Chapter 7

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 basics are often enough to understand many information-theoretic arguments that researchers make in fields like computer science, psychology and linguistics.

7.1 Surprisal and Entropy

Shannon's idea of information is as simple as it is compelling. The amount of *surprisal* of an event E is defined as the inverse probability $1/P(E)$. Intuitively, rare events (where $P(E)$ is small) are more surprising than those occurring with high probability (where $P(E)$ is high). If we are observing a realisation of a random variable, this realisation is surprising if it is unlikely to occur according to the distribution of that random variable. However, if the probability for the realisation is very low, then on average it does not occur very often, meaning that if we sample from the RV repeatedly, we are not surprised very often. We are not surprised when the probability mass of the distribution is concentrated on only a small subset of its support.

On the other hand, we quite often are surprised, if we cannot predict

what the outcome of our next draw from the RV might be. We are surprised when the distribution over values of the RV is (close to) 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, its 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, expressed in bits.

Definition 7.1 (Surprisal) *The surprisal (value) of an outcome $x \in \text{supp}(X)$ of some RV X is defined as $-\log_2(P(X=x)) = \log_2(\frac{1}{P(X=x)})$.*

Notice that we are using the logarithm of base 2 here. This is because surprisal and entropy are standardly measured in bits. Intuitively, the surprisal measures how many bits one needs to encode an observed outcome given that one knows the distribution underlying that outcome. Check this website to get a feeling for surprisal values measured in bits.

Definition 7.2 (Entropy) *The entropy $H(P_X)$ of a RV X with distribution P_X is defined as*

$$H(P_X) := \mathbb{E}[-\log_2(P(X=x))] = -\sum_{x \in \text{supp}(X)} P(X=x)\log_2(P(X=x))\,.$$

For the ease of notation, we often write $H(X)$ instead of $H(P_X)$.

The notational convenience of writing $H(X)$ instead of $H(P_X)$ can be confusing, because entropy is really assigning a (non-negative) real number to a distribution, i.e. $H(X)$ is **not a function** of the random variable X and it is **not a random variable** either! Formally, for any random variable X with distribution P_X over the set $\mathcal{X} = \text{supp}(X)$ (which might be categorical, i.e. X could for instance take on values "blue", "red" and "green"), we consider the surprisal function (in bits) $f(x) := -\log_2(P(X=x))$ mapping elements $x \in \mathcal{X}$ to real numbers $f(x) \in \mathbb{R}$. In that case, the surprisal $f(X)$ is a random variable over the reals and its expected value is well defined and called entropy $H(X) = H(P_X) := \mathbb{E}_X[f(X)]$.

As an example, we consider the categorical random variable X with distribution $P(X = \heartsuit) = P(X = \clubsuit) = 1/4, P(X = \spadesuit) = 1/2$. In that case, $\text{supp}(X) = \{\heartsuit, \clubsuit, \spadesuit\}$ and surprisal values in bits are $f(\heartsuit) = f(\clubsuit) = \log_2(4) = 2, f(\spadesuit) = \log_2(2) = 1$. The entropy is the expected surprisal value, i.e. the individual surprisal valuse weighted with their corresponding probabilities of occurring: $H(X) = \mathbb{E}_X[f(X)] = \frac{1}{4} \cdot 2 + \frac{1}{4} \cdot 2 + \frac{1}{2} \cdot 1 = 3/2$.

The entropy "does not care" about the actual outcomes or labels of a random variable, but only about the distribution! In fact, not even the order of the actual probabilities matter, as we are taking an expected value and the additive terms commute. You can verify that the calculation of $H(X) = 3/2$ in the example above does apply to all random variables X with distribution $(1/2, 1/4, 1/4)$, no matter what the actual outcomes are.

Exercise 7.3 *Compute the entropy of* $Y \sim Binomial(n = 2, p = 1/2)$.

The simplest and simultaneously most important example of entropy is given in Figure 7.1 which shows the entropy of the Bernoulli distribution as a function of the parameter $\theta \in [0, 1]$. The entropy function of the Bernoulli is often called the **binary entropy** $h(\theta) := -\theta \cdot \log_2(\theta) - (1-\theta)\log_2(1-\theta)$. It measures the information of a binary decision, like a coin flip or an answer to a yes/no-question. The entropy of the Bernoulli attains its maximum of 1 bit when the distribution is uniform, i.e. when both choices are equally probable. The entropy is 0 if and only if the coin is fully biased towards heads or tails. As explained above, the entropy of the distributions $(\theta, 1-\theta)$ and $(1-\theta, \theta)$ is the same and therefore $h(\theta) = h(1-\theta)$ and the graph is symmetric around $1/2$.

