Information Theory

Jan Olucha Fuentes

May 22, 2025

Dear Reader,

This is a set of lecture notes typed for the course *Information Theory* taught at the University of Cambridge during the academic year 2024-2025.
Yours falsely,

JOF.

Contents

1 Introduction: What is entropy? 4

1.1 The Asymptotic Equipartition Property 4

1.2 Fixed Rate Codes 8

2 Relative Entropy 11

3 Properties of Entropy and relative entropy 18

3.0.1 Properties of Entropy 18

3.1 Examples 23

3.2 Properties of relative entropy 25

4 Poisson approximation 28

4.1 Poisson distribution as a maximal entropy result 30

5 Mutual Information 33

5.1 Synergy 36

6 Entropy and additive combinatorics 39

7 Entropy rates of a stochastic process 43

7.1 Shannon-McMillan-Breiman Theorem 45

8 Types and large deviations 46

8.1 Method of types 46

8.2 Large Deviations 52

8.2.1 Sanov's Theorem 53

8.3 The Gibb's conditioning principle 60

8.4 Error probability in fixed-rate data compression 66

9 Variable rate Lossless data compression 71

9.1 Code-distribution correspondence 74

9.2 Shannon codes and properties 77

A Some nice results 79

Notation and how to study these notes

Difficulty notations:

☕: denotes a proof or idea that is hard or hard to reproduce.

♪: denotes a proof or idea that requires partially unmotivated tools or machinery that is hard to pull out of thin air, but the rest of the argument is easy.

♩: denotes a proof or idea that is easy to see knowing some other results that help motivate it.

♫: denotes a proof or idea that is easy and no particularly clever ideas are needed. For revision purposes reading it a couple times will suffice.

Other notation:

$x_1^n = (x_1, \cdots, x_n) \in \{0,1\}^n$: a message consisting of n bits x_i.

$\mathscr{A}$: an alphabet, finite unless otherwise specified.

$X_i^j = (X_i, X_{i+1}, \cdots, X_j)$: a block of random variables with $i \le j$, of length $j - i + 1$.

$0 \log 0$: a weird way of writing 0.

$\{0,1\}^*$: the set of all finite-length binary strings, i.e: $\{0,1\}^* = \bigcup_{n \ge 0} \{0,1\}^n$.

$\{X_i\}_\perp$: a collection of independent random variables.

$\mathbf{E}_{X \sim Q}[f(X)]$: means the expectation of $f(X)$ where $X \sim Q$. Also written as $\int_\Omega f(x) dQ(x)$

1 Introduction: What is entropy?

1.1 The Asymptotic Equipartition Property

Learning objectives:

- Definition of entropy.
- The AEP, which states the following: given a memoryless source, the probability of any string generated by the source becomes asymptotically uniform on the set of typical strings, which also carries all the mass. Conversely, any set that carries all the mass must be just as large.

Definition 1.1 A sequence of random variables $\mathbf{X} = (X_n : n \geq 1)$ with values on an alphabet $\mathscr{A}$ is called a source. If the random variables X_i are iid, the source is said to be memoryless.

Definition 1.2 Let X be a random variable taking values on $\mathscr{A}$, and let P denote its probability mass function. The entropy of X, $H(X)$ is defined as

$$H(X) := -\sum_{x \in \mathscr{A}} P(x) \log P(x) = \mathbf{E}[-\log P(X)]$$

Remark 1.3 Note that $P(X)$ is a random variable, simply defined by $P(X)(\omega) = \mathbf{P}(X = \omega)$. We will denote $P^n(X_1^n)$ for the joint PMF of a message X_1^n.

Theorem 1.4 (Shannon's computation ♫) Let X_1^n be a block of iid RVs with distribution P on a finite alphabet $\mathscr{A}$. Then

$$-n^{-1} \log\left(P^n(X_1^n)\right) \longrightarrow H$$

In probability. Where $H = H(X_i)$ is the common entropy.

Proof. Since the random variables are iid, we have

$$P^n(X_1^n) = \prod_{i=1}^{n} P(X_i) = 2^{\log\left(\prod_{i=1}^{n} P(X_i)\right)} = 2^{\sum_{i=1}^{n} \log(P(X_i))}$$

Therefore

$$-n^{-1}\log\big(P^n(X_1^n)\big) = n^{-1}\sum_{i=1}^{n} -\log P(X_i) \longrightarrow \mathbf{E}[-\log P(X_1)] = H$$

in probability by the WLLN. ♡

Remark 1.5 We have the following:

- $H(X) \in [0,\infty]$ because $-\log P(X) \geq 0$.
- $H(X)$ does not depend on the values of X, only on the probabilities with which it takes said values.
- $H(X) = H(f(X))$ for any injective function $f : \mathscr{A} \longrightarrow \mathscr{A}$ for the same reason as the second bulletpoint.
- $H(X) \geq 0$ with equality if and only if X is deterministic almost surely.

