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 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

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

$\square$

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

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

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

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

$\square$

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 <u>not</u> optimal for finite n. Indeed, for each n the optimal decision regions are the *Neyman-Pearson tests*

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

Proposition 0.6.

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

where

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

is the empirical distribution.

Proof. Note that

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

□

Proposition 0.7 (Log-sum inequality)**.** *For any $a_1, \dots, a_n, b_1, \dots, b_n \geq 0$,*

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}.$$

Moreover, we have equality if and only if a_i/b_i is constant over $i \in [n]$.

Proof. Let $f(x) = x \log x,\ x > 0$, which is strictly convex. Let $A = \sum_{i=1}^{n} a_i$ and $B = \sum_{i=1}^{n} b_i$. Define a random variable X which takes value a_i/b_i with probability b_i/B for $i \in [n]$. Then by Jensen's inequality

$$f(\mathbb{E}X) = f\left(\sum_{i=1}^{n} \frac{a_i}{b_i} \frac{b_i}{B} \right) = \frac{A}{B} \log \frac{A}{B}$$

so

$$\mathbb{E}(f(X)) = \sum_{i=1}^{n} \frac{a_i}{b_i} \log \frac{a_i}{b_i} \frac{b_i}{B} \geq f(\mathbb{E}X) = \frac{A}{B} \log \frac{A}{B}$$

by Jensen's inequality. We have equality if and only if X is constant, i.e a_i/b_i is constant for $i \in [n]$. $\square$

Proposition 0.8 (Basic entropy bounds)**.**

(i) If $X \sim P$ on a finite alphabet A, then $0 \leq H(X) \leq \log |A|$, with equality in the first inequality iff X is constant, and equality in the second indequality iff X is uniform on A.

(ii) If P, Q are PMFs on the same alphabet A then $D(P\|Q) \geq 0$ with equality if and only if $P = Q$.

Proof.

$$D(P\|Q) = \sum_{x \in A}^{n} P(x) \log \frac{P(x)}{Q(x)} \geq \left(\sum_{x \in A} P(x) \right) \log \frac{\sum_{x \in A} P(x)}{\sum_{x \in A} Q(x)} = 0$$

by the previous proposition, with equality if and only if $P(x)/Q(x)$ is constant over $x \in A$, i.e $P = Q$.

For (i), let Q be uniform on A and apply (ii):

$$0 \leq D(P\|Q) \leq \sum_{x \in A} P(x) \log \frac{P(x)}{1/|A|}$$

so

$$0 \leq \sum_{x \in A} P(x) \log P(x) + \sum_{x \in A} P(x) \log |A|$$

i.e $\log |A| - H(x) \geq 0$, with equality if and only if $P = Q$, i.e P is uniform on A. $\square$

Note. We saw that an iid sequence can at best be compressed to approximately $H(x_i)$ bits/symbol. The same source can be described, uncompressed using

$$\frac{1}{n}\lceil \log |A^n| \rceil \approx \log |A| \text{ bits/symbol.}$$

So compression is always possible, unless the source is "maximally" random, i.e iid uniform.

Recall our hypothesis testing setting. Data x_1^n generated iid either from P or Q. Then we had a decision region B_n (declaring P if $x_1^n \in B_n$ and Q otherwise) and error probabilities

$$e_1^{(n)}(B_n) = Q^n(B_n) \text{ and } e_2^{(n)} = P^n(B_n^c).$$

Stein's lemma told us that the likelihood ratio-typical decision regions

$$B_n^*(\varepsilon) = \left\{ x_1^n \in A^n : 2^{n(D-\varepsilon)} \leq \frac{P^n(x_1^n)}{Q^n(x_1^n)} \leq 2^{n(D+\varepsilon)} \right\} \text{ where } D = D(P\|Q)$$

are asymptotically optimal , i.e

$$e_1^{(n)}(B_n^*(\varepsilon)) \approx 2^{-nD} \text{ and } e_2^{(n)}(B_n^*(\varepsilon)) \to 0.$$

Recall the Neyman-Pearson decision regions

$$B_{\text{NP}} = \left\{ x_1^n : \frac{P(x_1^n)}{Q^n(x_1^n)} \geq T \right\} \text{ for } T > 0$$

turn out to be optimal for finite n.

Theorem 0.9 (Neyman-Pearson Lemma). *If $e_2^{(n)}(B_n) \leq e_2^{(n)}(B_{NP})$ then $e_1^{(n)}(B_n) \geq e_1^{(n)}(B_{NP})$.*

Proof. Observe that for all x_1^n:

$$[\mathbb{1}_{B_{\text{NP}}}(x_1^n) - \mathbb{1}_{B_n}(x_1^n)]\,[P^n(x_1^n) - TQ^n(x_1^n)] \geq 0$$

so summing over all x_1^n we get

$$P^n(B_{\text{NP}}) - TQ^n(B_{\text{NP}}) - P^n(B_n) + TQ^n(B_n) \geq 0$$

and so

$$1 - e_2^{(n)}(B_{\text{NP}}) - Te_1^{(n)}(B_{\text{NP}}) - \left[1 - e_2^{(n)}(B_n)\right] + Te_1^{(n)}(B_n) \geq 0$$

giving

$$e_2^{(n)}(B_n) - e_2^{(n)}(B_{\text{NP}}) \geq T\left[e_1^{(n)}(B_{\text{NP}}) - e_1^{(n)}(B_n)\right].$$

□

Definition. The *type* $\hat{P}_n$ r $\hat{P}_{x_1^n}$ of a string $x_1^n \in A^n$ is simply its empirical distribution, i.e

$$\hat{P}_n(a) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{a \in X_i\} \text{ for } a \in A.$$

Recall

Proposition. *We have*

$$B_{NP} = \{x_1^n \in A^n : D(\hat{P}_n \| Q) \geq D(\hat{P}_n \| P) + T'\} \textit{ where } T' = \frac{1}{n}\log T.$$

Definition. If X, Y are discrete random variables with values in A, B respectively and joint PMF $P_{X,Y}$, we define the *joint entropy*

$$H(X,Y) = \mathbb{E}[-\log P_{X,Y}(X,Y)] = \sum_{\substack{x\in A\\ y\in B}} P_{X,Y}(x,y)\log\frac{1}{P_{X,Y}(x,y)}$$

and similarly for n (not necessarily iid) random variables

$$H(X_1^n) = \mathbb{E}[-\log P_{X^n}(X_1^n)].$$

Example. Suppose $X \sim P_X$ and $Y \sim P_Y$ are independent. Then

$$\begin{aligned}H(X,Y) = \mathbb{E}[-\log(P_X(X)P_Y(Y))] &= \mathbb{E}[-\log P_X(X)] + \mathbb{E}[-\log P_Y(Y)]\\ &= H(X) + H(Y).\end{aligned}$$

In general, $P_{XY}(x,y) = P_X(x)P_{Y|X}(y|x)$, so

$$H(X,Y) = \mathbb{E}[-\log P_X(X)] + \mathbb{E}[-P_{Y|X}(Y|X)] = H(X) + H(Y|X).$$

Definition. The *conditional entropy* of Y given X is

$$H(Y|X) = \mathbb{E}[-\log P_{X|Y}(X|Y)] = \sum_{x,y} P_{XY}(x,y)\log P_{Y|X}(y|x).$$

Note. We also have

$$\begin{aligned}H(Y|X) &= \sum_x P_X(x)\sum_y P_{Y|X}(y|x)\log P_{Y|X}(y|x)\\ &= \sum_x P_X(x)H(Y|X=x).\end{aligned}$$

Hence if Y takes values in A_Y, we have $0 \le H(Y|X) \le \log|A_Y|$, since $0 \le H(Y|X=x) \le \log|A_Y|$.