From the plot is it also easy to see that entropy is never negative. It holds in general that entropy is non-negative, because entropy is defined as expectation of surprisal and surprisal is the negative logarithm of probabilities. Because $\log(x) \leq 0$ for $x \in (0, 1]$, it is clear that $-\log(x) \geq 0$ for x in the same interval. Notice that from here on we drop the subscript and by convention let $\log = \log_2$.

A standard interpretation of the entropy is that it quantifies uncertainty. As we have pointed out before, a uniform distribution means that you are most uncertain and indeed the uniform distribution maximizes the entropy. However, the more choices you have to pick from uniformly, the more uncertain you are going to be. The entropy function also captures this intuition. Notice that if a discrete distribution is uniform, all probabilities are $\frac{1}{|\text{supp}(X)|}$. Clearly, as we increase $|\text{supp}(X)|$, we decrease the probabilities. By decreasing the probabilities, we increase their negative logarithms, and hence their average surprisal. Let us make this intuition more formal.

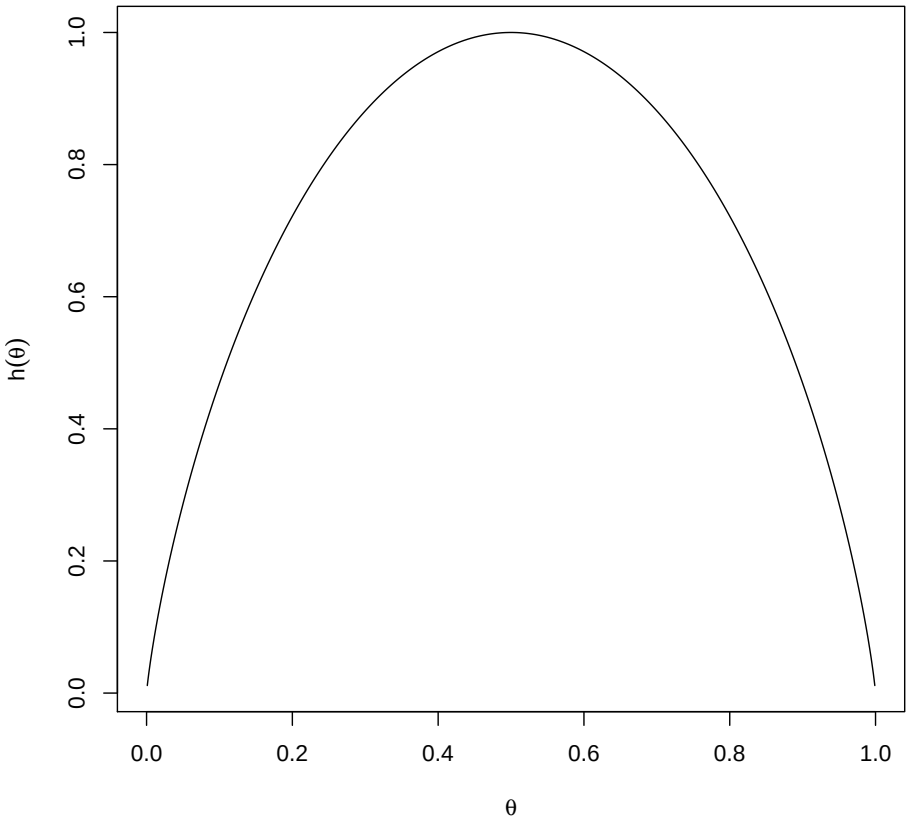

Figure 7.1: Binary entropy function

Theorem 7.4 *A discrete RV X with uniform distribution and support of size n has entropy $H(X) = \log(n)$.*