Informally speaking, what this Theorem tells us is that with a very high probability, the probability of a *typical long message* $P^n(X_1^n)$ is approximately $2^{-nH(X_1)}$. Or more precisely, given an error tolerance ϵ

$$\mathbf{P}\left(\underbrace{2^{-n(H+\epsilon)} \leq P^n(X_1^n) \leq 2^{-n(H-\epsilon)}}_{B_n^*(\epsilon)}\right) \longrightarrow 1$$

Definition 1.6 (Typical messages) We refer to the set $B_n^*(\epsilon)$ as the set of typical messages:

$$B_n^*(\epsilon) = \left\{x_1^n \in \mathscr{A}^n : 2^{-n(H+\epsilon)} \leq P^n(x_1^n) \leq 2^{-n(H-\epsilon)}\right\}$$

Corollary 1.7 (Asymptotic Equipartition Property ♫) Let $(X_n : n \geq 0)$ be iid on a finite alphabet $\mathscr{A}$, with common distribution P and common entropy $H = H(X_i)$. Then for any $\epsilon > 0$, we have

- $|B_n^*(\epsilon)| \leq 2^{n(H+\epsilon)}$ for all $n \geq 1$.
- $\mathbf{P}(X_1^n \in B_n^*(\epsilon)) \longrightarrow 1$.

And *conversely*, if (B_n) is a sequence of subsets of A^n for $n \geq 1$ with the property that $\mathbf{P}(X_1^n \in$

$B_n) \longrightarrow 1$ as $n \longrightarrow \infty$, then for any $\epsilon > 0$, eventually:

$$|B_n| \geq (1-\epsilon)2^{n(H-\epsilon)}$$

Main idea: To prove the convergence of the direct part is immediate from the construction before. To prove the size bound of the direct part we use the standard size counting trick. To prove the size bound of the converse part, we observe that since both (B_n) and (B_n^*) have that $\mathbf{P}(X_1^n \in B_n) \longrightarrow 1$ and $\mathbf{P}(X_1^n \in B_n^*(\epsilon)) \longrightarrow 1$ then it will be the case that $\mathbf{P}(X_1^n \in B_n \cap B_n^*(\epsilon)) \longrightarrow 1$. From here we can use a standard size bound trick.

Proof of Corollary 1.7. The fact that $\mathbf{P}(X_1^n \in B_n^*(\epsilon)) \longrightarrow 1$ as $n \longrightarrow \infty$ comes for free from the construction of $B_n^*(\epsilon)$ and Theorem 1.4. Indeed: by the Weak Law of Large numbers we have that for all $\epsilon > 0$:

$$\mathbf{P}\left(\left|-\frac{1}{n}\log P^n(X_1^n) - H\right| > \epsilon\right) \longrightarrow 0$$

So by rearranging:

$$\mathbf{P}\left(2^{-n(H+\epsilon)} < P^n(X_1^n) < 2^{-n(H-\epsilon)}\right) \longrightarrow 1$$

And the set on the inside is the set of typical strings.

To show the bound on the size, we note that

$$1 \geq \mathbf{P}(X_1^n \in B_n^*(\epsilon)) = \sum_{x_1^n \in B_n^*(\epsilon)} P^n(x_1^n) \geq |B_n^*(\epsilon)| 2^{-n(H+\epsilon)}$$

where in the last inequality we have used the defining property of $B_n^*(\epsilon)$. To show the *converse*, suppose $(B_n) \subseteq A^n$ is a sequence of sets of messages with the desired property, i.e: that all the mass is asymptotically concentrated on them. Then we start by noting that

$$P^n(B_n \cap B_n^*(\epsilon)) = P^n(B_n) + P^n(B_n^*(\epsilon)) - P^n(B_n \cup B_n^*(\epsilon)) \geq P^n(B_n) + P^n(B_n^*(\epsilon)) - 1$$

