Chapter 8

Information Theory

The subject of this chapter is a foray into **information theory**, which concerns itself with questions such as: "What is the shortest average description for a random variable?" (the **data compression problem**); and "What is the maximum rate at which we can send information through a noisy channel?" (the **data transmission problem**). Although we will not explore these questions in any depth, we will introduce the fundamental notions used in information theory to capture the illusive meaning of "information".

8.1 Entropy

Let X be a discrete random variable taking values in $\mathcal{X} \subseteq \mathbb{R}$. Throughout this chapter, we will use the abbreviated notation $\mathsf{p}_X(x) := \mathbb{P}(X = x)$.

Definition 8.1. The **entropy** of X, $H(X)$, is defined to be:

$$\boxed{H(X) := -\sum_{x \in \mathcal{X}} \mathsf{p}_X(x) \log_2 \mathsf{p}_X(x) = \mathbb{E}\left[\log_2 \frac{1}{\mathsf{p}_X(X)}\right].}$$

Observe that the entropy of X really only depends on the *distribution* of X, p_X, so we will also write $H(\mathsf{p}_X) := H(X)$.

We use the base-2 logarithm because we think of entropy as being measured in *bits*; indeed, we will give an interpretation of entropy as a measure of the *information* contained in the random variable. For now, though, suppose that we chose to measure entropy using a logarithm with a different base $b > 1$, that is, we define $H_b(\mathsf{p}_X) := -\sum_{x \in \mathcal{X}} \mathsf{p}_X(x) \log_b \mathsf{p}_X(x)$. Using the logarithm change-of-base rule,

$$\log_b x = \frac{\ln x}{\ln b} \qquad \text{and} \qquad \log_2 x = \frac{\ln x}{\ln 2},$$

so we find that

$$\log_b x = \frac{\ln 2}{\ln b} \log_2 x = (\log_b 2)(\log_2 x).$$

Therefore, we conclude that $H_b(\mathsf{p}_X) = (\log_b 2)H(\mathsf{p}_X)$. The moral of the story is that *using a different base for the logarithm just introduces a different constant factor in front of the entropy*, so regardless of which base we choose for the logarithm, the theory will remain essentially the same. From now on, we will stick to using base 2. [1]

Further, observe that since $\mathsf{p}_X \leq 1$, $-\log_2 \mathsf{p}_X \geq 0$, and so the entropy is always non-negative:

$$\boxed{H(\mathsf{p}_X) \geq 0}$$

(We consider this to be a desirable property of entropy, because we do not really know how to make sense of "negative information".)

Now, we will provide an interpretation of entropy. [2] Suppose that you are interested in measuring the value of a random variable X. You perform an experiment and observe the event $\{X = x\}$, where $x \in \mathcal{X}$. We define the **surprise** of this event to be $-\log_2 \mathsf{p}_X(x)$ and we regard it as the amount of information you gained from the observation. For example, if $\mathsf{p}_X(x) = 1$, then your surprise is 0; but this makes sense because $\mathsf{p}_X(x) = 1$ means the event $\{X = x\}$ was certain from the beginning! On the other hand, if $\mathsf{p}_X(x) = 0$, then your surprise is ∞, which again makes sense because $\mathsf{p}_X(x) = 0$ means that $\{X = x\}$ should have been impossible... uh-oh!

In this language, we see that the entropy is your *expected surprise*. Entropy is a strangely self-referential concept, where you make an observation, and then think about the probability that you would make the observation which you just made.

Example 8.2. Fix $n \in \mathbb{Z}_+$. Let $X \sim \text{Uniform}\{1, \ldots, n\}$, so $\mathsf{p}_X(x) = 1/n$ for each $x \in \{1, \ldots, n\}$.

$$H(X) = -\sum_{i=1}^{n} \frac{1}{n} \log_2 \frac{1}{n} = -\log_2 \frac{1}{n} = \log_2 n.$$

In fact, the uniform distribution is the distribution with the *largest* possible entropy over n symbols. (By "n symbols", we refer to the fact that $\{1, \ldots, n\}$ has n elements. Notice that the entropy of X does not depend on *what* possible values it can take, as the entropy only depends on the *probabilities* with which it takes on these values. So, saying that X has the largest entropy out of any random variable which takes values in $\{1, \ldots, n\}$ is equivalent to saying that X has the largest entropy out of any random variable which takes on at most n values, regardless of what those n values are.) Intuitively, this is true because the uniform distribution is "the most random"; in other words, before you measure X, you have the least

[1] Perhaps this could be called a "computer science mindset".

