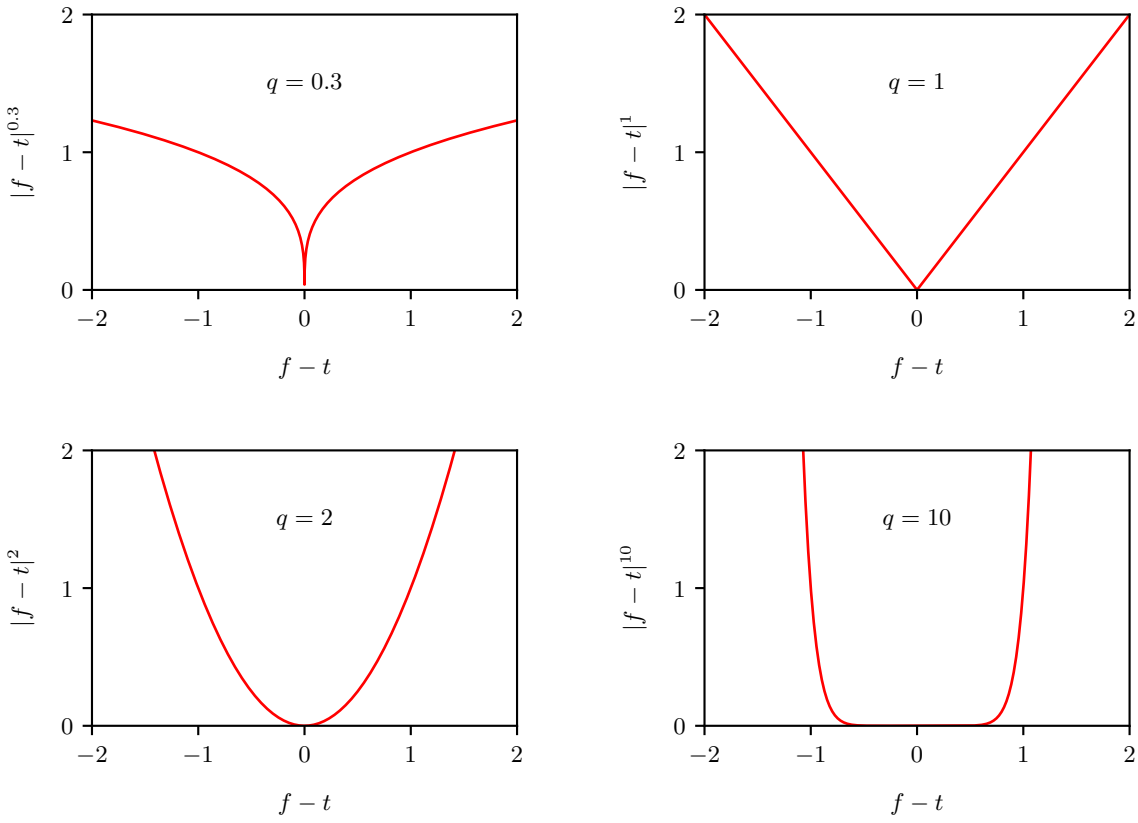

Figure 4.6 Plots of the quantity $L_q = |f - t|^q$ for various values of q.

Section 6.5 of Gaussians to give multimodal conditional distributions, which often arise in the solution of *inverse problems*. Our focus in this section has been on decision theory for regression problems, and in the next chapter we shall develop analogous concepts for classification tasks. *Section 5.2*

4.3. The Bias–Variance Trade-off

Section 1.2 So far in our discussion of linear models for regression, we have assumed that the form and number of basis functions are both given. We have also seen that the use of maximum likelihood can lead to severe over-fitting if complex models are trained using data sets of limited size. However, limiting the number of basis functions to avoid over-fitting has the side effect of limiting the flexibility of the model to capture interesting and important trends in the data. Although a regularization term can control over-fitting for models with many parameters, this raises the question of how to determine a suitable value for the regularization coefficient λ. Seeking the

solution that minimizes the regularized error function with respect to both the weight vector $\mathbf{w}$ and the regularization coefficient λ is clearly not the right approach, since this leads to the unregularized solution with $\lambda = 0$.

It is instructive to consider a frequentist viewpoint of the model complexity issue, known as the *bias–variance* trade-off. Although we will introduce this concept in the context of linear basis function models, where it is easy to illustrate the ideas using simple examples, the discussion has very general applicability. Note, however, that over-fitting is really an unfortunate property of maximum likelihood and does not arise when we marginalize over parameters in a Bayesian setting (Bishop, 2006).

Section 4.2

When we discussed decision theory for regression problems, we considered various loss functions, each of which leads to a corresponding optimal prediction once we are given the conditional distribution $p(t|\mathbf{x})$. A popular choice is the squared-loss function, for which the optimal prediction is given by the conditional expectation, which we denote by $h(\mathbf{x})$ and is given by

$$h(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}] = \int t p(t|\mathbf{x})\,\mathrm{d}t. \tag{4.41}$$

We have also seen that the expected squared loss can be written in the form

$$\mathbb{E}[L] = \int \{f(\mathbf{x}) - h(\mathbf{x})\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x} + \iint \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t. \tag{4.42}$$

Recall that the second term, which is independent of $f(\mathbf{x})$, arises from the intrinsic noise on the data and represents the minimum achievable value of the expected loss. The first term depends on our choice for the function $f(\mathbf{x})$, and we will seek a solution for $f(\mathbf{x})$ that makes this term a minimum. Because it is non-negative, the smallest value that we can hope to achieve for this term is zero. If we had an unlimited supply of data (and unlimited computational resources), we could in principle find the regression function $h(\mathbf{x})$ to any desired degree of accuracy, and this would represent the optimal choice for $f(\mathbf{x})$. However, in practice we have a data set $\mathcal{D}$ containing only a finite number N of data points, and consequently, we cannot know the regression function $h(\mathbf{x})$ exactly.

If we were to model $h(\mathbf{x})$ using a function governed by a parameter vector $\mathbf{w}$, then from a Bayesian perspective, the uncertainty in our model would be expressed through a posterior distribution over $\mathbf{w}$. A frequentist treatment, however, involves making a point estimate of $\mathbf{w}$ based on the data set $\mathcal{D}$ and tries instead to interpret the uncertainty of this estimate through the following thought experiment. Suppose we had a large number of data sets each of size N and each drawn independently from the distribution $p(t, \mathbf{x})$. For any given data set $\mathcal{D}$, we can run our learning algorithm and obtain a prediction function $f(\mathbf{x}; \mathcal{D})$. Different data sets from the ensemble will give different functions and consequently different values of the squared loss. The performance of a particular learning algorithm is then assessed by taking the average over this ensemble of data sets.

Consider the integrand of the first term in (4.42), which for a particular data set $\mathcal{D}$ takes the form

$$\{f(\mathbf{x}; \mathcal{D}) - h(\mathbf{x})\}^2. \tag{4.43}$$