

## Lecture 1: Entropy and Data Compression

The fundamental concepts of information theory can be motivated by the problem of data compression. Suppose that we have a countable set  $M$  of messages. Suppose that we want to transmit a sequence of  $b$  messages  $m_1, m_2, \dots, m_b$  where the messages  $m_i$  are drawn IID according to  $P$ . The main theorem we consider in this section is the following (stated somewhat informally).

**Shannon's Source Coding Theorem** [Informal Version] In the limit as the block size goes to infinity the number of bits required per message in the block is exactly the entropy  $H(P)$  of  $P$  defined as follows.

$$H(P) = E_{m \sim P} [\log_2(1/P(m))]$$

As a simple example suppose that  $P$  is the uniform distribution on  $2^k$  messages. In this case we have that  $H(P) = k$ . As another example suppose that  $P$  assigns nonzero weight to  $2^k$  messages but half of the weight is on a single message. In that case we can use the bit string 0 to represent the common message and codes of length  $k + 1$ , each starting with the the bit 1, for all other messages. In this case the average number of bits per message is  $1/2 + 1/2(k + 1) = 1 + k/2$ . In general if the distribution is non-uniform we get greater compression by assigning fewer bits to more common messages.

We now state the main theorem more precisely. For a countable set  $M$  of messages define a code for  $M$  to be an assignment  $c(m)$  of a bit string (code word) for each  $m \in M$  with the property that if  $m \neq m'$  then  $c(m) \neq c(m')$ . We will also consider coding a long sequence of messages by coding each message individually. In order for this to work one must be able to tell when one code word ends and the next begins. This can be done provided that no two distinct codes have the property that one is a prefix of the other.

**Definition 1** *A code is called prefix free if all code words are distinct and no code word is a prefix of any other code word.*

As an example of a prefix-free code we can consider the set of all null-terminated byte (or character) strings. In this case every code word is a certain byte string and hence the length of each code is a multiple of 8.

We can now state our main theorem more precisely. We now let  $|c(m)|$  be the length of the code word  $c(m)$  (the number of bits used in  $c(m)$ ). We let  $\mathcal{C}(M)$  be the set of all prefix-free codes on  $M$ . We let  $C^*(P)$  be defined as follows.

$$C^*(P) = \inf_{c \in \mathcal{C}(M)} \mathbb{E}_{m \sim P} [|c(m)|]$$

Let  $b$  be a positive integer block size. Let  $M^b$  be the set of all tuples  $\langle m_1, \dots, m_b \rangle$  with  $m_i \in M$ . Let  $P^b$  be the probability distribution on  $M^b$  where each  $m_i$  is selected independently with probability distribution  $P$ . (We say that the  $m_i$  are drawn “IID” for Independently Identically Distributed). The above definition implies the following.

$$C^*(P^b) = \inf_{c \in \mathcal{C}(M^b)} \mathbb{E}_{\langle m_1, \dots, m_b \rangle \sim P^b} [|c(\langle m_1, \dots, m_b \rangle)|]$$

Note that a code for  $M^b$  is coding for  $b$  messages from  $M$  so to get the number of bits per message we should divide the length of a code word by  $b$ . We can now state the main theorem as follows.

**Theorem 1 (Shannon’s Source Coding Theorem)**

$$\lim_{b \rightarrow \infty} \frac{1}{b} C^*(P^b) = H(P)$$

To prove this theorem we start with the following.

**Lemma 2 (Kraft Inequality)** *For any prefix-free code  $c$  we have the following.*

$$\sum_{m \in M} 2^{-|c(m)|} \leq 1$$

**Proof:** Suppose we generate a bit string by repeatedly flipping an unbiased coin and stopping as soon as the bit string we have generated is a code word. We then have that the probability of stopping with code word  $c(m)$  is exactly  $2^{-|c(m)|}$ . The Kraft inequality then follows from the fact that probabilities sum to 1 (and there can be some nonzero probability that we miss all the code words and never stop the process). ■

**Theorem 3** Let  $\ell$  be an assignment of a positive integer  $\ell(m)$  to each  $m \in M$  satisfying the Kraft inequality:

$$\sum_{m \in M} 2^{-\ell(m)} \leq 1$$

For any such assignment of lengths there exists a prefix-free code  $c$  with  $|c(m)| = \ell(m)$ .

