

information gained by observing an event of probability  $\frac{1}{e}$ . Other texts use base-2 logarithms and units called **bits** or **shannons**; information measured in bits is just a rescaling of information measured in nats.

When  $x$  is continuous, we use the same definition of information by analogy, but some of the properties from the discrete case are lost. For example, an event with unit density still has zero information, despite not being an event that is guaranteed to occur.

Self-information deals only with a single outcome. We can quantify the amount of uncertainty in an entire probability distribution using the **Shannon entropy**:

$$H(x) = \mathbb{E}_{x \sim P}[I(x)] = -\mathbb{E}_{x \sim P}[\log P(x)]. \quad (3.49)$$

also denoted  $H(P)$ . In other words, the Shannon entropy of a distribution is the expected amount of information in an event drawn from that distribution. It gives a lower bound on the number of bits (if the logarithm is base 2, otherwise the units are different) needed on average to encode symbols drawn from a distribution  $P$ . Distributions that are nearly deterministic (where the outcome is nearly certain) have low entropy; distributions that are closer to uniform have high entropy. See figure 3.5 for a demonstration. When  $x$  is continuous, the Shannon entropy is known as the **differential entropy**.

If we have two separate probability distributions  $P(x)$  and  $Q(x)$  over the same random variable  $x$ , we can measure how different these two distributions are using the **Kullback-Leibler (KL) divergence**:

$$D_{\text{KL}}(P\|Q) = \mathbb{E}_{x \sim P} \left[ \log \frac{P(x)}{Q(x)} \right] = \mathbb{E}_{x \sim P} [\log P(x) - \log Q(x)]. \quad (3.50)$$

In the case of discrete variables, it is the extra amount of information (measured in bits if we use the base 2 logarithm, but in machine learning we usually use nats and the natural logarithm) needed to send a message containing symbols drawn from probability distribution  $P$ , when we use a code that was designed to minimize the length of messages drawn from probability distribution  $Q$ .

The KL divergence has many useful properties, most notably that it is non-negative. The KL divergence is 0 if and only if  $P$  and  $Q$  are the same distribution in the case of discrete variables, or equal “almost everywhere” in the case of continuous variables. Because the KL divergence is non-negative and measures the difference between two distributions, it is often conceptualized as measuring some sort of distance between these distributions. However, it is not a true distance measure because it is not symmetric:  $D_{\text{KL}}(P\|Q) \neq D_{\text{KL}}(Q\|P)$  for some  $P$  and  $Q$ . This