

Note: in this course, \log denotes \log_2 .

Shannon's computation

Suppose we wish to compress a binary message $x_1^n = (x_1, \dots, x_n) \in \{0, 1\}^n$. Assume x_1^n is generated by n iid random variables $X_1^n = (X_1, \dots, X_n)$ where each X_i is Bernoulli of parameter p , for some $p \in (0, 1)$. We write P for the probability mass function of the X_i , i.e $P(x) = \mathbb{P}(X_i = x)$ for $x \in \{0, 1\}$.

Idea: give more likely strings shorter descriptions.

Question: how is the probability distributed among all such x_1^n ?

Let P^n denote the joint pmf of X_1^n . Then

$$\begin{aligned} \mathbb{P}(X_1^n = x_1^n) &= P^n(x_1^n) = \prod_{i=1}^n P(x_i) = 2^{\log \prod_{i=1}^n P(x_i)} \\ &= 2^{\sum_{i=1}^n \log P(x_i)} \\ &= 2^{k \log p + (n-k) \log(1-p)} \\ &= 2^{-n \left[-\frac{k}{n} \log p - \frac{n-k}{n} \log(1-p) \right]} \\ &\approx 2^{-n[-p \log p - (1-p) \log(1-p)]}. \quad (\text{LLN}) \end{aligned}$$

Where we have defined k to be the number of 1's in x_1^n . Now we define

$$h(p) = -p \log p - (1-p) \log(1-p)$$

so for large n we have

$$\mathbb{P}(X_1^n = x_1^n) \approx 2^{-nh(p)}$$

with high probability.

This means that for large n , the space $\{0, 1\}^n$ of all possible messages consists of:

1. non typical strings that have negligible probability of showing up;
2. approximately $2^{nh(p)}$ each of similar probability.

Note that the *binary entropy function* $h(p)$ has a maximum at $p = \frac{1}{2}$ with $h(1/2) = 1$ and is symmetric through $p = \frac{1}{2}$.

Back to data compression. Consider the following algorithm. Let $B_n \subseteq \{0, 1\}^n$ consist of the "typical" strings. Given x_1^n to compress:

- If $x_1^n \notin B_n \rightarrow$ declare "error";
- If $x_1^n \in B_n$, then describe it by describing its index j in B_n , where $1 \leq j \leq |B_n|$. This takes $\log |B_n| \approx nh(p)$ bits

Asymptotic Equipartition Property

Suppose X_1, X_2, \dots are iid random variables with values in a finite set, or *alphabet*, A . Let P denote the PMF of these variables, i.e $P(x) = \mathbb{P}(X_i = x)$, $x \in A$.

Theorem 0.1. Write $X_1^n = (X_1, X_2, \dots, X_n)$. Then

$$-\frac{1}{n} \log P^n(X_1^n) = -\frac{1}{n} \log \prod_{i=1}^n P(X_i) = \frac{1}{n} \sum_{i=1}^n [-\log P(X_i)] \xrightarrow{\mathbb{P}} H \text{ as } n \rightarrow \infty$$

where H is the entropy of X .

Proof. Law of large numbers. □

Definition. If $X \sim P$ on a finite alphabet A , the *entropy* of X is defined as

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

Notes.

1. $H(X) = \sum_{x \in A} P(x) \log(1/P(x))$;
2. By convention $0 \log 0 = 0$;
3. $H(X)$ is a function of P only, and in fact only depends on the probabilities $P(x)$, not the values of the random variable. In particular, if F is a bijection then $H(F(X)) = H(X)$;
4. $H(X) \geq 0$ with equality if and only if X is almost-surely constant;
5. For large n , $P^n(X_1^n) \approx 2^{-nH}$, with high probability. More formally,

$$\mathbb{P}\left(\left|-\frac{1}{n} \log P^n(X_1^n) - H\right| \leq \varepsilon\right) \rightarrow 1 \text{ as } n \rightarrow \infty.$$

Equivalently,

$$\mathbb{P}\left(\left\{x_1^n \in A^n : \left|-\frac{1}{n} \log P^n(x_1^n) - H\right| \leq \varepsilon\right\}\right) \rightarrow 1 \text{ as } n \rightarrow \infty$$

or,

$$P^n(B_n^*(\varepsilon)) \rightarrow 1 \text{ as } n \rightarrow \infty \forall \varepsilon > 0$$

where $B_n^*(\varepsilon) = \{x_1^n \in A^n : 2^{-n(H+\varepsilon)} \leq P^n(x_1^n) \leq 2^{-n(H-\varepsilon)}\}$ are the “typical strings”.

