

Now consider a data set of inputs  $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$  with corresponding target values  $t_1, \dots, t_N$ . We group the target variables  $\{t_n\}$  into a column vector that we denote by  $\mathbf{t}$  where the typeface is chosen to distinguish it from a single observation of a multivariate target, which would be denoted  $\mathbf{t}$ . Making the assumption that these data points are drawn independently from the distribution (4.8), we obtain an expression for the likelihood function, which is a function of the adjustable parameters  $\mathbf{w}$  and  $\sigma^2$ :

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^N \mathcal{N}(t_n | \mathbf{w}^T \phi(\mathbf{x}_n), \sigma^2) \quad (4.9)$$

where we have used (4.3). Taking the logarithm of the likelihood function and making use of the standard form (2.49) for the univariate Gaussian, we have

$$\begin{aligned} \ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) &= \sum_{n=1}^N \ln \mathcal{N}(t_n | \mathbf{w}^T \phi(\mathbf{x}_n), \sigma^2) \\ &= -\frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi) - \frac{1}{\sigma^2} E_D(\mathbf{w}) \end{aligned} \quad (4.10)$$

where the sum-of-squares error function is defined by

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{t_n - \mathbf{w}^T \phi(\mathbf{x}_n)\}^2. \quad (4.11)$$

The first two terms in (4.10) can be treated as constants when determining  $\mathbf{w}$  because they are independent of  $\mathbf{w}$ . Therefore, as we saw previously, maximizing the likelihood function under a Gaussian noise distribution is equivalent to minimizing the sum-of-squares error function (4.11).

Section 2.3.4

### 4.1.3 Maximum likelihood

Having written down the likelihood function, we can use maximum likelihood to determine  $\mathbf{w}$  and  $\sigma^2$ . Consider first the maximization with respect to  $\mathbf{w}$ . The gradient of the log likelihood function (4.10) with respect to  $\mathbf{w}$  takes the form

$$\nabla_{\mathbf{w}} \ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \frac{1}{\sigma^2} \sum_{n=1}^N \{t_n - \mathbf{w}^T \phi(\mathbf{x}_n)\} \phi(\mathbf{x}_n)^T. \quad (4.12)$$

Setting this gradient to zero gives

$$0 = \sum_{n=1}^N t_n \phi(\mathbf{x}_n)^T - \mathbf{w}^T \left( \sum_{n=1}^N \phi(\mathbf{x}_n) \phi(\mathbf{x}_n)^T \right). \quad (4.13)$$

Solving for  $\mathbf{w}$  we obtain

$$\mathbf{w}_{\text{ML}} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{t}, \quad (4.14)$$