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

$$y(x, \mathbf{u}) = u_0 + \sum_{j=1}^{M} u_j \tanh\left(\frac{x - \mu_j}{2s}\right) \tag{4.58}$$

and find expressions to relate the new parameters $\{u_1, \ldots, u_M\}$ to the original parameters $\{w_1, \ldots, w_M\}$.

4.4 ($\star\star\star$) Show that the matrix

$$\mathbf{\Phi}(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi})^{-1}\mathbf{\Phi}^{\mathrm{T}} \tag{4.59}$$

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 (4.14) corresponds to an orthogonal projection of the vector $\mathbf{t}$ onto the manifold $\mathcal{S}$, as shown in Figure 4.3.

4.5 ($\star$) 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

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} r_n \left\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\right\}^2. \tag{4.60}$$

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.

4.6 ($\star$) By setting the gradient of (4.26) with respect to $\mathbf{w}$ to zero, show that the exact minimum of the regularized sum-of-squares error function for linear regression is given by (4.27).

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

$$p(\mathbf{t}|\mathbf{W}, \mathbf{\Sigma}) = \mathcal{N}(\mathbf{t}|\mathbf{y}(\mathbf{x}, \mathbf{W}), \mathbf{\Sigma}) \tag{4.61}$$

where

$$\mathbf{y}(\mathbf{x}, \mathbf{W}) = \mathbf{W}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}) \tag{4.62}$$

together with a training data set comprising 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 $\mathbf{W}_{\mathrm{ML}}$ for the parameter matrix $\mathbf{W}$ has the property that each column is given by an expression of the form (4.14), which was the solution for an isotropic noise distribution. Note that this is independent of the covariance matrix $\mathbf{\Sigma}$. Show that the maximum likelihood solution for $\mathbf{\Sigma}$ is given by

$$\mathbf{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{t}_n - \mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\right)\left(\mathbf{t}_n - \mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\right)^{\mathrm{T}}. \tag{4.63}$$

4.8 ($\star$) Consider the generalization of the squared-loss function (4.35) for a single target variable t to multiple target variables described by the vector $\mathbf{t}$ given by

$$\mathbb{E}[L(\mathbf{t}, \mathbf{f}(\mathbf{x}))] = \iint \|\mathbf{f}(\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{t}. \tag{4.64}$$

Using the calculus of variations, show that the function $\mathbf{f}(\mathbf{x})$ for which this expected loss is minimized is given by

$$\mathbf{f}(\mathbf{x}) = \mathbb{E}_t[\mathbf{t}|\mathbf{x}]. \tag{4.65}$$

4.9 ($\star$) By expansion of the square in (4.64), derive a result analogous to (4.39) and, hence, show that the function $\mathbf{f}(\mathbf{x})$ that minimizes the expected squared loss for a vector $\mathbf{t}$ of target variables is again given by the conditional expectation of $\mathbf{t}$ in the form (4.65).

4.10 ($\star\star$) Rederive the result (4.65) by first expanding (4.64) analogous to (4.39).

4.11 ($\star\star$) The following distribution

$$p(x|\sigma^2, q) = \frac{q}{2(2\sigma^2)^{1/q}\Gamma(1/q)} \exp\left(-\frac{|x|^q}{2\sigma^2}\right) \tag{4.66}$$

is a generalization of the univariate Gaussian distribution. Here $\Gamma(x)$ is the gamma function defined by

$$\Gamma(x) = \int_{-\infty}^{\infty} u^{x-1} e^{-u} \, \mathrm{d}u. \tag{4.67}$$

Show that this distribution is normalized so that

$$\int_{-\infty}^{\infty} p(x|\sigma^2, q) \, \mathrm{d}x = 1 \tag{4.68}$$

and that it reduces to the Gaussian when $q = 2$. Consider a regression model in which the target variable is given by $t = y(\mathbf{x}, \mathbf{w}) + \epsilon$ and ϵ is a random noise variable drawn from the distribution (4.66). Show that the log likelihood function over $\mathbf{w}$ and σ^2, for an observed data set of input vectors $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and corresponding target variables $\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$, is given by

$$\ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{n=1}^{N} |y(\mathbf{x}_n, \mathbf{w}) - t_n|^q - \frac{N}{q} \ln(2\sigma^2) + \text{const} \tag{4.69}$$

where 'const' denotes terms independent of both $\mathbf{w}$ and σ^2. Note that, as a function of $\mathbf{w}$, this is the L_q error function considered in Section 4.2.

4.12 ($\star\star$) Consider the expected loss for regression problems under the L_q loss function given by (4.40). Write down the condition that $y(\mathbf{x})$ must satisfy 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 as for $t \geqslant y(\mathbf{x})$. Also show that the minimum expected L_q loss for $q \to 0$ is given by the conditional mode, i.e., by the function $y(\mathbf{x})$ being equal to the value of t that maximizes $p(t|\mathbf{x})$ for each $\mathbf{x}$.