Proof:

$$H(X) = \sum_{x \in \operatorname{supp}(X)} -\log(P(X = x))P(X = x) \tag{7.1}$$

$$= \sum_{x \in \operatorname{supp}(X)} -\log(\frac{1}{|\operatorname{supp}(X)|})P(X = x) \tag{7.2}$$

$$= \sum_{x \in \operatorname{supp}(X)} \log(n)P(X = x) = \log(n)\,. \tag{7.3}$$

□

Exercise 7.5 *You are trying to learn chess and you start by studying where chess grandmasters move their king when it is positioned in one of the middle fields of the board. The king can move to any of the adjoining 8 fields. Since you do not know a thing about chess yet, you assume that each move is equally probable. In this situation, what is the entropy of moving the king?*

One of the first important results in information theory is Shannon's source-coding theorem which states that the entropy $H(X)$ of a random variable X measures how many bits one will need on average to encode an outcome that is generated by the distribution P_X. This result applies to the real-world problem of data compression. Assume that N data points are drawn iid from the distribution P_X. In that case, the source-coding theorem tells us that on average, we will need $N \cdot H(X)$ bits to store the (optimally compressed) data. For example, let P_X be the $Bernoulli(\theta)$ distribution over bits. In the case $\theta = 1/2$, we have N perfectly random bits which cannot be compressed, and hence we need $N \cdot H(X) = N \cdot h(\theta) = N \cdot h(1/2) = N$ bits of storage. For the general case $\theta \neq 1/2$ when the individual bits are biased, the graph of the binary entropy $h(\theta)$ in Figure 7.1 tells us exactly what the compression ratio will be. We will not cover the proof of the source-coding theorem here, but refer to the literature instead.

7.2 Conditional Entropy

At the outset of this section we promised you that you could easily transfer results from probability theory to information theory. We will not be able to show any kind of linearity for entropy because it contains log-terms and the logarithm is not linear. We can however find alternative expressions for joint entropy (where the joint entropy is simply the entropy of a joint RV). Before we do so, let us also define the notion of conditional entropy. We have seen in Section **??** that $P_{X|Y=y}$ is a valid probability distribution for any $y \in \text{supp}(Y)$ such that $P(Y = y) > 0$. Hence, we can also define its conditional entropy.

Definition 7.6 (Conditional Entropy) *For two jointly distributed RVs X, Y and $y \in \text{supp}(Y)$ such that $P(Y = y) > 0$, the conditional entropy of X given that $Y = y$ is defined as*

$$\begin{aligned} H(X|Y=y) &:= \mathbb{E}_X[-\log_2(P(X=x|Y=y))] \\ &= -\sum_{x \in \text{supp}(X)} P(X=x|Y=y) \log_2(P(X=x|Y=y))\,. \end{aligned}$$

The conditional entropy of X given Y is defined as

$$H(X|Y) := \mathbb{E}_Y[H(X|Y)] = \sum_{y \in \text{supp}(Y)} P(Y=y) H(X|Y=y)\,.$$

Intuitively, $H(X|Y)$ is the (average) uncertainty of X after learning Y. Intuitively, learning Y (and in fact any information) cannot increase your uncertainty about X. Formally, one can prove the following

Lemma 7.7 (see e.g. Proposition 4 of this script) *For any two random variables X, Y with joint distribution P_{XY}, it holds that $H(X|Y) \leq H(X)$.*

Note however, that this non-increase of uncertainty only holds on average, as illustrated by the following example:

Example Consider the binary random variables X and Y, with joint distribution

$$P(X=0,Y=0) = \frac{1}{2}, \quad P(X=0,Y=1) = \frac{1}{4}$$
$$P(X=1,Y=0) = 0, \quad P(X=1,Y=1) = \frac{1}{4}.$$

By marginalization, we find that $P(X=0) = \frac{3}{4}$ and $P(X=1) = \frac{1}{4}$, while $P(Y=0) = P(Y=1) = \frac{1}{2}$. This allows us to make the following computations:

$$\begin{aligned}
H(X,Y) &= \frac{1}{2}\log 2 + \frac{1}{4}\log 4 + \frac{1}{4}\log 4 = \frac{3}{2} \\
H(X) &= h\left(\frac{1}{4}\right) = h\left(\frac{3}{4}\right) \approx 0.81 \\
H(Y) &= h\left(\frac{1}{2}\right) = 1 \\
H(X|Y) &= P(Y=0)\cdot H(X|Y=0) + P(Y=1)\cdot H(X|Y=1) \\
&= \frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 1 = \frac{1}{2} \\
H(Y|X) &= P(X=0)\cdot H(Y|X=0) + P(X=1)\cdot H(Y|X=1) \\
&= \frac{3}{4}\cdot h\left(\frac{1}{3}\right) + \frac{1}{4}\cdot 0 \approx 0.69
\end{aligned}$$

