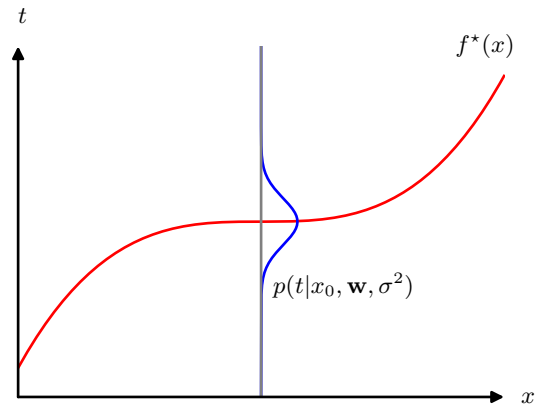

Figure 4.5 The regression function $f^\star(x)$, which minimizes the expected squared loss, is given by the mean of the conditional distribution $p(t|x)$.

given by

$$\mathbb{E}[L] = \iint L(t, f(\mathbf{x}))p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t \tag{4.34}$$

where we are averaging over the distribution of both input and target variables, weighted by their joint distribution $p(\mathbf{x}, t)$. A common choice of loss function in regression problems is the squared loss given by $L(t, f(\mathbf{x})) = \{f(\mathbf{x}) - t\}^2$. In this case, the expected loss can be written

$$\mathbb{E}[L] = \iint \{f(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t. \tag{4.35}$$

It is important not to confuse the squared-loss function with the sum-of-squares error function introduced earlier. The error function is used to set the parameters during training in order to determine the conditional probability distribution $p(t|\mathbf{x})$, whereas the loss function governs how the conditional distribution is used to arrive at a predictive function $f(\mathbf{x})$ specifying a prediction for each value of $\mathbf{x}$.

Appendix B Our goal is to choose $f(\mathbf{x})$ so as to minimize $\mathbb{E}[L]$. If we assume a completely flexible function $f(\mathbf{x})$, we can do this formally using the calculus of variations to give

$$\frac{\delta \mathbb{E}[L]}{\delta f(\mathbf{x})} = 2\int \{f(\mathbf{x}) - t\}p(\mathbf{x}, t)\,\mathrm{d}t = 0. \tag{4.36}$$

Solving for $f(\mathbf{x})$ and using the sum and product rules of probability, we obtain

$$f^\star(\mathbf{x}) = \frac{1}{p(\mathbf{x})}\int tp(\mathbf{x}, t)\,\mathrm{d}t = \int tp(t|\mathbf{x})\,\mathrm{d}t = \mathbb{E}_t[t|\mathbf{x}], \tag{4.37}$$

which is the conditional average of t conditioned on $\mathbf{x}$ and is known as the *regression function*. This result is illustrated in Figure 4.5. It can readily be extended to multiple target variables represented by the vector $\mathbf{t}$, in which case the optimal solution is the *Exercise 4.8* conditional average $\mathbf{f}^\star(\mathbf{x}) = \mathbb{E}_t[\mathbf{t}|\mathbf{x}]$. For a Gaussian conditional distribution of the

form (4.8), the conditional mean will be simply

$$\mathbb{E}[t|\mathbf{x}] = \int tp(t|\mathbf{x})\,\mathrm{d}t = y(\mathbf{x}, \mathbf{w}). \tag{4.38}$$

The use of calculus of variations to derive (4.37) implies that we are optimizing over all possible functions $f(\mathbf{x})$. Although any parametric model that we can implement in practice is limited in the range of functions that it can represent, the framework of deep neural networks, discussed extensively in later chapters, provides a highly flexible class of functions that, for many practical purposes, can approximate any desired function to high accuracy.

We can derive this result in a slightly different way, which will also shed light on the nature of the regression problem. Armed with the knowledge that the optimal solution is the conditional expectation, we can expand the square term as follows

$$\begin{aligned}\{f(\mathbf{x}) - t\}^2 &= \{f(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}] + \mathbb{E}[t|\mathbf{x}] - t\}^2 \\ &= \{f(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 + 2\{f(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}\{\mathbb{E}[t|\mathbf{x}] - t\} + \{\mathbb{E}[t|\mathbf{x}] - t\}^2\end{aligned}$$

where, to keep the notation uncluttered, we use $\mathbb{E}[t|\mathbf{x}]$ to denote $\mathbb{E}_t[t|\mathbf{x}]$. Substituting into the loss function (4.35) and performing the integral over t, we see that the cross-term vanishes and we obtain an expression for the loss function in the form

$$\mathbb{E}[L] = \int \{f(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x} + \int \mathrm{var}\,[t|\mathbf{x}]\, p(\mathbf{x})\,\mathrm{d}\mathbf{x}. \tag{4.39}$$

The function $f(\mathbf{x})$ we seek to determine appears only in the first term, which will be minimized when $f(\mathbf{x})$ is equal to $\mathbb{E}[t|\mathbf{x}]$, in which case this term will vanish. This is simply the result that we derived previously, and shows that the optimal least-squares predictor is given by the conditional mean. The second term is the variance of the distribution of t, averaged over $\mathbf{x}$, and represents the intrinsic variability of the target data and can be regarded as noise. Because it is independent of $f(\mathbf{x})$, it represents the irreducible minimum value of the loss function.

The squared loss is not the only possible choice of loss function for regression. Here we consider briefly one simple generalization of the squared loss, called the *Minkowski* loss, whose expectation is given by

$$\mathbb{E}[L_q] = \iint |f(\mathbf{x}) - t|^q p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t, \tag{4.40}$$

which reduces to the expected squared loss for $q = 2$. The function $|f - t|^q$ is plotted against $f - t$ for various values of q in Figure 4.6. The minimum of $\mathbb{E}[L_q]$ is given by the conditional mean for $q = 2$, the conditional median for $q = 1$, and the conditional mode for $q \to 0$.

Exercise 4.12

Note that the Gaussian noise assumption implies that the conditional distribution of t given $\mathbf{x}$ is unimodal, which may be inappropriate for some applications. In this case a squared loss can lead to very poor results and we need to develop more sophisticated approaches. For example, we can extend this model by using mixtures