### 2.5.7 Mutual information

When two variables $\mathbf{x}$ and $\mathbf{y}$ are independent, their joint distribution will factorize into the product of their marginals $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$. If the variables are not independent, we can gain some idea of whether they are 'close' to being independent by considering the Kullback–Leibler divergence between the joint distribution and the product of the marginals, given by

$$\begin{aligned} \mathrm{I}[\mathbf{x}, \mathbf{y}] &\equiv \mathrm{KL}(p(\mathbf{x}, \mathbf{y}) \| p(\mathbf{x})p(\mathbf{y})) \\ &= -\iint p(\mathbf{x}, \mathbf{y}) \ln \left( \frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} \right) \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y}, \end{aligned} \tag{2.109}$$

which is called the *mutual information* between the variables $\mathbf{x}$ and $\mathbf{y}$. From the properties of the Kullback–Leibler divergence, we see that $I[\mathbf{x}, \mathbf{y}] \geqslant 0$ with equality if, and only if, $\mathbf{x}$ and $\mathbf{y}$ are independent. Using the sum and product rules of probability, we see that the mutual information is related to the conditional entropy through

*Exercise 2.38*

$$\mathrm{I}[\mathbf{x}, \mathbf{y}] = \mathrm{H}[\mathbf{x}] - \mathrm{H}[\mathbf{x}|\mathbf{y}] = \mathrm{H}[\mathbf{y}] - \mathrm{H}[\mathbf{y}|\mathbf{x}]. \tag{2.110}$$

Thus, the mutual information represents the reduction in the uncertainty about $\mathbf{x}$ by virtue of being told the value of $\mathbf{y}$ (or vice versa). From a Bayesian perspective, we can view $p(\mathbf{x})$ as the prior distribution for $\mathbf{x}$ and $p(\mathbf{x}|\mathbf{y})$ as the posterior distribution after we have observed new data $\mathbf{y}$. The mutual information therefore represents the reduction in uncertainty about $\mathbf{x}$ as a consequence of the new observation $\mathbf{y}$.

## 2.6. Bayesian Probabilities

When we considered the bent coin in Figure 2.2, we introduced the concept of probability in terms of the frequencies of random, repeatable events, such as the probability of the coin landing concave side up. We will refer to this as the *classical* or *frequentist* interpretation of probability. We also introduced the more general *Bayesian* view, in which probabilities provide a quantification of uncertainty. In this case, our uncertainty is whether the concave side of the coin is heads or tails.

The use of probability to represent uncertainty is not an ad hoc choice but is inevitable if we are to respect common sense while making rational and coherent inferences. For example, Cox (1946) showed that if numerical values are used to represent degrees of belief, then a simple set of axioms encoding common sense properties of such beliefs leads uniquely to a set of rules for manipulating degrees of belief that are equivalent to the sum and product rules of probability. It is therefore natural to refer to these quantities as (Bayesian) probabilities.

For the bent coin we assumed, in the absence of further information, that the probability of the concave side of the coin being heads is 0.5. Now suppose we are told the results of flipping the coin a few times. Intuitively, it seems that this should provide us with some information as to whether the concave side is heads. For instance, suppose we see many more flips that land tails than land heads. Given