observations of data point $(\mathbf{x}_n, t_n)$; this becomes particularly clear if we consider (3.104) with r_n taking positive integer values, although it is valid for any $r_n > 0$.

3.7 From Bayes' theorem we have

$$p(\mathbf{w}|\mathbf{t}) \propto p(\mathbf{t}|\mathbf{w})p(\mathbf{w}),$$

where the factors on the r.h.s. are given by (3.10) and (3.48), respectively. Writing this out in full, we get

$$\begin{aligned}
p(\mathbf{w}|\mathbf{t}) &\propto \left[\prod_{n=1}^{N} \mathcal{N}\left(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right)\right] \mathcal{N}\left(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0\right) \\
&\propto \exp\left(-\frac{\beta}{2}(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w})^{\mathrm{T}}(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w})\right) \\
&\qquad \exp\left(-\frac{1}{2}(\mathbf{w} - \mathbf{m}_0)^{\mathrm{T}}\mathbf{S}_0^{-1}(\mathbf{w} - \mathbf{m}_0)\right) \\
&= \exp\Bigg(-\frac{1}{2}\Big(\mathbf{w}^{\mathrm{T}}\left(\mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)\mathbf{w} - \beta\mathbf{t}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{w} - \beta\mathbf{w}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t} + \beta\mathbf{t}^{\mathrm{T}}\mathbf{t} \\
&\qquad \mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{w} - \mathbf{w}^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0 + \mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0\Big)\Bigg) \\
&= \exp\Bigg(-\frac{1}{2}\Big(\mathbf{w}^{\mathrm{T}}\left(\mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)\mathbf{w} - \left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}\right)^{\mathrm{T}}\mathbf{w} \\
&\qquad -\mathbf{w}^{\mathrm{T}}\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}\right) + \beta\mathbf{t}^{\mathrm{T}}\mathbf{t} + \mathbf{m}_0^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0\Big)\Bigg)
\end{aligned}$$

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

where we have used (3.50) and (3.51) when completing the square in the last step. The first exponential corrsponds to the posterior, unnormalized Gaussian distribution over $\mathbf{w}$, while the second exponential is independent of $\mathbf{w}$ and hence can be absorbed into the normalization factor.

3.8 Combining the prior

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

and the likelihood

$$p(t_{N+1}|\mathbf{x}_{N+1}, \mathbf{w}) = \left(\frac{\beta}{2\pi}\right)^{1/2}\exp\left(-\frac{\beta}{2}(t_{N+1}-\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_{N+1})^2\right) \tag{130}$$

where $\boldsymbol{\phi}_{N+1} = \boldsymbol{\phi}(\mathbf{x}_{N+1})$, we obtain a posterior of the form

$$p(\mathbf{w}|t_{N+1}, \mathbf{x}_{N+1}, \mathbf{m}_N, \mathbf{S}_N) \\ \propto \exp\left(-\frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^{\mathrm{T}}\mathbf{S}_N^{-1}(\mathbf{w}-\mathbf{m}_N)-\frac{1}{2}\beta(t_{N+1}-\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_{N+1})^2\right).$$

We can expand the argument of the exponential, omitting the $-1/2$ factors, as follows

$$\begin{aligned}
(\mathbf{w}-\mathbf{m}_N)^{\mathrm{T}}&\mathbf{S}_N^{-1}(\mathbf{w}-\mathbf{m}_N)+\beta(t_{N+1}-\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_{N+1})^2 \\
&= \mathbf{w}^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{w}-2\mathbf{w}^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N \\
&\quad + \beta\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_{N+1}^{\mathrm{T}}\boldsymbol{\phi}_{N+1}\mathbf{w}-2\beta\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_{N+1}t_{N+1}+\text{const} \\
&= \mathbf{w}^{\mathrm{T}}(\mathbf{S}_N^{-1}+\beta\boldsymbol{\phi}_{N+1}\boldsymbol{\phi}_{N+1}^{\mathrm{T}})\mathbf{w}-2\mathbf{w}^{\mathrm{T}}(\mathbf{S}_N^{-1}\mathbf{m}_N+\beta\boldsymbol{\phi}_{N+1}t_{N+1})+\text{const},
\end{aligned}$$

where const denotes remaining terms independent of $\mathbf{w}$. From this we can read off the desired result directly,

$$p(\mathbf{w}|t_{N+1}, \mathbf{x}_{N+1}, \mathbf{m}_N, \mathbf{S}_N) = \mathcal{N}(\mathbf{w}|\mathbf{m}_{N+1}, \mathbf{S}_{N+1}),$$

with

$$\mathbf{S}_{N+1}^{-1} = \mathbf{S}_N^{-1}+\beta\boldsymbol{\phi}_{N+1}\boldsymbol{\phi}_{N+1}^{\mathrm{T}}. \tag{131}$$

and

$$\mathbf{m}_{N+1} = \mathbf{S}_{N+1}(\mathbf{S}_N^{-1}\mathbf{m}_N+\beta\boldsymbol{\phi}_{N+1}t_{N+1}). \tag{132}$$