Since this right hand side converges to 1 (by hypothesis and B_n^* being the set of typical messages), we have that

$$P^n(B_n \cap B_n^*(\epsilon)) \longrightarrow 1.$$

Therefore, for any $\epsilon > 0$ it will be the case that eventually

$$\begin{aligned}(1-\epsilon) &\le P^n(B_n \cap B_n^*(\epsilon)) \\ &= \sum_{x_1^n \in B_n \cap B_n^*(\epsilon)} \underbrace{P^n(x_1^n)}_{\in B_n^*(\epsilon)} \\ &\le \left|B_n \cap B_n^*(\epsilon)\right| 2^{-n(H-\epsilon)} \\ &\le |B_n| 2^{-n(H-\epsilon)}\end{aligned}$$

Rearranging gives the claim. ♡

Remark 1.8 There is something to say about the indiscriminate use of the notation $P^n(A)$ that we are using. Here, by P^n we mean the *measure induced by* X_1^n, so whenever we say $P^n(A)$ what we really mean is $\mathbf{P}(X_1^n \in A)$

Let us make a few comments on the AEP. First note the format of the result. We first show the existence of an object with a certain quantitative performance, and then show that no other object can outperform it. This result gives an interpretation of the meaning of entropy. Indeed: it tells us that the smallest set of strings that carry almost all the probability on $\mathscr{A}^n$ has a size of approximately 2^{nH} strings. As such, a higher entropy means a higher number of typical messages, which means that RVs with high entropy are "more random".

1.2 Fixed Rate Codes

Learning objectives: In this section we study the fundamental limitations of fixed-rate compression. In particular we learn:

- Definition of fixed-rate codes, their error probability and their rate, which quantifies how good they perform.
- The fixed-rate coding Theorem, which states that one can always achieve a performance at least as good as the entropy, but actually one can't do any better.

Definition 1.9 (Fixed Rate Codes) Let $\mathbf{X} = (X_n)$ be a memoryless source with distribution P. A fixed rate lossless compression code for $\mathbf{X}$ is a sequence of *codebooks* (B_n) with $B_n \subseteq \mathscr{A}^n$.

Let us explain what we mean by a codebook. Suppose we are trying to compress a specific message x_1^n. What we could do is write a *codebook*, i.e: a list of all messages we are interested in compressing, call it B_n. Suppose both sender and receiver agree on B_n and the ordering of the messages in B_n. Then if Alice wants to send x_1^n to Bob, she first checks whether $x_1^n \in B_n$, and if it is, she describes the message to Bob by writing a *flag*, say 1, followed by the index of x_1^n in B_n. This takes

$$1 + \lceil \log_2(|B_n|) \rceil \quad \text{bits}$$

If the message $x_1^n \notin B_n$, then she indicates a 0 flag, and sends x_1^n uncompressed. This takes

$$1 + \lceil \log_2(|\mathscr{A}^n|) \rceil = 1 + \lceil n \log_2(|\mathscr{A}|) \rceil \quad \text{bits}$$

The interplay here is that you want to make B_n large to avoid having to send an entire uncompressed message, but then you don't want to make B_n too large, because then otherwise this whole business becomes redundant.

Definition 1.10 (Rate and error probability) The rate of a fixed-rate code (B_n) for a source $\mathbf{X}$ is defined as

$$R_n = n^{-1}\left(1 + \lceil \log_2(|B_n|) \rceil\right)$$

The probability of error is defined as $P_e^{(n)} = \mathbf{P}(X_1^n \notin B_n)$

The rate captures how many bits of information you need to transmit one symbol through the fixed-rate codes. A lower rate implies a better compression. The probability of error is pretty self-explanatory. The question we may ask now is: if we require the probability of error to be negligible asymptotically, what is the best (i.e: the lowest) rate we can achieve? It turns out that entropy is the answer to this problem.

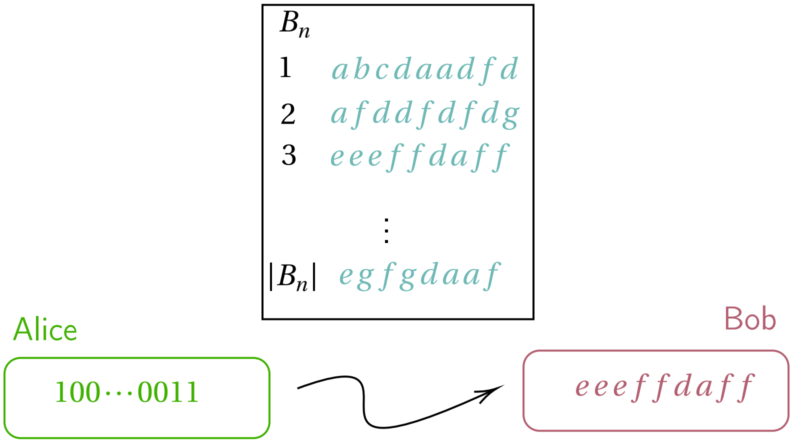


Figure 1: Codebook illustration

Theorem 1.11 (Fixed Rate Coding Theorem ♩) Let $\mathbf{X} = (X_n)$ be a memoryless source with distribution P and entropy $H(X_i) = H$. Then