Proposition 0.10 ('Chain rule')**.** *If X_1^n are n arbitrary discrete random variables, then*

$$\begin{aligned}H(X_1^n) &= H(X_1) + H(X_2|X_1) + \ldots + H(X_n|X_1^{n-1})\\ &= \sum_{i=1}^n H(X_i|X_1^{i-1}).\end{aligned}$$

If the random variables are independent, then $H(X_1^n) = \sum_{i=1}^n H(X_i)$.

Proof. Since $P_{X_1^n}(x_1^n) = \prod_{i=1}^n P_{X_i|X_1^{i-1}}(x_i|x_1^{i-1})$ we can just take log-expectations. □

Proposition 0.11 ('Conditioning reduces entropy')**.** *We have $H(Y|X) \le H(Y)$, with equality if and only if X, Y are independent.*

Proof.

$$
\begin{aligned}
H(Y) - H(Y|X) &= \mathbb{E}[-\log P_Y(Y)] - \mathbb{E}[-\log P_{Y|X}(Y)] \\
&= \mathbb{E}\left(\log\left(\frac{P_{Y|X}(Y)}{P_Y(Y)}\frac{P_X(X)}{P_X(X)}\right)\right) \\
&= \mathbb{E}\left(\log\frac{P_{XY}(X,Y)}{P_{X(X)P_Y(Y)}}\right) \\
&= D(P_{XY}\|P_X P_Y) \geq 0
\end{aligned}
$$

with equality if and only if $P_{XY} = P_X P_Y$, i.e X, Y are independent. □

Corollary 0.12 (Subadditivity of entropy)**.** *$H(X_1^n) \leq H(X_1) + H(X_2) + \ldots + H(X_n)$, with equality if and only if the X_i are independent.*

Proposition 0.13 (Data processing inequalities for entropy)**.** *For any discrete random variable X on A and function f on A:*

(a) $H(f(X)|X) = 0$;

(b) $H(f(X)) \leq H(X)$ with equality iff f is injective.

Proof.

(a) We have $H(X) = H(X, f(X))$ since $x \mapsto (x, f(x))$ is injective. Then $H(f(X)|X) = H(X, F(X)) - H(X) = 0$;

(b) We have $H(f(X)) = H(X, f(X)) - H(X|f(X)) \leq H(X, f(X)) = H(X)$ with equality if and only if $H(X|f(X)) = 0$, i.e f is injective.

□

Proposition 0.14 (Properties of conditional entropy)**.**

(a) $H(X, Y|Z) = H(X|Z) + H(Y|X, Z)$;

(b) $H(Y|X, Z) = H(Y|Z)$;

(c) $H(X, Y|Z) \leq H(X|Z) + H(Y|Z)$.

Furthermore we have equality in (b) and (c) if and only if X and Y are conditionally independent given Z.

Proof. Exercise. □

Theorem 0.15 (Fano's inequality)**.** *Suppose X, Y are discrete random variables taking values in A, B respectively. Let $\hat{X} = f(Y)$ for some function $f : B \to A$ and let $p_e = \mathbb{P}(\hat{X} \neq X)$. Then*

$$H(X|Y) \leq h(p_e) + p_e \log(|A| - 1)$$

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

Proof. Let $E = \mathbb{1}\{X \neq \hat{X}\}$ so that $E \sim \text{Bern}(p_e)$. Then by the chain rule

$$\begin{aligned} H(X,E|Y) &= H(X|Y) + \underbrace{H(E|X,Y)}_{=0} \\ &= H(E|Y) + H(X|E,Y) \end{aligned}$$

hence

$$\begin{aligned} H(X|Y) &= H(E|Y) + H(X|E,Y) \\ &\leq H(E) + \mathbb{P}(E=1)\underbrace{H(X|E=1,Y)}_{\leq \log(|A|-1)} + \mathbb{P}(E=0)\underbrace{H(X|E=0,Y)}_{=0} \\ &\leq h(p_e) + p_e \log(|A|-1). \end{aligned}$$

□

Proposition 0.16 (Data processing for relative entropy)**.** *Suppose $X \sim P_X$ and $Y \sim P_Y$ on A. Let $f : A \to B$ and $f(X) \sim P_{f(X)}$, $f(Y) \sim P_{f(Y)}$. Then $D(P_{f(X)}||P_{f(Y)}) \leq D(P_X||P_Y)$.*

Proof. For $z \in B$ define $A_z = f^{-1}(\{z\})$. Then

$$\begin{aligned} D(P_X \| P_Y) &= \sum_{x\in A} P_X(x) \log \frac{P_X(x)}{P_Y(x)} \\ &= \sum_{z\in B} \sum_{x \in A_z} P_X(x) \log \frac{P_X(x)}{P_Y(x)} \\ &\geq \sum_{z \in B} \left(\sum_{x\in A_z} P_X(x) \right) \log \left(\frac{\sum_{x\in A_z} P_X(x)}{\sum_{x\in A_z} P_Y(x)} \right) \\ &= \sum_{z\in B} P_{f(X)}(y) \log \frac{P_{f(X)}(y)}{P_{f(Y)}(y)} \\ &= D(P_{f(X)} \| P_{f(Y)}). \end{aligned}$$

□

Definition. The *total variation distance* between two PMF's P, Q on the same alphabet A is

$$\|P - Q\|_{TV} = \sum_{x\in A} |P(x) - Q(x)|.$$

Theorem 0.17 (Pinsker's inequality)**.** *For PMF's P, Q on the same alphabet A we have*

$$\|P - Q\|_{TV}^2 \leq (2\log_e(2)) D(P\|Q) = 2D_e(P\|Q)$$

where $D_e(P\|Q) = \sum_{x\in A} P(x) \ln(P(x)/Q(x))$.

