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

Shannon's computation

Suppose we wish to compress a binary message $x_1^n = (x_1, \ldots, x_n) \in \{0,1\}^n$. Assume x_1^n is generated by n iid random variables $X_1^n = (X_1, \ldots, X_n)$ where each X_i is Bernouilli 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

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

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 $x_1^n \notin B_n \to$ 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 $X_1, X_2, \ldots$ 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 $X_1^n = (X_1, X_2, \ldots, X_n)$. Then*

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

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,

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

   Equivalently,

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

   or,

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

   where $B_n^*(\varepsilon) = \{x_1^n \in A : 2^{-n(H+\varepsilon)} \leq P^n(x_1^n) \leq 2^{-n(H-\varepsilon)}\}$ 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 $\mathbb{P}(X_1^n \in B_n^*(\varepsilon)) \to 1$ as $n \to \infty$.*

- ($\Leftarrow$) *if* $(B_n)_{n\geq 1}$ *is a sequence of sets with* $B_n \subseteq A^n$ *for all* $n \geq 1$ *such that* $\mathbb{P}(X_1^n \in B_n) \to 1$ *as* $n \to \infty$, *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 $\mathbb{P}(x_1^n \in B_n^*(\varepsilon)) \to 1$ by the previous.

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

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

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 $x_1^n \in A^n$:

(i) 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;

(ii) 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

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

Question: if we require $P_e^{(n)} \to 0$, 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 $P_e^{(n)} \to 0$ and rate less that 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 $P_e^{(n)} = P^n(B_n^*(\varepsilon)^c) \to 0$ by the AEP and the resulting rate is

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

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

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

□

Relative Entropy & Hypothesis Testing

Definition. Let P, Q be two PMFs on a discrete alphabet A. The *relative entropy* between P&Q is

$$D(P\|Q) = \sum_{x \in A} P(x) \log \frac{P(x)}{Q(x)}.$$

Notes. $D(P||Q)$ is not symmetric and it does not satisfy the triangle inequality. Despite this, we do think of this as a 'distance'.

Theorem 0.4 (Basic entropy bounds)**.**

(i) If X takes values in A, then

$$0 \leq H(x) \leq \log A$$

with equality in the first inequality if and only if X is uniform.

(ii) $D(P\|Q) \geq 0$ with equality if and only if $P = Q$.

Binary or simple-vs-simple hypothesis testing

Suppose X_1^n has iid entries from either P or Q on A. A *hypothesis test* is a decision region $B_n \subseteq A^n$ such that

$$\begin{aligned} x_1^n \in B_n &\rightarrow \text{ declare } X_1^n \sim P^n \text{ and} \\ x_1^n \notin B_n &\rightarrow \text{ declare } X_1^n \sim Q^n. \end{aligned}$$

The probabilities of error are

$$\begin{aligned} e_1^{(n)} &= \mathbb{P}(\text{declare } P | X_1^n \sim Q^n) = Q^n(B_n) \\ e_2^{(n)} &= \mathbb{P}(\text{declare } Q | X_1^n \sim P^n) = P^n(B_n^c). \end{aligned}$$

Question: if we require that $e_2^{(n)} \to 0$ as $n \to \infty$, how small can $e_1^{(n)}$ be?

Theorem 0.5 (Stein's Lemma)**.** *Suppose P, Q are PMFs on the same alphabet A such that $D(P\|Q) \neq 0, \infty$. Then for all $\varepsilon > 0$*

- *($\Rightarrow$) There are decision regions $B_n^*(\varepsilon)$ such that*

  $$e_1^{(n)} \leq 2^{-(D-\varepsilon)n} \text{ for all } n$$

  and $e_2^{(n)} \to 0$ as $n \to \infty$.

- *($\Leftarrow$) For any decision regions (B_n) such that*

  $$e_2^{(n)} \to 0 \text{ as } n \to \infty$$

  we have $e_1^{(n)} \geq 2^{-n(D+\varepsilon+\frac{1}{n})}$ eventually, where $D = D(P\|Q)$.