**Proof:** Arrange the messages  $m$  in a sequence  $m_1, m_2, m_3, \dots$  of nondecreasing length according to the assignment  $\ell$ , i.e., such that  $\ell(m_{i+1}) \geq \ell(m_i)$ . Pick code words in the order given subject to the constraint the selected code word can not have as a prefix any previously selected code word. We view a code word  $c$  as having probability mass  $2^{-|c|}$ . When a code word  $c(m)$  is selected the amount of probability mass that becomes unavailable for future assignments is  $2^{-\ell(m)}$ . By the Kraft inequality the amount of probability mass remaining must be sufficient for the remainder of the code words. Furthermore, when selecting a code word for  $m$  we can consider the remaining mass to be uniformly distributed among the remaining code words of length  $\ell(m)$  and hence such a code word can always be selected. ■

**Theorem 4** For probability distribution  $P$  on a countable set of messages  $M$  there exists a code  $c$  assigning code word  $c(m)$  to each  $m \in M$  satisfying the following.

$$E_{m \sim P} [|c(m)|] \leq H(P) + 1$$

**Proof:** Let  $\ell(m)$  be  $\lceil \log_2 1/P(m) \rceil$ . We have the following.

$$\begin{aligned} \sum_{m \in M} 2^{-\ell(m)} &= \sum_{m \in M} 2^{-\lceil \log_2 1/P(m) \rceil} \\ &\leq \sum_{m \in M} P(m) \\ &= 1 \end{aligned}$$

Therefore by theorem 3 there exists a code  $c$  with  $|c(m)| = \ell(m)$  and hence we have the following.

$$\begin{aligned} |c(m)| &\leq \log_2(1/P(m)) + 1 \\ E_{m \sim P} [|c(m)|] &\leq E_{m \sim P} [\log_2(1/P(m)) + 1] \\ &= E_{m \sim P} [\log_2(1/P(m))] + E_{m \sim P} [1] \\ &= H(P) + 1 \end{aligned}$$

■

Now consider the distribution  $P^b$  on the message blocks  $M^b$ . The expected number of bits per message in a message block using a Shannon code for message blocks is the following.

$$\begin{aligned} \frac{1}{b} \mathbb{E}_{\langle m_1, \dots, m_b \rangle \sim P^b} [|\mathcal{C}(\langle m_1, \dots, m_b \rangle)|] &\leq 1/b(H(P^b) + 1) \\ &= H(P) + 1/b \end{aligned}$$

This implies that for arbitrarily long blocks the number of bits per message can be made arbitrarily close to  $H(P)$ . In the next lecture we will prove that no code can achieve fewer bits per code word than  $H(P)$ .

## 1 Cross Entropy and KL Divergence

We will now prove the other half of the Shannon source coding theorem by applying Jensen's inequality to KL divergence. We will have to consider functions  $g(x)$  which can be infinity. When  $g$  can be infinite we define the expectation of  $g$  as follows.

$$\mathbb{E}_{m \sim P} [g(m)] = \sum_{m \in M, P(m) > 0} P(m)g(m)$$

Under this definition, if  $g(m)$  is infinite only when  $P(m)$  is zero then we have that the expectation ignores the infinite values of  $g$  and is still finite (for finite  $M$ ).