Note. If we let $B = \{x : P(x) > Q(x)\}$ we can write

$$\begin{aligned}\|P-Q\|_{TV} &= \sum_{x\in B} |P(x)-Q(x)| + \sum_{x\in B^c} |P(x)-Q(x)| \\ &= \sum_{x\in B} (P(x)-Q(x)) + \sum_{x\in B^c} (Q(x)-P(x)) \\ &= P(B) - Q(B) + Q(B^c) + P(B^c) \\ &= 2(P(B)-Q(B)).\end{aligned}$$

Proof. First suppose $P \sim \text{Bern}(p)$ and $Q \sim \text{Bern}(q)$ with $0 \le q \le p \le 1$ wlog (otherwise take $p \mapsto 1-p$ and $q \mapsto 1-q$). Let $\Delta(p,q) = 2D_e(P\|Q) - \|P-Q\|_{TV}^2$. Fix p and note that $\Delta(p,p) = 0$. Then (using the previous note to simplify $\|P-Q\|_{TV}$)

$$\Delta(p,q) = 2p\log p - 2p\log q + 2(1-p)\log(1-p) - 2(1-p)\log(1-q) - (2(p-q))^2$$

so differentiating Δ with respect to q gives

$$-2\frac{p}{q} + 2\frac{1-p}{1-q} + 8(p-q) = 2(q-p)\left[\frac{1}{q(1-q)} - 4\right] \le 0.$$

Therefore $\Delta(p,q) \ge 0$, so we have the Bernouilli case.

In the general case $X \sim P$ and $Y \sim Q$, let $B = \{x : P(x) > Q(x)\}$ and $x' = \mathbb{1}\{X \in B\}$, $Y' = \mathbb{1}\{Y \in B\}$, so that $X' \sim \text{Bern}(P(B))$, $Y' \sim \text{Bern}(Q(B))$. Then

$$\begin{aligned}\|P-Q\|_{TV}^2 = (2(P(B)-Q(B)))^2 &= \|P_{X'} - P_{Y'}\|_{TV}^2 \\ &\le 2D_e(P_{X'}\|P_{Y'}) && \text{(Bernouilli case)} \\ &\le 2D_e(P\|Q). && \text{(Data processing)}\end{aligned}$$

□

Poisson Appoximation

Suppose $X_1, \dots, X_n \sim \text{Bern}(\lambda/n)$ are iid. Then $S_n = \sum_{i=1}^n X_i \sim \text{Bin}(n, \lambda/n)$ and we have $P_{S_n} \to \text{Poi}(\lambda)$ as $n \to \infty$. This phenomenon is in fact much more general.

If $X_1, \dots, X_n \sim \text{Bern}(p_i)$ and $S_n = \sum_{i=1}^n X_i \sim P_{S_n}$. Then $P_{S_n} \approx P_0(\lambda)$ as long as:

(i) The p_i are small;

(ii) The X_i ae only weakly dependent.

Theorem 0.18 (Poisson Approximation)**.** *Suppose* $X_i \sim \mathrm{Bern}(p_i)$, $i \in [n]$, *and let* $S_n = \sum_{i=1}^n X_i \sim P_{S_n}$ *and* $\lambda = \sum_{i=1}^n p_i$. *Then*

$$D_e(P_{S_n} \| \mathrm{Poi}(\lambda)) \leq \sum_{i=1}^n p_i^2 + \left[\sum_{i=1}^n H(X_i) - H(X_1^n) \right].$$

Example. In the classical case this gives

$$\|P_{S_n} - \mathrm{Poi}(\lambda)\|_{TV} \leq \frac{2\lambda}{\sqrt{n}}.$$

Proof. Let $Z_i \sim \mathrm{Poi}(p_i)$ be independent for $i \in [n]$. Then $T_n = \sum_{i=1}^n Z_i \sim \mathrm{Poi}(\lambda)$. Now

$$\begin{aligned}
&D_e(P_{S_n} \| \mathrm{Poi}(\lambda)) = D_e(P_{S_n} \| P_{T_n}) \\
&\leq D_e(P_{X_1^n} \| P_{Z_1^n}) \\
&= \mathbb{E}\left(\ln \left(\frac{P_{X_1^n}(X_1^n)}{P_{Z_1^n}(X_1^n)} \times \frac{\prod_{i=1}^n P_{X_i}(X_i)}{\prod_{i=1}^n P_{X_i}(X_i)} \right) \right) \\
&= \mathbb{E}\left(\ln \prod_{i=1}^n \frac{P_{X_i}(X_i)}{P_{Z_i}(X_i)} \right) - \mathbb{E}\left(\ln \left(\prod_{i=1}^n P_{X_i}(X_i) \right) \right) + \mathbb{E}\left(\ln P_{X_1^n}(X_1^n) \right) \\
&= \sum_{i=1}^n \mathbb{E}\left(\ln \frac{P_{X_i}(X_i)}{P_{Z_i}(X_i)} \right) + \sum_{i=1}^n \mathbb{E}\left(-\ln P_{X_i}(X_i) \right) - H(X_1^n) \\
&= \sum_{i=1}^n \underbrace{D_e(\mathrm{Bern}(p_i) \| \mathrm{Poi}(p_i))}_{\leq p_i^2} + \sum_{i=1}^n H(X_i) - H(X_1^n).
\end{aligned}$$

□

Mutual Information

Definition. If X, Y are two discrete random variables, the *mutual information between X and Y is*

$$I(X;Y) = H(X) - H(X|Y).$$

Proposition 0.19.

$$\begin{aligned} I(X;Y) = H(X) + H(Y) - H(X,Y) &= \mathbb{E}\left[\log \frac{P_{X,Y}(X,Y)}{P_X(X)P_Y(Y)}\right] \\ &= D(P_{XY} \| P_X P_Y). \end{aligned}$$

Proof. Trivial. □

Note. This implies the mutual information is symmetric, i.e $I(X;Y) = I(Y;X)$.

Proposition 0.20.

1. *$I(X;Y) \geq 0$ with equality if and only if X, Y are independent;*
2. *$I(X;Y) \leq H(X)$.*

Proof. Trivial. □

Definition. The *conditional mututal information* $H(X;Y|Z)$ is defined by

$$H(X;Y|Z) = H(X|Z) - H(X|Y,Z).$$

Note. Conditional mutual information satisfies properties analogous to those of the usual mutual information. For example $I(X;Y|Z) \geq 0$ with equality iff X, Y are conditionally independent given Z.