[2] In reality, it is possible that you will not be satisfied by any of the interpretations we give. After all, why should we believe that a concept as illusive as "information" can be captured by mathematics? Personally, I think the *true* justification for the concept of entropy is provided by looking at the various situations in which it arises and the various results we can prove about entropy which align with our intuition.

information about X when X is uniformly distributed on the set $\{1, \ldots, n\}$.

Proving the above assertion (that if X takes on at most n values, then $H(X) \leq \log_2 n$) is tricky, so we will wait until we have more tools.

Example 8.3. Let $X \sim \text{Bernoulli}(p)$. Then,

$$H(X) = -p \log_2 p - (1-p) \log_2 (1-p).$$

This is known as the **binary entropy function** and it is commonly denoted simply as $H(p)$. See Figure 8.1.

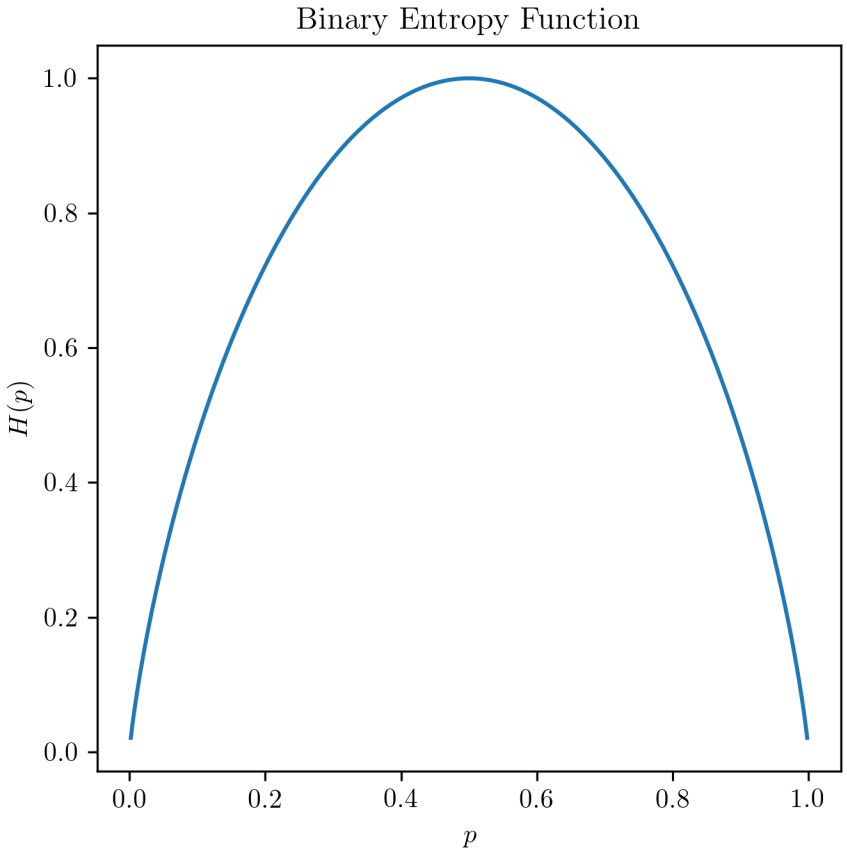

Figure 8.1: Plot of the binary entropy function. Observe that it is non-negative and concave.

8.2 Relative Entropy

Look back at the interpretation of surprise mentioned as an interpretation for entropy. Now, consider the situation where you have an incorrect belief about the probability distribution of the random variable X, that is, X has the probability distribution p but you mistakenly believe that X has the probability distribution q. Upon seeing the event $\{X = x\}$, your surprise is now $-\log_2 \mathsf{q}(x)$, and your expected surprise overall is $-\sum_{x\in\mathcal{X}} \mathsf{p}(x) \log_2 \mathsf{q}(x)$.

