

# Definition: Cross Entropy

Suppose you draw samples from a set  $\mathcal{X}$ , according to a distribution  $P$ , while *thinking* you are drawing those samples according to a distribution  $Q$ . How surprised do you expect to be? This surprisal value is expressed by the **cross entropy**, a quantity that is very closely related to the relative entropy.

## Definition: Cross Entropy

The cross entropy of two probability distributions  $P$  and  $Q$  over the same  $\mathcal{X}$  is defined by

$$H_C(P, Q) := - \sum_{x \in \mathcal{X}, P(x) > 0} P(x) \log Q(x).$$

There are a few things to note about this definition:

- Note that if  $Q(x) = 0$  for some  $x$  with  $P(x) > 0$ , then  $H_C(P, Q) = \infty$ .
- Also note that  $H_C$  is a function of the *distributions*  $P$  and  $Q$ . With regular Shannon entropy, we can be sloppy with the notation, and write  $H(X)$  instead of the more correct  $H(P)$ . Here, that sloppy notation would lead to ambiguous expressions.
- In the literature, the notation  $H(P, Q)$  is often used to denote the cross entropy. This notation can potentially be confused with the notation for joint entropy, so we use the subscript  $C$  to make the distinction explicit.

The cross entropy often pops up in topics related to machine learning: a system usually learns from a set of *training data*, from which it hypothesizes a certain distribution  $Q$  (on letter frequencies, cluster sizes, or whatever the system is learning about). When the system is released into the real world, it encounters a (possibly) different distribution  $P$ . The cross entropy  $H_C(P, Q)$  quantifies how well the system performs in the real world when it was trained on the training set.