Proposition 0.21 (Chain rule for mutual information)**.**

$$I(X_1^n;Y) = \sum_{i=1}^{n} H(X_i;Y|X_1^{i-1}).$$

Proof. Trivial. □

Proposition 0.22 (Data processing)**.** *If $Z = f(Y)$ or, more generally, if X-Y-Z (X, Z are conditionally independent given Y), then*

1. *$I(X;Y) \geq I(X;Z)$;*
2. *$I(X;Y) \geq I(X;Y|Z)$.*

Proof.

$$\begin{aligned} I(X,Y;Z) &= I(X;Y) + \underbrace{I(X;Z|Y)}_{=0} && \text{(chain rule)} \\ &= I(X;Z) + I(X;Y|Z). && \text{(chain rule)} \end{aligned}$$

Hence

$$I(X;Y) = I(X;Z) + I(X;Y|Z).$$

□

Synergy

Definition. The *synergy* between X and Y_1, Y_2 is

$$\begin{aligned} S(X;Y_1,Y_2) &= I(X;Y_1,Y_2) - [I(X;Y_1) + I(X;Y_2)] \\ &= I(X;Y_2|Y_1) - I(X;Y_2). \end{aligned}$$

Remark. The synergy can be either positive or negative.

Proposition 0.23. *Consider the following scheme*

$$\begin{tikzcd} & X \arrow[dl] \arrow[dr] \arrow[d, phantom, "P(y|x)"] & \\ Y_1 \arrow[d] & & Y_2 \arrow[d, "Q(z|y)"'] \\ Z_1 & & Z_2 \end{tikzcd} \qquad \begin{tikzcd} & X \arrow[d, "P(y|x)"'] & \\ & Y \arrow[dl] \arrow[dr] \arrow[d, phantom, "Q(w|y)"] & \\ W_1 & & W_2 \end{tikzcd} .$$

Then if $S(X;W_1,W_2) > 0$, we have

$$I(X;W_1,W_2) > I(X;Z_1,Z_2).$$

Proof. We have

$$I(X;W_2|W_1) > I(X;W_2) = I(X;Z_2).$$

Hence

$$I(X;W_2|W_1) \geq I(X;Z_2|Z_1) \qquad \text{(data processing)}$$

also

$$I(X;W_1) = I(X;Z_1)$$

which, by combining and the chain rule, these we have

$$I(X;W_1,W_2) > I(X;Z_1,Z_2).$$

$\square$

Theorem 0.24 (Maximum Entropy Property of Poisson)**.**

$$H(\mathrm{Po}(\lambda)) = \sup\left\{H(P_{S_n}) : S_n = \sum_{i=1}^n X_i,\ X_i \sim \mathrm{Bern}(p_i)\text{indep}, \sum_{i=1}^n p_i = \lambda,\ n \geq 1\right\}.$$

Proof.

$$\begin{aligned} &\sup\left\{H(P_{S_n}) : S_n = \sum_{i=1}^n X_i,\ X_i \sim \mathrm{Bern}(p_i)\text{indep}, \sum_{i=1}^n p_i = \lambda\right\} \\ &= \sup_{n\geq 1} H(\mathrm{Bin}(n, \lambda/n)) && (1) \\ &= \lim_{n\to\infty} H(\mathrm{Bin}(n, \lambda/n)) && (2) \\ &= H(\mathrm{Po}(\lambda)) \end{aligned}$$

$\square$

Entropy & Additive Combinatorics

In this section, all random variables take values in $\mathbb{Z}$.
Suppose A, B are finite subsets of $\mathbb{Z}$. Define $A+B = \{a+b : a \in A, b \in B\}$ and $A - B = \{a - b : a \in A, b \in B\}$. Then $|A| \leq |A+B| \leq |A||B|$.

Proposition 0.25 (Ruzsa triangle inequality)**.** *We have* $|A-C| \leq \frac{|A-B||B-C|}{|B|}$.

Proof. It suffices to construct an injective map $f : B \times (A-C) \to (A-B) \times (B-C)$. For any $y \in A-C$ there exist $a \in A$ and $c \in C$ such that $y = a - c$. Choose and fix such a pair a_y, c_y for each $y \in A - C$, and define

$$f(x, y) = (a_y - x, x - c_y).$$

This is injective since $(a_y - x) + (x - c_y) = a_y - c_y = y$ so we can recover y, which gives c_y and then $(x - c_y) + c_y = x$ so we can recover x and thus (x, y). $\square$

Observe that the above proof uses the "data-processing like" property that $a - x + x - c = a - c$.

Idea: suppose $X_1, \dots, X_n$ are iid copies of $X \sim P$ on A. Then the AEP tells us that their joint PMF P^n is essentially supported on the set of $\approx 2^{nH} = (2^H)^n$ typical strings, instead of the full $|A|^n$ collection of all possible strings. Therefore we think of 2^H as the *essential support size of the PMF P.*
Rusza-Tao Correspondence: given a bound on cardinalities of subsets and different sets, replace sets by independent random variables and log-cardinalities by entropies, to get a candidate entropy bound!

Example. The bound $|A| \leq |A+B| \leq |A||B|$ corresponds to $H(X) \leq H(X+Y) \leq H(X) + H(Y)$. In the latter the first inequality follows from $H(X) + H(Y) = H(X,Y) = H(X, X+Y)$ (data processing) then $H(X, X+Y) = H(X+Y) + H(Y|X+Y) \leq H(X+Y) + H(Y)$. The second inequality follows from $H(X+Y) \leq H(X,Y)$ (data processing) and $H(X,Y) = H(X) + H(Y)$.

The Rusza triangle inequality motivates

Theorem 0.26 (Rusza triangle inequality for entropy)**.** *If X, Y, Z are independent, then*

$$H(X-Z) + H(Y) \leq H(X-Y) + H(Y-Z).$$

Proof. First observe that $(X, (X-Y, Y-Z), (X-Z))$ form a Markov chain of the form $(u, v, f(v))$. So by the data processing inequality for mutual information,

$$I(X; (X-Y, Y-Z)) \geq I(X; X-Z)$$

i.e

$$\begin{aligned} H(X-Z)-H(Z) &= H(X-Z)-H(X-Z|X) \\ &= I(X;X-Z) \\ &\leq I(X;(X-Y,Y-Z)) \\ &= H(X)+H(X-Y,Y-Z)-H(X,X-Y,Y-Z) \\ &= H(X)+H(X-Y)+H(Y-Z)-H(X,Y,Z) \\ &= H(X-Y)+H(Y-Z)-H(Y)-H(Z). \end{aligned}$$

□

Theorem 0.27 (Doubling-difference inequality). *If X_1, X_2 are iid then*

$$\frac{1}{2} \leq \frac{H(X_1+X_2)-H(X_1)}{H(X_1-X_2)-H(X_1)} \leq 2.$$

We need a couple of lemmas before proving this:

Lemma 0.28. *If X, Y, Z are independent, then*

$$H(X-Z)+H(Y) \leq H(X+Y)+H(Y+Z).$$

Proof. This is the Rusza triangle inequality with Y replaced by $-Y$. □

Lemma 0.29. *For X, Y, Z independent we have*

$$H(X+Y+Z)+H(Y) \leq H(X+Y)+H(Y+Z).$$

Proof. Since $(X, X+Y, X+Y+Z)$ forms a Markov chain, we have

$$I(X;X+Y) \geq I(X;X+Y+Z).$$

Hence

$$\begin{aligned} H(X+Y)-H(X+Y|X) &= H(X+Y)-H(Y) \\ &\geq H(X+Y+Z)-H(X+Y+Z|X) \\ &= H(X+Y+Z)-H(Y+Z). \end{aligned}$$

□

Now we can prove:

Theorem 0.30 (Doubling-difference inequality). *If X_1, X_2 are iid then*

$$\frac{1}{2} \leq \frac{H(X_1+X_2)-H(X_1)}{H(X_1-X_2)-H(X_1)} \leq 2.$$

Proof. For the lower bound, take X, Y, Z to be iid so by the first lemma

$$H(X - Z) + H(X) \leq 2H(X + Z)$$

and therefore

$$H(X - Z) - H(X) \leq 2[H(X + Z) - H(X)]$$

giving the lower bound. For the upper bound, replacing Y by $-Y$ in the second lemma gives

$$H(X - Y + Z) + H(Y) \leq H(X - Y) + H(Z - Y)$$

so if X, Y, Z are iid

$$H(X+Z)+H(X) = H(X+Z)+H(Y) \leq H(X-Y+Z)+H(Y) \leq 2H(X-Z).$$

$\square$

Entropy Rate

Definition. The *entropy rate* of a source $X = (X_n)_{n\geq 1}$ with alphabet A is

$$H = H(X) = \lim_{n\to\infty} \frac{H(X_1^n)}{n} \text{ bits/symbol}$$

whenever the limit exists.

Example. If X is memoryless (i.e X_n are iid) then $H(X_1^n) = nH(X_1)$ so $H(X)$ is $H(X_1)$.

Example. Suppose X is an ergodic (i.e irreducible and aperiodic) markov chain on the state space A, with $X_1 \sim P_{X_1}$ and transition matrix $Q = (Q_{xx'})_{x,x'\in A}$ where $Q_{xx'}\mathbb{P}(X_{n+1} = x'|X_n = x)$. Let π denote the unique stationary distribution of X. Let $\bar{X} = (\bar{X}_n)_{n\geq 1}$ be the stationary version of X (i.e $\bar{X}_1 \sim \pi$ and $\bar{X}$ has the same transition matrix as X). Then

$$\begin{aligned} H(X_1^n) = \sum_{i=1}^n H(X_i|X_1^{i-1}) \\ &= \sum_{i=1}^n H(X_i|X_{i-1}) && \text{(Markov property)} \\ &= H(X_1) - H(X_{n+1}|X_n) + \sum_{i=2}^{n+1} H(X_i|X_{i-1}). \end{aligned}$$

Since X is ergodic, $P_{X_n} \to \pi$ as $n \to \infty$ and $P_{X_n,X_{n+1}}(x,x') \to \pi_x Q_{xx'} = P_{\bar{X}_1,\bar{X}_2}(x,x')$ for all $x,x' \in A$. Since conditional entropy is a continuous functional of the joint PMF, $H(X_{n+1}|X_n) \to H(\bar{X}_2|\bar{X}_1)$. So

$$\frac{1}{n}H(X_1^n) \to H(\bar{X}_2|\bar{X}_1)$$

i.e $H(X) + H(\bar{X}_2|\bar{X}_1)$.

Definition. $X = (X_n)_{n\geq 1}$ is *stationary* if for each n, X_{k+1}^{k+n} is independent of k.

Proposition 0.31. *If X is stationary then the entropy rate exists and is*

$$H(X) = \lim_{n\to\infty} \frac{H(X_1^n)}{n} = \lim_{n\to\infty} H(X_n|X_1^{n-1}) \textit{ bits/symbol.}$$

Proof. Note that

$$\begin{aligned} H(X_n|X_1^{n-1}) &= H(X_{n+1}|X_2^n) && \text{(stationarity)} \\ &\geq H(X_{n+1}|X_1^n). && \text{(conditioning reduces entropy)} \end{aligned}$$

Hence the sequence $(H(X_n|X_1^{n-1}))_{n\geq 1}$ is decreasing and bounded below, so the limit $\lim_{n\to\infty} H(X_n|X_1^{n-1})$ exists. Also

$$\frac{1}{n}H(X_1^n) = \frac{1}{n}\sum_{i=1}^{n} H(X_i|X_1^{i-1}) \xrightarrow{n\to\infty} \lim_{n\to\infty} H(X_n|X_1^{n-1}).$$

□

Recall: if $\bar{X} = (X_n)_{n\geq 1}$ is stationary, then it always admits a unique two-sided extension to $(X_n)_{n\in\mathbb{Z}}$ (by Kolmogorov's extension theorem).

Proposition 0.32. *If X is a stationary source then its entropy rate can also be expressed as*

$$H(X) = \lim_{n\to\infty} H(X_0|X_{-n}^{-1}) = H(X_0|X_{-\infty}^{-1}) := \mathbb{E}[-\log P(X_0|X_{-\infty}^{-1})].$$

The following proof is **non-examinable**:

Proof. First let $P(x_0|X_{-\infty}^{-1}) = \mathbb{P}(X_0 = x_0|X_{-\infty}^{-1})$ denote the regular conditional distribution of X_0 given $X_{-\infty}^{-1}$. Then by martingale convergence, we know that $P(x|X_{-n}^{-1}) \to P(x|X_{-\infty}^{-1})$ almost-surely as $n\to\infty$. Since $p \mapsto p\log p$ is bounded on $(0,1)$, by the bounded convergence theorem we have

$$\begin{aligned} H(X_0|X_{-n}^{-1}) &= -\mathbb{E}\left[-\sum_x P(x|X_{-n}^{-1})\log P(x|X_{-n}^{-1})\right] \\ &\to \mathbb{E}\left[-\sum_x P(x|X_{-\infty}^{-1})\log P(x|X_{-\infty}^{-1})\right] \\ &= H(X_0|X_{-\infty}^{-1}). \end{aligned}$$

And finally, by stationarity

$$H(X_n|X_1^{n-1}) = H(X_0|X_{-n+1}^{-1}) \to H(X_0|X_{-\infty}^{-1}).$$

□

Ergodic Theorem

Consider the space $A^{\mathbb{Z}}$ of all doubly-infinite strings $x = (x_n)_{n=-\infty}^{\infty}$ with values in A, and define the shift map $T : A^{\mathbb{Z}} \to A^{\mathbb{Z}}$ by $(Tx)_n = x_{n+1}$. Then a stationary source X is ergodic if and only if the following holds:

Theorem 0.33 (Birkhoff's Ergodic Theorem)**.** *If $f : A^{\mathbb{Z}} \to \mathbb{R}$ has $\mathbb{E}|f(X_{-\infty}^{\infty})| < \infty$ then*

$$\frac{1}{n}\sum_{i=1}^{n} f(T^i X_{-\infty}^{\infty}) \to \mathbb{E}(f(X_{-\infty}^{\infty})) \text{ almost-surely.}$$

Example. If $f(x_{-\infty}^{\infty}) = g(x_0)$ and X is IID, we recover the SLLN.