Since your expected surprise would have been $H(\mathsf{p}) = -\sum_{x\in\mathcal{X}} \mathsf{p}(x) \log_2 \mathsf{p}(x)$ (if you had correctly known the true distribution p), we see that your *additional* expected surprise from your incorrect belief is

$$-\sum_{x\in\mathcal{X}} \mathsf{p}(x) \log_2 \mathsf{q}(x) - \left(-\sum_{x\in\mathcal{X}} \mathsf{p}(x) \log_2 \mathsf{p}(x)\right) = \sum_{x\in\mathcal{X}} \mathsf{p}(x) \log_2 \frac{\mathsf{p}(x)}{\mathsf{q}(x)}.$$

We now formulate a definition based on these ideas.

Definition 8.4. The **relative entropy of q from p** is:

$$\boxed{D_{\mathrm{KL}}(\mathsf{p} \,\|\, \mathsf{q}) := \sum_{x\in\mathcal{X}} \mathsf{p}(x) \log_2 \frac{\mathsf{p}(x)}{\mathsf{q}(x)} = \mathbb{E}_{\mathsf{p}}\left[\log_2 \frac{\mathsf{p}(X)}{\mathsf{q}(X)}\right]}$$

This is also commonly called the **Kullback-Leibler divergence of q from p**, which explains the subscript in the notation.

Observe that if there is any $x \in \mathcal{X}$ such that $\mathsf{p}(x) > 0$ but $\mathsf{q}(x) = 0$, then the term in the summation for x yields $\mathsf{p}(x) \log_2(\mathsf{p}(x)/0)$ which we interpret as ∞.

Often, the relative entropy is used as a measure of "distance" between two probability distributions. It is not a true distance function because it is not symmetric: in general, $D_{\mathrm{KL}}(\mathsf{p} \,\|\, \mathsf{q}) \neq D_{\mathrm{KL}}(\mathsf{q} \,\|\, \mathsf{p})$. [3]

Exercise 22 Consider the following two distributions on $\{0, 1\}$:

$$\begin{aligned} \mathsf{p}(0) &= 1 - p \\ \mathsf{p}(1) &= p \end{aligned} \qquad \text{and} \qquad \begin{aligned} \mathsf{q}(0) &= 1 - q \\ \mathsf{q}(1) &= q \end{aligned}$$

Calculate $D_{\mathrm{KL}}(\mathsf{p} \,\|\, \mathsf{q})$ and $D_{\mathrm{KL}}(\mathsf{q} \,\|\, \mathsf{p})$. [Note: Here, p corresponds to the Bernoulli(p) distribution and q corresponds to the Bernoulli(q) distribution; in this special case, the relative entropy of q from p is often denoted more simply as $D_{\mathrm{KL}}(p \,\|\, q)$.] Show that for $p = 1/2$ and $q = 1/4$, they are not equal.

However, we can at least show that it is always non-negative:

[3] Moreover, it does not necessarily satisfy the triangle inequality.

Theorem 8.5 (Relative Entropy Inequality)**.** *For all probability distributions* p *and* q *on the countable set* $\mathcal{X}$*,* $D_{\mathrm{KL}}(\mathsf{p} \,\|\, \mathsf{q}) \geq 0$*, with equality if and only if* $\mathsf{p} = \mathsf{q}$*.*