Theorem 0.2 (Asymptotic Equipartition Property). Suppose $(X_n)_{n \geq 1}$ is a sequence of iid random variables with PMF P on A . Then for any $\varepsilon > 0$:

- (\Rightarrow) : $|B_n^*(\varepsilon)| \leq 2^{n(H+\varepsilon)}$ for all $n \geq 1$, and $\mathbb{P}(X_1^n \in B_n^*(\varepsilon)) \rightarrow 1$ as $n \rightarrow \infty$.

- (\Leftarrow) if $(B_n)_{n \geq 1}$ is a sequence of sets with $B_n \subseteq A^n$ for all $n \geq 1$ such that $\mathbb{P}(X_1^n \in B_n) \rightarrow 1$ as $n \rightarrow \infty$, then $|B_n| \geq (1 - \varepsilon)2^{n(H - \varepsilon)}$ eventually.

Proof. For (\Rightarrow) we have

$$1 \geq P^n(B_n^*(\varepsilon)) = \sum_{x_1^n \in B_n^*(\varepsilon)} P^n(x_1^n) \geq |B_n^*(\varepsilon)| 2^{-n(H + \varepsilon)}$$

and $\mathbb{P}(x_1^n \in B_n^*(\varepsilon)) \rightarrow 1$ by the previous.

For (\Leftarrow), suppose $P^n(B_n) \rightarrow 1$ as $n \rightarrow \infty$. Then

$$P^n(B_n \cap B_n^*(\varepsilon)) = P^n(B_n) + P^n(B_n^*(\varepsilon)) - P^n(B_n \cup B_n^*(\varepsilon)) \rightarrow 1 + 1 - 1 = 1.$$

So eventually,

$$\begin{aligned} (1 - \varepsilon) &\leq P^n(B_n \cap B_n^*(\varepsilon)) \\ &\leq \sum_{x_1^n \in B_n \cap B_n^*(\varepsilon)} P^n(x_1^n) \\ &\leq |B_n \cap B_n^*(\varepsilon)| 2^{-n(H - \varepsilon)} \\ &\leq |B_n| 2^{-n(H - \varepsilon)}. \end{aligned}$$

□

Fixed-rate (lossless) data compression

Definition. A *source* (X_n) with alphabet A is a collection of random variables taking values in A . The source is *memoryless* if the X_i are iid with some common PMF P on A .

Definition. A *fixed-rate code* of block length n on a finite alphabet A is a collection of codebooks (B_n) where $B_n \subseteq A^n$. To compress $x_1^n \in A^n$:

- If $x_1^n \notin B_n$, then send “0” followed by x_1^n in binary. This will take $1 + \lceil \log |A^n| \rceil$ bits;
- If $x_1^n \in B_n$ then describe it by sending a “1” followed by the index of x_1^n in B_n , in binary. This takes $1 + \lceil \log |B_n| \rceil$ bits.

The *error probability* of the code is

$$P_e^{(n)} = \mathbb{P}(X_1^n \notin B_n) = P^n(B_n^c)$$

and its *rate* is

$$\frac{1}{n} (1 + \lceil \log |B_n| \rceil) \text{ bits/symbol.}$$

Question: if we require $P_e^{(n)} \rightarrow 0$, what is the best (i.e smallest possible) compression rate.

Theorem 0.3 (Fixed-rate coding theorem). *If (X_n) is a memoryless source with PMF P on A then for all $\varepsilon > 0$:*

- (\Rightarrow) *There is a code $(B_n^*(\varepsilon))$ with $P_e^{(n)} \rightarrow 0$ and rate less than or equal to $H + \varepsilon + \frac{2}{n}$ bits/symbol;*
- (\Leftarrow) *Any code has rate larger than $H - \varepsilon$ eventually, where $H = H(X_i)$ is the entropy.*

Proof. (\Rightarrow) Let $B_n^*(\varepsilon)$ be the typical sets. Then $P_e^{(n)} = P^n(B_n^*(\varepsilon)^c) \rightarrow 0$ by the AEP and the resulting rate is

$$\frac{1}{n} (1 + \lceil \log |B_n^*(\varepsilon)| \rceil) \leq \frac{1}{n} + \frac{1}{n} + \frac{1}{n} \log \left(2^{n(H+1)} \right) \leq H + \varepsilon + \frac{2}{n}.$$

(\Leftarrow) By the AEP, any code with $P_e^{(n)} \rightarrow 0$ has $|B_n| \geq (1 - \varepsilon)2^{n(H - \varepsilon)}$ eventually, so its rate is

$$\frac{1}{n} (1 + \lceil \log |B_n| \rceil) \geq \frac{1}{n} + \frac{1}{n} \log (1 - \varepsilon) + H - \varepsilon \geq H - \varepsilon.$$

□