Definition. A stationary source $(X_n)_{n\geq 1} = X$ on A is *ergodic* if and only if all invariant events are trivial, i.e whenever $T^{-1}(B) = B$ we have $\mathbb{P}(X_{-\infty}^{\infty} \in B) \in \{0, 1\}$.

Theorem 0.34 (Shannon-McMillan-Breiwan Theorem). *If $X = (X_n)_{n\geq 1}$ is a stationary and ergodic source on a finite alphabet A, with entropy rate H, and P_n denotes the PMF of X_1^n, then*

$$-\frac{1}{n}\log P_n(X_1^n) \xrightarrow{n\to\infty} H \text{ almost-surely.}$$

So for large n, $P_n(x_1^n) \approx 2^{-nH}$ with high probability.

Exercise: prove the AEP for stationary and ergodic sources, as well as the fixed-rate coding theorem.

Proof. We have

$$-\frac{1}{n}\log P_n(X_1^n) = \frac{1}{n}\sum_{i=1}^n [-\log P(X_i|X_1^{i-1})]$$

we would like to apply the Ergodic Theorem, but this is not of the form $\frac{1}{n}\sum_{i=1}^b f(T^i x)$. Instead, we first consider an "infinite-memory" version. Note

$$\begin{aligned} -\frac{1}{n}\log P(X_1^n|X_{-\infty}^0) &= \frac{1}{n}\sum_{i=1}^n [-\log P(X_i|X_{-\infty}^{i-1})] \\ &\xrightarrow{n\to\infty} \mathbb{E}[-\log P(X_0|X_{-\infty}^{-1})] \\ &= H(X_0|X_{-\infty}^{-1}) = H. \end{aligned}$$

Then we consider a fixed-memory version: define a new sequence of PMFs Q_n by $Q_n = P_n$ for $n \leq k$, and for $n \geq k+1$, $Q_n(x_1^n) = Q_k(x_1^k)\prod_{i=k+1}^n P(X_i|X_{i-k}^{i-1})$. Then

$$\begin{aligned} &-\frac{1}{n}\log Q_n(X_1^n) \\ &= -\frac{1}{n}\log Q_k(X_1^k) + \frac{1}{n}\sum_{i=1}^n [-\log P(X_i|X_{i-k}^{i-1})] - \frac{1}{n}\sum_{i=1}^k [-\log P(X_i|X_{i-k}^{i-1})] \end{aligned}$$

so by the Ergodic Theorem

$$\begin{aligned} \lim_{n\to\infty} -\frac{1}{n}\log Q_n(X_1^n) &= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n [-\log P(X_i|X_{i-k}^{i-1})] \\ &= \mathbb{E}(-\log P(X_0|X_{-k}^{-1})) \\ &= H(X_0|X_{-k}^{-1}) \end{aligned}$$

almost-surely. Since $H(X_0|X_{-k}^{-1}) \to H(X_0|X_{-\infty}^{-1}) = H$ by a previous lemma, the rest of the proof is given in the next two lemmas. $\square$

Lemma 0.35.

$$\limsup_{n\to\infty}\left[-\frac{1}{n}\log P(X_1^n|X_{-\infty}^0) - \left[-\frac{1}{n}\log P_n(X_1^n)\right]\right] \leq 0$$

almost-surely.

Proof. Let $\varepsilon > 0$. Then

$$\begin{aligned}
&\mathbb{P}\left(-\frac{1}{n}\log P(X_1^n|X_{-\infty}^0) - \left[-\frac{1}{n}\log P_n(X_1^n)\right] > \varepsilon\right) \\
&= \mathbb{P}\left(\frac{1}{n}\log\frac{P_n(X_1^n)}{P(X_1^n|X_{-\infty}^0)} > \varepsilon\right) \\
&= \mathbb{P}\left(\frac{P_n(X_1^n)}{P(X_1^n|X_{-\infty}^0)} > 2^{n\varepsilon}\right) \\
&\leq 2^{-n\varepsilon}\mathbb{E}\left[\frac{P_n(X_1^n)}{P(X_1^n|X_{-\infty}^0)}\right] \\
&= 2^{-n\varepsilon}\mathbb{E}\left[\mathbb{E}\left(\frac{P_n(X_1^n)}{P(X_1^n|X_{-\infty}^0)}|X_{-\infty}^0\right)\right] \\
&= 2^{-n\varepsilon}\mathbb{E}\left[\sum_{x_1^n} P(x_1^n|x_{-\infty}^0)\frac{P_n(x_1^n)}{P(x_1^n|x_{-\infty}^0)}\right] \\
&= 2^{-n\varepsilon}
\end{aligned}$$

which is summable. So the result follows by Borel-Cantelli. □

Lemma 0.36.

$$\limsup_{n\to\infty}\left[-\frac{1}{n}\log P_n(X_1^n) - \left[-\frac{1}{n}\log Q_n(X_1^n)\right]\right] \leq 0$$

almost-surely.

Proof. Let $\varepsilon > 0$. Then

$$\begin{aligned}
&\mathbb{P}\left(-\frac{1}{n}\log P_n(X_1^n) - \left[-\frac{1}{n}\log Q_n(X_1^n)\right] > \varepsilon\right) \\
&= \mathbb{P}\left(\frac{1}{n}\log\frac{Q_n(X_1^n)}{P_n(X_1^n)} > \varepsilon\right) \\
&\leq 2^{-n\varepsilon}\mathbb{E}\left(\frac{Q_n(X_1^n)}{P_n(X_1^n)}\right) \\
&= 2^{-n\varepsilon}\sum_{x_1^n} P_n(x_1^n)\frac{Q_n(X_1^n)}{P_n(X_1^n)} \\
&= 2^{-n\varepsilon}
\end{aligned}$$

and again since this is summable, the result follows. □

Method of Types

Suppose X_1^n are random variables on a finite alphabet $A = \{a_1, \ldots, a_m\}$. Let $\mathcal{P}$ denote the set of all PMFs on A, which we identify as a subset of $[0,1]^m$. The *type* of a string x_1^n is simply its empirical distribution $\hat{P}_n(a) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{X_i = a\}$. For each n, let $\mathcal{P}_n \subseteq \mathcal{P}$ denote the set of all n-types. Then

$$\mathcal{P}_n = \{P \in \mathcal{P} : P(a) = k/n \text{ for some } 0 \le k \le n, \text{ for all } a\}.$$

A (bad) bound is $|\mathcal{P}_n| \le (n+1)^m$.