Proof. We will use the inequality $\ln x \leq x - 1$ for $x > 0$, with equality if and only if $x = 1$. [a] So,

$$\begin{aligned} D_{\mathrm{KL}}(\mathsf{p} \,\|\, \mathsf{q}) &= \frac{1}{\ln 2} \sum_{x \in \mathcal{X}} \mathsf{p}(x) \ln \frac{\mathsf{p}(x)}{\mathsf{q}(x)} = -\frac{1}{\ln 2} \sum_{x \in \mathcal{X}} \mathsf{p}(x) \ln \frac{\mathsf{q}(x)}{\mathsf{p}(x)} \\ &\geq -\frac{1}{\ln 2} \sum_{x \in \mathcal{X}} \mathsf{p}(x) \left(\frac{\mathsf{q}(x)}{\mathsf{p}(x)} - 1 \right) = -\frac{1}{\ln 2} \left(\sum_{x \in \mathcal{X}} \mathsf{q}(x) - \sum_{x \in \mathcal{X}} \mathsf{p}(x) \right) = 0, \end{aligned}$$

with equality if and only if $\mathsf{q}(x)/\mathsf{p}(x) = 1$ for all $x \in \mathcal{X}$, that is, $\mathsf{p} = \mathsf{q}$. □

[a] Indeed, $f(x) = x - 1 - \ln x$ has $f''(x) = x^{-2} > 0$ for all $x > 0$, so f is strictly convex and it has a unique minimum. Since $f'(x) = 1 - x^{-1} = 0$ when $x = 1$, it follows that f attains its minimum value of 0 at $x = 1$, which proves that $x - 1 \geq \ln x$ for all $x > 0$, with equality if and only if $x = 1$.

Corollary 8.6 (Maximum Entropy Distribution)**.** *Let* X *be a random variable taking on values in* $\mathcal{X}$*, with* $|\mathcal{X}| = n$*. Then,* $H(X) \leq \log_2 n$*.*

Proof. Let $\mathsf{p}(x) := \mathbb{P}(X = x)$ and q be the uniform distribution on $\mathcal{X}$. Then,

$$D_{\mathrm{KL}}(\mathsf{p} \,\|\, \mathsf{q}) = \sum_{x \in \mathcal{X}} \mathsf{p}(x) \log_2 \frac{\mathsf{p}(x)}{1/n} = \sum_{x \in \mathcal{X}} \mathsf{p}(x) \log_2 n + \sum_{x \in \mathcal{X}} \mathsf{p}(x) \log_2 \mathsf{p}(x) = \log_2 n - H(X),$$

but from Theorem 8.5, $D_{\mathrm{KL}}(\mathsf{p} \,\|\, \mathsf{q}) \geq 0$, and so we have $H(X) \leq \log_2 n$. □

Corollary 8.6 verifies our earlier claim that the uniform distribution on n symbols (which has $H(X) = \log_2 n$) has the maximum entropy out of all distributions on n symbols.

8.3 Chernoff Bounds

Take a moment to recall (3.32), which is reproduced here for convenience: For $\theta > 0$,

$$\mathbb{P}(X \geq x) \leq \frac{\mathbb{E}[\exp(\theta X)]}{\exp(\theta x)}. \tag{8.1}$$

We will focus on the simple case of coin flips and demonstrate a connection between the relative entropy and the Chernoff bound.

Theorem 8.7 (Chernoff Bound for Coin Flips)**.** *Let* $n \in \mathbb{Z}_+$ *and* $S_n := X_1 + \cdots + X_n$*,*

where the X_i are i.i.d. Bernoulli(p) *random variables. Then, for $\varepsilon > 0$,*

$$\mathbb{P}\left(\frac{S_n}{n} \geq p + \varepsilon\right) \leq \exp(-nD_{\mathrm{KL}}(p + \varepsilon \,\|\, p)),$$

where the relative entropy is calculated using the natural logarithm.

Proof. We will apply (3.32) to the random variable S_n/n:

$$\begin{aligned}\mathbb{P}\left(\frac{S_n}{n} \geq p + \varepsilon\right) &= \mathbb{P}(S_n \geq (p+\varepsilon)n) \leq \frac{\mathbb{E}[\exp(\theta(X_1 + \cdots + X_n))]}{\exp(\theta(p+\varepsilon)n)} \\ &= \frac{\mathbb{E}[\exp(\theta X_1) \cdots \exp(\theta X_n)]}{\exp(\theta(p+\varepsilon)n)} = \frac{\mathbb{E}[\exp(\theta X_1)] \cdots \mathbb{E}[\exp(\theta X_n)]}{\exp(\theta(p+\varepsilon)n)} \\ &= \frac{M_X(\theta)^n}{\exp(\theta(p+\varepsilon)n)} = \exp(n(\ln M_X(\theta) - \theta(p+\varepsilon))).\end{aligned}$$