Now we assume, without loss of generality, that  $M$  contains a special element  $\perp$  and  $P(\perp) = 0$  (we can always add one if no such element exists). Without loss of generality we can also restrict our attention to prefix-free codes  $c$  that assign code words to elements of  $M$  other than  $\perp$ . For any such code  $c$  we can define  $P_c(m)$  as follows.

$$P_c(m) = 2^{-|c(m)|} \text{ if } m \neq \perp$$

$$P_c(\perp) = 1 - \sum_{m \neq \perp} P_c(m)$$

The Kraft inequality implies that  $P_c(\perp)$  as defined above is non-negative. By introducing  $\perp$  we have arranged that  $P_c$  is a probability distribution on  $M$  (it sums to one). Although we are particularly interested in distributions of the form  $P_c$  for some code  $c$ , we now consider an arbitrary distribution  $Q$  on  $M$  and define the cross entropy  $H(P, Q)$  as follows.

$$H(P, Q) = E_{m \sim P} \left[ \log_2 \frac{1}{Q(m)} \right]$$

Note that messages with zero probability under  $P$  (such as  $\perp$ ) are ignored in this definition. If  $Q$  assigns zero probability to a message that has nonzero probability under  $P$  then  $H(P, Q)$  is infinite. We now have the following.

$$E_{m \sim P} [|c(m)|] = H(P, P_c)$$

In general  $H(P, Q)$  can be viewed as the expected code length when we select messages according to  $P$  but use a code that is optimal for  $Q$ . If  $Q$  is different from  $P$  then we are using the wrong code so we should expect that  $H(P, Q)$  is at least  $H(P)$ . We will now prove this as a consequence of Jensen's inequality. First we define the Kulback-Leibler divergence,  $KL(P, Q)$  as follows.

$$KL(P, Q) = H(P, Q) - H(P)$$

It now suffices to show that  $KL(P, Q) \geq 0$ . We can do this as follows where the first inequality is derived from Jensen's inequality and the fact that the function  $f(x) = -\log_2(x)$  is convex.

$$\begin{aligned} KL(P, Q) &= H(P, Q) - H(P) \\ &= E_{m \sim P} \left[ \log_2 \left( \frac{1}{Q(m)} \right) \right] - E_{m \sim P} \left[ \log_2 \left( \frac{1}{P(m)} \right) \right] \\ &= E_{m \sim P} \left[ \log_2 \left( \frac{1}{Q(m)} \right) - \log_2 \left( \frac{1}{P(m)} \right) \right] \end{aligned}$$

$$\begin{aligned}
&= \mathbb{E}_{m \sim P} \left[ \log_2 \left( \frac{P(m)}{Q(m)} \right) \right] \\
&= \mathbb{E}_{m \sim P} \left[ -\log_2 \left( \frac{Q(m)}{P(m)} \right) \right] \\
&\geq -\log_2 \left( \mathbb{E}_{m \sim P} \left[ \frac{Q(m)}{P(m)} \right] \right) \\
&= -\log_2 \left( \sum_{m \in M, P(m) > 0} P(m) \frac{Q(m)}{P(m)} \right) \\
&= -\log_2 \left( \sum_{m \in M, P(m) > 0} Q(m) \right) \\
&\geq 0
\end{aligned}$$

We have now proved that for any prefix-free code  $c$  we have that  $H(P, P_c) \geq H(P)$  which proves that any code must use at least  $H(P)$  bits per message.

Kulback-Leibler divergence is one of the central concepts of information theory. Intuitively,  $KL(P, Q)$  measures the degree to which  $P$  and  $Q$  are different. But in general we have that  $KL(P, Q)$  does not equal  $KL(Q, P)$ . Note that if  $U$  is the uniform distribution on a finite set of size  $k$  then  $H(P, U) = H(U) = \log k$  while  $H(U, P) = \infty$  whenever there exists a value  $\perp$  with  $P(\perp) = 0$ .

## 2 Mutual Information

For two random variables  $X$  and  $Y$  we define the conditional entropy  $H(Y|X)$  as follows.

$$\begin{aligned}
H(Y|X) &= \mathbb{E} \left[ \log_2 \frac{1}{P(y|x)} \right] \\
&= \sum_x P(x) \sum_y P(y|x) \log_2 \frac{1}{P(y|x)}
\end{aligned}$$

$$= \sum_x P(x)H(Y|x)$$

In the last line above  $H(Y|x)$  is a quantity that is different for different values of  $x$  while  $H(Y|X)$  is a fixed quantity independent of any particular value for  $X$  or  $Y$ .