Proposition 0.37. *If x_1^n has type $\hat{P}_n$ and Q is any PMF, then*

(a) $Q^n(x_1^n) = 2^{-n(H(\hat{P}_n)+D(\hat{P}_n\|Q))}$*;*

(b) $\hat{P}_n^n(x_1^n) = 2^{-nH(\hat{P}_n)}$.

Proof.

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

So we have (a), and (b) follows from taking $Q = \hat{P}_n$. $\square$

Definition. If $P \in \mathcal{P}_n$, the *type-class* of P is

$$T(P) = \{x_1^n \in A^n : x_1^n \text{ has type } P\}.$$

Note $|T(P)| = \binom{n}{nP(a_1),nP(a_2),\ldots,nP(a_m)} = \frac{n!}{(nP(a_1))!\ldots(nP(a_m))!}$.

Lemma 0.38. *If $P \in \mathcal{P}_n$, then*

$$\max_{P' \in \mathcal{P}_n} P^n(T(P')) = P^n(T(P)).$$

Proof. Note that for $P' \in \mathcal{P}_n$

$$\begin{aligned}
\frac{P^n(T(P))}{P^n(T(P'))} &= \frac{|T(P)| \prod_{j=1}^m P(a_j)^{nP(a_j)}}{|T(P')| \prod_{j=1}^m P(a_j)^{nP'(a_j)}} \\
&= \frac{\prod_{j=1}^m (nP'(a_j))!}{\prod_{j=1}^m (nP(a_j))!} \prod_{j=1}^m P(a_j)^{n(P(a_j)-P'(a_j))} \\
&\geq \prod_{j=1}^m (nP(a_j))^{n(P'(a_j)-P(a_j))} \prod_{j=1}^m P(a_j)^{n(P(a_j)-P'(a_j))} \\
&= n^{n\sum_{j=1}^m (P'(a_j)-P(a_j))} = 1
\end{aligned}$$

where we used that $\frac{k!}{\ell!} \geq \ell^{k-\ell}$. ☐

Proposition 0.39 (Size of type class)**.** *If $P \in \mathcal{P}_n$, then*

$$(n+1)^{-m} 2^{nH(p)} \leq |T(P)| \leq 2^{nH(P)}.$$

Proof. We have

$$1 \geq P^n(T(P)) = |T(P)| 2^{-nH(P)}$$

and

$$\begin{aligned}
1 = \sum_{x_1^n \in A^n} P^n(x_1^n) &= \sum_{P' \in \mathcal{P}_n} P^n(T(P')) \\
&\leq |\mathcal{P}_n| |T(P)| 2^{-nH(P)}.
\end{aligned}$$

☐

Proposition 0.40 (Probability of a type class)**.** *If $P \in \mathcal{P}_n$ and $Q \in \mathcal{P}$ then*

$$(n+1)^{-m} 2^{-nD(P\|Q)} \leq Q^n(T(P)) \leq 2^{-nD(P\|Q)}.$$

Proof. We have

$$Q^n(T(P)) = |T(P)| 2^{-n(H(P)+D(P\|Q))}$$

so using the previous proposition for bounding $|T(P)|$ we are done. ☐

Example. Suppose X_1^n are iid with distribution Q on A, and let $f : A \to \mathbb{R}$ be such that $\mu = \mathbb{E}[f(X)]$. Then, we look at

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n f(X_i) \geq \mu + \varepsilon\right)$$

for some $\varepsilon > 0$ with $\mu + \varepsilon < \max_{a \in A} f(a)$.

Writing $S_n = \sum_{i=1}^n f(X_i)$ we have

$$\begin{aligned}\mathbb{P}(S_n \geq n(\mu+\varepsilon)) &= \mathbb{P}\left(e^{\lambda S_n} \geq e^{\lambda n(\mu+\varepsilon)}\right) \\ &\leq e^{-n\lambda(\mu+\varepsilon)}\mathbb{E}[e^{\lambda S_n}] \\ &= e^{-n\lambda(\mu+\varepsilon)}(\mathbb{E}e^{\lambda f(X_1)})^n\end{aligned}$$

for any $\lambda > 0$. Hence

$$\mathbb{P}(S_n \geq n(\mu+\varepsilon)) \leq \exp\{-n[\lambda(\mu+\varepsilon) - \Lambda(\lambda)]\}$$

where $\Lambda(\lambda) = \log_e \mathbb{E}\left[e^{\lambda f(X_1)}\right]$. This gives the Chernoff bound

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n f(X_i) \geq \mu+\varepsilon\right) \leq e^{-n\Lambda^*(\mu+\varepsilon)}$$

where $\Lambda^*(x) = \sup_{\lambda>0}[\lambda x - \Lambda(\lambda)]$. But note that we also have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n f(X_i) &= \frac{1}{n}\sum_{i=1}^n \sum_{a\in A} \mathbb{1}\{x_i = a\}f(a) \\ &= \sum_{a\in A}\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{x_i = a\}f(a) \\ &= \sum_{a\in A}\hat{P}_n(a)f(a) \\ &= \mathbb{E}_{\hat{P}_n}(f(X))\end{aligned}$$

where $\hat{P}_n$ is the (random) type of X_1^n. So

$$\left\{\frac{1}{n}\sum_{i=1}^n f(X_i) \geq \mu+\varepsilon\right\} = \{\hat{P}_n \in E\}$$

where $E = \{P \in \mathcal{P} : \mathbb{E}_P f(X) \geq \mu+\varepsilon\}$.

Theorem 0.41 (Sanov's theorem)**.** *Suppose X_1^n are iid with distribution Q on a finite alphabet A, where Q has full support. Let $\hat{P}_n$ denote the (random) type of X_1^n. Then for any $E \subseteq \mathcal{P}$*

$$Q^n(\hat{P}_n \in E) \leq (n+1)^m 2^{-n \inf_{P \in E} D(P \| Q)}$$

so that in particular,

$$\limsup_{n \to \infty} \frac{1}{n} \log Q^n(\hat{P}_n \in E) \leq -\inf_{P \in E} D(P \| Q).$$

Moreover, if E is equal to the closure of its interior, then

$$\lim_{n \to \infty} \frac{1}{n} \log Q^n(\hat{P}_n \in E) = -D(P^* \| Q)$$

where P^* achieves $\inf_{P \in E} D(P \| Q)$*.*

Proof. We have

$$\begin{aligned}
Q^n(\hat{P}_n \in E) &= Q^n(\hat{P}_n \in E \cap \mathcal{P}_n) \\
&= \sum_{P \in E \cap \mathcal{P}_n} Q^n(T(P)) \\
&\leq |E \cap \mathcal{P}_n| \max_{P \in E \cap \mathcal{P}_n} 2^{-nD(P \| Q)} \\
&\leq |\mathcal{P}_n| \sup_{P \in E} 2^{-nD(P \| Q)} \\
&\leq (n+1)^m 2^{-n \inf_{P \in E} D(P \| Q)}.
\end{aligned}$$

