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

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

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

(ii) $D(P\|Q) \ge 0$ 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*

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

  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 $e_1^{(n)} \ge 2^{-n(D+\varepsilon+\frac{1}{n})}$ 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*

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

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 $i \in [n]$. □

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. □

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 $\Delta(p, q) = 2D_e(P\|Q) - \|P - Q\|^2_{TV}$. 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

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

□

Poisson Appoximation

Suppose $X_1, \ldots, X_n \sim \text{Bern}(\lambda/n)$ 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 $X_1, \ldots, X_n \sim \text{Bern}(p_i)$ 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

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

□

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.

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

□