where we have used the i.i.d. assumption, and we used the definition

$$M_X(\theta) := \mathbb{E}[\exp(\theta X_1)] = 1 - p + p\exp\theta.$$

Now, we seek the best possible bound over all $\theta > 0$, so we differentiate the following quantity with respect to θ:

$$\frac{\mathrm{d}}{\mathrm{d}\theta}(\ln(1 - p + p\exp\theta) - \theta(p+\varepsilon)) = \frac{p\exp\theta}{1 - p + p\exp\theta} - (p+\varepsilon),$$

and by setting the above quantity to 0, we have:

$$\begin{aligned}\frac{p\exp\theta}{1-p+p\exp\theta} = p + \varepsilon &\implies \frac{1-p+p\exp\theta}{p\exp\theta} = \frac{1}{p+\varepsilon} \implies \frac{1-p}{p\exp\theta} + 1 = \frac{1}{p+\varepsilon} \\ &\implies \frac{1-p}{p\exp\theta} = \frac{1-p-\varepsilon}{p+\varepsilon} \implies \exp\theta = \frac{p+\varepsilon}{p} \cdot \frac{1-p}{1-p-\varepsilon} \\ &\implies \theta = \ln\frac{p+\varepsilon}{p} + \ln\frac{1-p}{1-p-\varepsilon}.\end{aligned}$$

Now, we plug in the above result into $\ln M_X(\theta) - \theta(p+\varepsilon)$:

$$\begin{aligned}&\ln M_X(\theta) + \theta(p+\varepsilon) \\ &\quad= \ln\left(1 - p + (1-p)\frac{p+\varepsilon}{1-p-\varepsilon}\right) - (p+\varepsilon)\ln\frac{p+\varepsilon}{p} - (p+\varepsilon)\ln\frac{1-p}{1-p-\varepsilon} \\ &\quad= \ln\frac{1-p}{1-p-\varepsilon} - (p+\varepsilon)\ln\frac{p+\varepsilon}{p} - (p+\varepsilon)\ln\frac{1-p}{1-p-\varepsilon}\end{aligned}$$

$$= (1-p-\varepsilon)\ln\frac{1-p}{1-p-\varepsilon} + (p+\varepsilon)\ln\frac{p}{p+\varepsilon} = -D_{\mathrm{KL}}(p+\varepsilon \,\|\, p).$$

We have the remarkable result that $\mathbb{P}(S_n/n \geq p+\varepsilon) \leq \exp(-nD_{\mathrm{KL}}(p+\varepsilon \,\|\, p))$. □

We could go on to upper bound $D_{\mathrm{KL}}(p+\varepsilon \,\|\, p)$, but at this stage we are mainly interested in knowing that there is a bound with decays exponentially with the number of random variables. Therefore, it suffices to observe that for $\varepsilon > 0$, $D_{\mathrm{KL}}(p+\varepsilon \,\|\, p) > 0$ due to Theorem 8.5, and so (8.1) is indeed exponentially decaying.

8.4 Solutions to Exercises

Exercise 22 From the definition, we have

$$D_{\mathrm{KL}}(\mathsf{p} \,\|\, \mathsf{q}) = (1-p)\log_2\frac{1-p}{1-q} + p\log_2\frac{p}{q},$$
$$D_{\mathrm{KL}}(\mathsf{q} \,\|\, \mathsf{p}) = (1-q)\log_2\frac{1-q}{1-p} + q\log_2\frac{q}{p}.$$

Plugging in $p = 1/2$ and $q = 1/4$, we find (numerically) that $D_{\mathrm{KL}}(\mathsf{p} \,\|\, \mathsf{q}) \approx 0.208$ and $D_{\mathrm{KL}}(\mathsf{q} \,\|\, \mathsf{p}) \approx 0.189$, which gives an example where the relative entropy is not symmetric.