For the lower bound note that since Q has full support, $D(P \| Q)$ is continuous in $P \in E$, so $P^* \in E$ exists. Also $\bigcup \mathcal{P}_n$ is dense in $\mathcal{P}$, and $\mathcal{P}_n$ eventually intersects every open subset of $\mathcal{P}$, so we can pick a sequence of PMFs (P_n) such that $P_n \in \mathcal{P}_n \cap E$ for all n and $P_n \to P^*$. Then

$$\begin{aligned}
Q^n(\hat{P}_N \in E) &= \sum_{P \in \mathcal{P}_n \cap E} Q^n(T(P)) \\
&\geq Q^n(T(P_n)) \geq 2^{-nD(P_n \| Q)}
\end{aligned}$$

so taking logs, dividing by n and taking $n \to \infty$ gives

$$\liminf_{n \to \infty} \frac{1}{n} \log Q^n(\hat{P}_n \in E) \geq -D(P^* \| Q).$$

□

Example. Let X_1^n be iid with distribution Q, $f : A \to \mathbb{R}$ and $\mu = \mathbb{E}[f(X_1)]$. Then by a Chernoff bound

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} f(X_i) \geq \mu + \varepsilon \right) \leq e^{-n\Lambda^*(\mu + \varepsilon)}$$

where $mu + \varepsilon < f^* = \max_{a\in A} f(a)$ and $\Lambda^*(x) = \sup_{\lambda>0}[\lambda x - \Lambda(\lambda)]$ for $x > 0$ and $\Lambda(\lambda) = \log_e \mathbb{E}(e^{\lambda f(X_1)})$ for $\lambda > 0$. But

$$\left\{\frac{1}{n}\sum_{i=1}^{n} f(X_i) \geq \mu + \varepsilon\right\} = \{\hat{P}_n \in \mathbb{E}\}$$

where $E = \{P \in \mathcal{P} : \mathbb{E}_P[f(X)] \geq \mu + \varepsilon\}$. So we have

$$\limsup_{n\to\infty} \frac{1}{n} \log_e Q^n(\hat{P}_n \in E) \leq -\Lambda^*(\mu + \varepsilon)$$

and

$$\liminf_{n\to\infty} \frac{1}{n} \log_e Q^n(\hat{P}_n \in E) \geq -D_e(P^*\|Q)$$

so $\Lambda^*(\mu + \varepsilon) \leq D_e(P^*\|Q)$.

In fact

Proposition 0.42. $\Lambda^*(\mu + \varepsilon) = D_e(P^*\|Q)$.

Proof. Let

$$P_\lambda(x) = \frac{e^{\lambda f(x)}Q(x)}{\mathbb{E}(e^{\lambda f(X_1)})},\ x \in A.$$

Then $\Lambda'(\lambda) = \frac{\mathbb{E}(f(X_1)e^{\lambda f(X_1)})}{\mathbb{E}(e^{\lambda f(X_1)})} = \mathbb{E}_{P_\lambda}[f(X)]$. Similarly it is easy to see $\Lambda''(\lambda) = \mathrm{Var}_{P_\lambda}(f(X)) \geq 0$. Therefore $\Lambda'(\lambda)$ increases from $\Lambda'(0+) = \mathbb{E}(f(X)) = \mu$ to $\Lambda'(+\infty)$. Note

$$\Lambda'(\lambda) = \frac{\sum_x Q(x)f(x)e^{\lambda f(x)}}{\sum_x Q(x)e^{\lambda f(x)}} \xrightarrow{\lambda\to\infty} f^*$$

so there exists $\lambda^* > 0$ such that $\Lambda'(\lambda^*) = \mu+\varepsilon = \mathbb{E}_{P_{\lambda^*}}[f(X_1)]$. Also $\Lambda^*(\mu+\varepsilon) = \lambda^*(\mu + \varepsilon) - \Lambda(\lambda^*)$. Hence $P_{\lambda^*} \in E$, so

$$\begin{aligned} D_e(P^*\|Q) \leq D_e(P_{\lambda^*}\|Q) &= \sum_x P_{\lambda^*}(x) \log_e \frac{P_{\lambda^*}(x)}{Q(x)} \\ &= \sum_x P_{\lambda^*}(x) \log \frac{e^{\lambda^* f(x)}}{\mathbb{E}(e^{\lambda^* f(X_1)})}. \end{aligned}$$

So $D_e(P^*\|Q) \leq \lambda^* bbE_{P^*}[f(X)] - \log_e \mathbb{E}[e^{\lambda^* f(X_1)}] = \lambda^*(\mu+\varepsilon) - \Lambda(\lambda^*) = \Lambda^*(\mu+\varepsilon)$. ☐

Note. If E is closed and convex, $P^* \in E$ exists and is unique, since $D(\cdot\|Q)$ is strictly convex.

Theorem 0.43 (Pythagorean identity)**.** *Suppose $E \subseteq \mathcal{P}$ is closed and convex, Q has full support annd let P^* achieve* $\inf_{P\in E} D(P\|Q)$*. Then for any $P \in E$*

$$D(P\|Q) \geq D(P\|P^*) + D(P^*\|Q).$$

Proof. Let $P \in E$ and define $P_\lambda = \lambda P + (1-\lambda)P^* \in E$ for $\lambda \in [0,1]$. Since $P_\lambda|_{\lambda=0} = P^*$ and P^* achieves the inequality, we must have

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} D_e(P_\lambda \| Q)|_{\lambda=0^+} \geq 0$$

so

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\lambda} \sum_x P_\lambda(x) \log_e \frac{P_\lambda(x)}{Q(x)}\Bigg|_{\lambda=0^+} \\
&= \sum_x (P(x) - P^*(x)) \log_e \frac{P_\lambda(x)}{Q(x)}\Bigg|_{\lambda=0^+} + \underbrace{\sum_x P_\lambda(x) \frac{Q(x)}{P_\lambda(x)} \frac{P(x) - P^*(x)}{Q(x)}\Bigg|_{\lambda=0^+}}_{=0} \\
&= \sum_x P(x) \log_e \left(\frac{P^*(x)}{Q(x)} \frac{P(x)}{P(x)}\right) - \sum_x P^*(x) \log_e \frac{P^*(x)}{Q(x)} \\
&= D_e(P\|Q) - D_e(P\|P^*) - D_e(P^*\|Q) \geq 0.
\end{aligned}$$

$\square$