Note that for this specific distribution, learning the outcome $Y=1$ increases the uncertainty about X, $H(X|Y=1) > H(X)$, but on average, we always have $H(X|Y) \leq H(X)$. It is important to remember that Lemma 7.7 only holds on average, not for specific values of Y. Note also that in this example, $H(X|Y) \neq H(Y|X)$.

It is not a coincidence that the joint entropy $H(X,Y)$ in the example above is equal to $H(X|Y) + H(Y)$ and $H(Y|X) + H(X)$. One can prove this chain rule in general:

$$\begin{aligned}
H(X,Y) &= \sum_{\substack{x\in\text{supp}(X)\\ y\in\text{supp}(Y)}} -\log(P(X=x,Y=y)) \times P(X=x,Y=y) \\
&= \sum_{\substack{x\in\text{supp}(X)\\ y\in\text{supp}(Y)}} -\log(P(X=x \mid Y=y)) \times P(X=x,Y=y) \\
&\quad - \sum_{y\in\text{supp}(Y)} \log(P(Y=y)) \times \sum_{x\in\text{supp}(X)} P(X=x,Y=y) \\
&= \sum_{y\in\text{supp}(Y)} P(Y=y) \times \sum_{x\in\text{supp}(X)} -\log(P(X=x \mid Y=y)) \times P(X=x \mid Y=y) \\
&\quad - \sum_{y\in\text{supp}(Y)} \log(P(Y=y)) \times P(Y=y) \\
&= H(X|Y) + H(Y)\ .
\end{aligned}$$

Exercise 7.8 *Prove that $H(X,Y|Z) = H(X|Z)+H(Y|Z)$ if $X \perp Y \mid Z$.*

As corollary, we get that $H(X,Y) = H(X)+H(Y)$ for independent random variables X and Y. More generally, the entropy of n independent random variables is $H(X_1^n) = \sum_{i=1}^n H(X_i)$.

7.3 An Information-Theoretic View on EM

Now that we have seen some information-theoretic concepts, you may be happy to hear that there is an information-theoretic interpretation of EM. This interpretation helps us to get a better intuition for the algorithm. To formulate that interpretation we need one more concept, however.

Definition 7.9 (Relative Entropy) *The relative entropy of RVs X, Y with distributions P_X, P_Y and $\text{supp}(X) \subseteq \text{supp}(Y)$ is defined as*

$$D(P_X||P_Y) := \sum_{x\in\text{supp}(X)} P(X=x) \log \frac{P(X=x)}{P(Y=x)}\ .$$

If $P(Y=x) = 0$ for any $x \in \text{supp}(X)$ we define $D(P_X||P_Y) = \infty$. As with entropy, we often abbreviate $D(P_X||P_Y)$ with $D(X||Y)$.

The relative entropy is commonly known as **Kullback-Leibler (KL)** divergence. It measures the entropy of X as scaled to Y. Intuitively, it gives

a measure of how "far away" P_X is from P_Y. To understand "far away", recall that entropy is a measure of uncertainty. This uncertainty is low if both distributions place most of their mass on the same outcomes. Since $\log(1) = 0$ the relative entropy is 0 if $P_X = P_Y$.

It is worthwhile to point out the difference between relative and conditional entropy. Conditional entropy is the average entropy of X given that you know what value Y takes on. In the case of relative entropy you do not know the value of Y, only its distribution.

Exercise 7.10 *Show that* $D(X, Y||Y) = H(X|Y)$. *Furthermore show that* $D(X, Y||Y) = H(X)$ *if* $X \perp Y$.

Let us start by remembering why we need EM. We have a model that defines a joint distribution over observed (x) and latent data (z). Such a model generally looks as follows:

(7.4)
$$P(X = x, Z = z \mid \Theta = \theta) = P(X = x \mid Z = z, \Theta = \theta)P(Z = z \mid \Theta = \theta)$$

where we have chosen a factorization that provides a separate term for a distribution over only the latent data.

