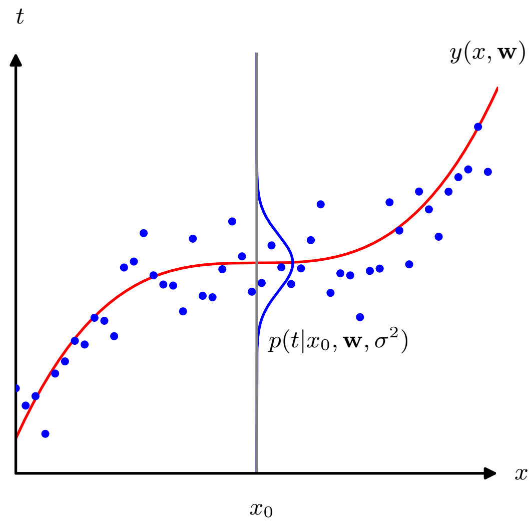

Figure 2.11 Schematic illustration of a Gaussian conditional distribution for t given x, defined by (2.64), in which the mean is given by the polynomial function $y(x, \mathbf{w})$, and the variance is given by the parameter σ^2.

independently from the distribution (2.64), then the likelihood function is given by

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(t_n|y(x_n, \mathbf{w}), \sigma^2\right). \tag{2.65}$$

As we did for the simple Gaussian distribution earlier, it is convenient to maximize the logarithm of the likelihood function. Substituting for the Gaussian distribution, given by (2.49), we obtain the log likelihood function in the form

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi). \tag{2.66}$$

Consider first the evaluation of the maximum likelihood solution for the polynomial coefficients, which will be denoted by $\mathbf{w}_{\mathrm{ML}}$. These are determined by maximizing (2.66) with respect to $\mathbf{w}$. For this purpose, we can omit the last two terms on the right-hand side of (2.66) because they do not depend on $\mathbf{w}$. Also, note that scaling the log likelihood by a positive constant coefficient does not alter the location of the maximum with respect to $\mathbf{w}$, and so we can replace the coefficient $1/2\sigma^2$ with $1/2$. Finally, instead of maximizing the log likelihood, we can equivalently minimize the negative log likelihood. We therefore see that maximizing the likelihood is equivalent, so far as determining $\mathbf{w}$ is concerned, to minimizing the *sum-of-squares error function* defined by

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2. \tag{2.67}$$

Thus, the sum-of-squares error function has arisen as a consequence of maximizing the likelihood under the assumption of a Gaussian noise distribution.