**Definition 2** *The mutual information between  $X$  and  $Y$ , denoted  $I(X, Y)$  is defined as follows.*

$$I(X, Y) = H(X) + H(Y) - H(\langle X, Y \rangle)$$

If  $X$  and  $Y$  are correlated then  $H(\langle X, Y \rangle)$  will be smaller than  $H(X) + H(Y)$ .

**Lemma 5**

$$\begin{aligned} I(X, Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \end{aligned}$$

**Proof:**

$$\begin{aligned} I(X, Y) &= H(X) + H(Y) - H(\langle X, Y \rangle) \\ &= H(X) + \mathbb{E} \left[ \log_2 \frac{1}{P(y)} \right] - \mathbb{E} \left[ \log_2 \frac{1}{P(\langle x, y \rangle)} \right] \\ &= H(X) + \mathbb{E} \left[ \log_2 \frac{1}{P(y)} - \log_2 \frac{1}{P(\langle x, y \rangle)} \right] \\ &= H(X) + \mathbb{E} \left[ \log_2 \frac{P(\langle x, y \rangle)}{P(y)} \right] \\ &= H(X) + \mathbb{E} [\log_2 P(x|y)] \\ &= H(X) - \mathbb{E} \left[ \log_2 \frac{1}{P(x|y)} \right] \\ &= H(X) - H(X|Y) \end{aligned}$$

■

### 3 Problems

1. Let the set of messages  $M$  be the positive integers  $1, 2, 3, \dots$ . Suppose we pick an integer by flipping an unbiased coin and stopping as soon as we get the first heads. We then output the number of flips. This gives  $P(i) = (1/2)^i$ . Give a prefix-free code  $c(i)$  with  $|c(i)| = \log_2(1/P(i))$ . What is the entropy of this distribution?

2. Suppose that there is a popular 10 megapixel digital camera where a pixel is represented by 16 bits. Consider the probability distribution over images to be taken by this model of camera over the next year (we call these “natural” images). Give an upper bound on the entropy of this distribution. Ignore the question of whether the distribution of natural images is really well defined.

3. Suppose that we use rendering software to construct images of solid models where a model consists of a set of objects each at a certain configuration and under certain lighting conditions and from a certain camera position. Suppose that the models are generated by kids using modeling software on the web where each model must fit in a single ten kilobyte message. Consider the probability distribution over the images rendered in this process. Give an upper bound on the entropy of this distribution. (Again ignore the question of whether the distribution is well defined).

4. Consider climate simulation software that samples future weather patterns by using a random number generator to add noise to the process of weather formation. If the program always starts with the same current state but uses a 32 bit random number seed (selected uniformly from all such seeds) what is the entropy of the probability distribution over future weather patterns produced by this program assuming that each different random seed produces a different weather pattern.

5. Use the fact that  $K(P, Q) \geq 0$  to prove that the expected length of any prefix code is greater than  $H(X)$ , i.e.:

$$\sum_x p(x)l(x) \geq H(p)$$

Hint: Use the Kraft inequality and set  $q(x) = 2^{-l(x)}$ . Make sure you add a “dummy” element so that  $q(x)$  is a valid distribution.

1. Show that  $H(\langle X, Y \rangle) = H(X) + H(Y|X)$ .
2. Show that  $I(X, Y) = KL(P_{\langle X, Y \rangle}, P_X P_Y)$  where  $P_{\langle X, Y \rangle}$  is the distribution on pairs  $\langle x, y \rangle$  given by world states and  $P_X P_Y$  is the distribution on pairs  $\langle x, y \rangle$  given by  $P(\langle x, y \rangle) = P_X(x)P_Y(y)$ .
3. Explain why 2. implies that  $I(X, Y) \geq 0$
4. Explain why problem 3. implies that  $H(Y|X) \leq H(Y)$ .