Proof. ($\Rightarrow$) Let us look at the likelihood ratio $\frac{P^n(x_1^n)}{Q^n(x_1^n)}$. If $X_1^n \sim P^n$, then

$$\frac{1}{n}\log\frac{P^n(X_1^n)}{Q^n(X_1^n)} = \frac{1}{n}\sum_{i=1}^{n}\log\frac{P(X_i)}{Q(X_i)} \xrightarrow{\mathbb{P}} D(P\|Q)$$

by the Law of Large Numbers.

This motivates the definition

$$B_n^*(\varepsilon) = \{x_1^n : 2^{n(D-\varepsilon)} \leq \frac{P^n(x_1^n)}{Q^n(x_1^n)} \leq 2^{n(D+\varepsilon)}\}$$

so we have $P^n(B_n^*(\varepsilon)) \to 1$. Hence $e_2^{(n)} = P^n(B_n^*(\varepsilon)^c) \to 0$. Also

$$\begin{aligned} 1 \geq P^n(B_n^*(\varepsilon)) = \sum_{x_1^n \in B_n^*(\varepsilon)} P^n(x_1^n) &= \sum_{x_1^n \in B_n^*(\varepsilon)} Q^n(x_1^n)\frac{P^n(x_1^n)}{Q^n(x_1^n)} \\ &\geq 2^{n(D-\varepsilon)}Q^n(B_n^*(\varepsilon)). \end{aligned}$$

($\Leftarrow$) Suppose $e_2^{(n)}(B_n) = P^n(B_n^c) \to 0$ and recall that also $e_2^{(n)}(B_n^*(\varepsilon)) = P^n(B_n^*(\varepsilon)^c) \to 0$ as $n \to \infty$. Then $P^n(B_n \cap B_n^*(\varepsilon)) \to 1$ as $n \to \infty$, and in particular

$$\begin{aligned} \frac{1}{2} \leq P^n(B_n \cap B_n^*(\varepsilon)) &= \sum_{x_1^n \in B_n \cap B_n^*(\varepsilon)} Q^n(x_1^n)\frac{P^n(x_1^n)}{Q^n(x_1^n)} \\ &\leq 2^{n(D+\varepsilon)}Q^n(B_n \cap B_n^*(\varepsilon)) \\ &\leq 2^{n(D+\varepsilon)}e_1^{(n)}(B_n). \end{aligned}$$

□

Note. The "likelihood-ratio typical" sets $B_n^*(\varepsilon)$ are *asymptotically* optimal, in that they achieve the best possible exponent for $e_1^{(n)}$, namely $D = D(P\|Q)$. But they are not optimal for finite n. Indeed, for each n the optimal decision regions are the *Neyman-Pearson tests*

$$B_{\mathrm{NP}} = \{x_1^n \in A^n : P^n(x_1^n) \geq T\} \text{ for some threshold } T.$$

Proposition 0.6.

$$B_{NP} = \{x_1^n : D(\hat{P}_n\|Q) \geq D(\hat{P}_n\|P) + \frac{1}{n}\log T\}$$

where

$$\hat{P}_n(a) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{x_i = a\}$$

is the empirical distribution.

Proof. Note that

$$
\begin{aligned}
\frac{1}{n}\log\frac{P^n(x_1^n)}{Q^n(x_1^n)} &= \frac{1}{n}\sum_{i=1}^{n}\log\frac{P(x_i)}{Q(x_i)} \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in A}\mathbb{1}\{x_i=a\}\log\frac{P(a)}{Q(a)} \\
&= \sum_{a\in A}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{x_i=a\}\log\frac{P(a)}{Q(a)} \\
&= \sum_{a\in A}\hat{P}_n(a)\log\left(\frac{P(a)}{Q(a)}\frac{\hat{P}_n(a)}{\hat{P}_n(a)}\right) \\
&= \sum_{a\in A}\hat{P}_n(a)\log\frac{\hat{P}_n(a)}{Q(a)} - \sum_{a\in A}\hat{P}_n(a)\log\frac{\hat{P}_n(a)}{P(a)} \\
&= D(\hat{P}_n\|Q) - D(\hat{P}_n\|P)
\end{aligned}
$$

□