We can also use maximum likelihood to determine the variance parameter σ^2. Maximizing (2.66) with respect to σ^2 gives

Exercise 2.18

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}_{\mathrm{ML}}) - t_n\}^2 . \tag{2.68}$$

Note that we can first determine the parameter vector $\mathbf{w}_{\mathrm{ML}}$ governing the mean, and subsequently use this to find the variance σ^2_{ML} as was the case for the simple Gaussian distribution.

Having determined the parameters $\mathbf{w}$ and σ^2, we can now make predictions for new values of x. Because we now have a probabilistic model, these are expressed in terms of the *predictive distribution* that gives the probability distribution over t, rather than simply a point estimate, and is obtained by substituting the maximum likelihood parameters into (2.64) to give

$$p(t|x, \mathbf{w}_{\mathrm{ML}}, \sigma^2_{\mathrm{ML}}) = \mathcal{N}\left(t|y(x, \mathbf{w}_{\mathrm{ML}}), \sigma^2_{\mathrm{ML}}\right) . \tag{2.69}$$

2.4. Transformation of Densities

We turn now to a discussion of how a probability density transforms under a nonlinear change of variable. This property will play a crucial role when we discuss a class of generative models called *normalizing flows*. It also highlights that a probability density has a different behaviour than a simple function under such transformations.

Chapter 18

Consider a single variable x and suppose we make a change of variables $x = g(y)$, then a function $f(x)$ becomes a new function $\widetilde{f}(y)$ defined by

$$\widetilde{f}(y) = f(g(y)). \tag{2.70}$$

Now consider a probability density $p_x(x)$, and again change variables using $x = g(y)$, giving rise to a density $p_y(y)$ with respect to the new variable y, where the suffixes denote that $p_x(x)$ and $p_y(y)$ are different densities. Observations falling in the range $(x, x + \delta x)$ will, for small values of δx, be transformed into the range $(y, y + \delta y)$, where $x = g(y)$, and $p_x(x)\delta x \simeq p_y(y)\delta y$. Hence, if we take the limit $\delta x \to 0$, we obtain

$$\begin{aligned} p_y(y) &= p_x(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| \\ &= p_x(g(y))\left|\frac{\mathrm{d}g}{\mathrm{d}y}\right| . \end{aligned} \tag{2.71}$$

Here the modulus $|\cdot|$ arises because the derivative $\mathrm{d}y/\,\mathrm{d}x$ could be negative, whereas the density is scaled by the ratio of lengths, which is always positive.