5.1 Basics of Information Theory

When we talk about *information*, we often use the term in qualitative sense. We say things like *This is valuable information* or *We have a lack of information*. We can also make statements about some information being more helpful than other. For a long time, however, people have been unable to quantify information. The person who succeeded in this endeavour was Claude E. Shannon who with his famous 1948 article *A Mathematical Theory of Communication* single-handedly created a new discipline: Information Theory! He also revolutionised digital communication and can be seen as one of the main contributors to our modern communication systems like the telephone, the internet etc.

The beauty about information theory is that it is based on probability theory and many results from probability theory seamlessly carry over to information theory. In this chapter, we are going to discuss the bare basics of information theory. These basic will often be enough to understand many information theoretic arguments that researchers make in fields like machine learning, psychology and linguistics.

Shannon's idea of information is as simple as it is compelling. Intuitively, if we are observing a realisation of a random variable, this realisation will surprise if it is unlikely to occur according to the distribution of that random variable. However, if the probability for the realisation is very low, than on average it will not occur very often, meaning that if we sample from the RV repeatedly, we will not be surprised very often. This will be the case when the probability mass of the distribution is concentrated on only a small subset of its support.

On the other hand, we will quite often be surprised, if we cannot predict what the outcome of our next draw from the RV might be. This is exactly the case when the distribution over values of the RV is uniform. Thus, we are going to be most surprised on average if we are observing realisations of a uniformly distributed RV.

Shannon's idea was that observing RVs that cause a lot of surprises is informative because we cannot predict the outcomes and with each new outcome we have effectively learned something (namely that the i^{th} outcome took on the value that it did). Observing RVs with very concentrated distributions is not very informative under this conception because by just choosing the most probable outcome we can correctly predict most actually observed outcomes. Obviously, if I manage to predict an outcome beforehand, it's occurrence is not teaching me anything.

The goal of Shannon was to find a function that captures this intuitive idea. He eventually found it and showed that it is the only function to have properties that encompass the intuition. This function is called the **entropy** of a RV and it is simply the expected **surprisal** value.

Definition 5.1 (Surprisal) *The surprisal of an outcome $x \in supp(X)$ of some RV X is*

$$-\log(P(X = x)) \ .$$

Definition 5.2 (Entropy) *The entropy of a RV X is $H(X)$ where $H : X \to \mathbb{R}_{0+}$ is defined as*

$$H(X) := \mathbb{E}[-\log(P(X = x))] \ .$$