Recall that the goal of the EM algorithm is to iteratively increase the likelihood through consecutive updates of parameter estimates. These updates are achieved through maximum-likelihood estimation based on expected sufficient statistics. We are now going to show that a) EM computes a lower bound on the marginal log-likelihood of the data in each iteration and b) that this lower bound becomes tight when the expected sufficient statistics are taken with respect to the model posterior. The latter implies that EM performs the optimal update in each iteration.

Let us start by expanding the data log-likelihood and then lower-bounding it.

$$\log(P(X = x \mid \Theta = \theta)) = \log(\sum_y P(X = x, Y = y \mid \Theta = \theta)) \tag{7.5}$$

$$= \log\left(\sum_y Q(Y = y \mid \Phi = \phi)\frac{P(X = x, Y = y \mid \Theta = \theta)}{Q(Y = y \mid \Phi = \phi)}\right) \tag{7.6}$$

$$\geq \sum_y Q(Y = y \mid \Phi = \phi) \log\left(\frac{P(X = x, Y = y \mid \Theta = \theta)}{Q(Y = y \mid \Phi = \phi)}\right) \tag{7.7}$$

Here, we have used Jensen's Inequality to derive the lower bound. Observe that the log is indeed a concave function.

We also have introduced an auxiliary distribution Q over the latent variables with parameters ϕ. For reasons that we will explain shortly, this distributions is often called the **variational distribution** and its parameters the **variational parameters**. The letter Q is slightly non-standard to

denote distributions but we are are following conventions from the field of **variational inference** here.

In the next step, we factorise the model distribution in order to recover a KL divergence term between the variational distribution and the model posterior over latent variables.

(7.8)
$$\sum_y Q(Y = y \mid \Phi = \phi) \log \left(\frac{P(X = x, Y = y \mid \Theta = \theta)}{Q(Y = y \mid \Phi = \phi)} \right)$$

(7.9)
$$= \sum_y Q(Y = y \mid \Phi = \phi) \log \left(\frac{P(Y = y \mid X = x, \Theta = \theta) P(X = x \mid \Theta = \theta)}{Q(Y = y \mid \Phi = \phi)} \right)$$

(7.10)
$$= \sum_y Q(Y = y \mid \Phi = \phi) \log \left(\frac{P(Y = y \mid X = x, \Theta = \theta)}{Q(Y = y \mid \Phi = \phi)} \right) + \log(P(X = x \mid \Theta = \theta))$$

(7.11)
$$= -D(Q||P) + \log(P(X = x \mid \Theta = \theta))$$

Equation (7.11) gives us two insights. First it quantifies the gap between the lower bound and the actual data likelihood. This gap is equal to the KL divergence between the variational distribution and the model posterior over latent variables. Second, since KL divergence is always positive, the bound only becomes tight when $P = Q$. But this is exactly what is happening in the E-step! The E-step sets $P = Q$ and then computes expectations under that distribution (see Equation (7.7)). Thus, the E-step increases the lower bound on the marginal log-likelihood.

Looking back at Equation (7.7), we also see that the M-step increases the lower bound because it maximises $\mathbb{E}\left[P(X = x, Y = y \mid \Theta = \theta)\right]$. We conclude that both steps are increasing the lower bound on the log-likelihood. We therefore conclude that EM increases the data likelihood in every iteration (or leaves it unchanged at worst).

We will finish with a quick rejoinder on variational inference. EM is a special case of variational inference. Variational inference is any inference procedure which uses an auxiliary distribution Q to compute a lower bound on the likelihood. In the general setting, the auxiliary distribution can be different from the model posterior. This means that the bound never gets tight. However, in models in which the exact posterior is hard (read: impossible) to compute, using a non-tight lower bound instead can be incredibly useful!

The reason this inference procedure is called *variational* is because it is based on the calculus of variations. This works mostly like normal calculus except that standard operations like differentiation are done with respect to functions instead of variables.

Further Material

At the ILLC, there is a whole course about information theory, currently taught by Christian Schaffner. David MacKay also offers a free book on the subject. Finally, Coursera also offers an online course on information theory.

The information-theoretic formulation of EM was pioneered in this paper. A very recent and intelligible tutorial on variational inference can be found on the archive.