- For any $\epsilon > 0$, there exists a fixed-rate codebook (B_n^*) with vanishing error probability and with rate
$$R_n \leq H + \epsilon + \frac{2}{n} \quad \forall n \geq 1$$
- If a fixed-rate codebook (B_n) has vanishing probability, then for any $\epsilon > 0$, its rate must eventually satisfy
$$R_n > H - \epsilon$$

Main idea: The direct part is a corollary of the AEP: Indeed, the hinted codebook is simply the set of typical messages, and to prove the bound on the rate we use the size upper bound given by the AEP. The converse follows by using the lower bound given by the AEP on the size of a set that carries asymptotically all mass.

Proof. For the first part, we take the codebook consisting of typical messages $(B_n^*(\epsilon))$, we already know that this set accumulates all the mass of the messages, so $\mathbf{P}(X_1^n \in B_n^*(\epsilon)) \longrightarrow 1$, showing $P_e^{(n)} \longrightarrow 0$. We also know that $|B_n^*(\epsilon)| \leq 2^{n(H+\epsilon)}$ so it follows that the rate of this code is

$$R_n \leq \frac{1}{n}(1 + \underbrace{\log|B_n^*| + 1}_{\geq\lceil\log|B_n^*\rceil}) \leq H + \epsilon + \frac{2}{n}$$

Conversely, take $\epsilon > 0$, without loss of generality assume $\epsilon < 1/2$. Then

$$\begin{aligned} R_n &:= \frac{1}{n}\lceil \log|B_n| \rceil + \frac{1}{n} \\ &\geq \frac{1}{n}\log|B_n| + \frac{1}{n} \\ &\overset{(!)}{\geq} \frac{1}{n}\log(1-\epsilon) + (H-\epsilon) + \frac{1}{n} \\ &= \frac{1}{n}\log(2(1-\epsilon)) + H - \epsilon \\ &> H - \epsilon \end{aligned}$$

Where step (!) comes from the fact that since by assumption $|B_n|$ carries all the probability, we then know from the AEP, (Theorem 1.7) that $|B_n| \geq (1-\epsilon)2^{n(H-\epsilon)}$ so plugging in gives the result. I should also mention just in case I have lost neurons when I read this in the future that the equality after (!) comes from writing $1 \equiv \log 2$. ♡

2 Relative Entropy

Learning Objectives: in this section we learn about

- Relative entropy ♫: its definition and interpretation as a "distance" between probability distributions.
- Stein's Lemma ♩: which states that in a hypothesis test, the best asymptotic performance a simple hypothesis test can achieve is described by relative entropy.
- Neymann-Pearson Lemma: gives the optimal decision regions for a hypothesis test for a finite sample size.
- An information theoretic interpretation of Neymann-Pearson Lemma.

In this section we argue that relative entropy is the answer to the question: how statistically indistinguishable are two probability distributions? Suppose $x_1^n \in \mathscr{A}^n$ are observations generated by a memoryless source X_1^n and we want to decide if $X_1^n \sim P^n$ or $X_1^n \sim Q^n$. At its most fundamental level, hypothesis test is described by a decision region $B_n \subseteq \mathscr{A}^n$. If our observation $x_1^n \in B_n$ then we declare the outcome of our test to be that $X_1^n \sim P^n$. Otherwise we declare the other distribution to be the data generating mechanism. There are two associated error probabilities

Definition 2.1 (Type I and Type II error probabilities) We distinguish the probabilities

$$e_1^{(n)} = \mathbf{P}_Q(x_1^n \in B_n) = Q^n(B_n) \qquad e_2^{(n)} = \mathbf{P}_P(x_1^n \notin B_n) = P^n(B_n^c)$$

That is: the probability that we declared our data as coming from P when it came form Q, and the probability of declaring it came from Q when it came from P.

There is of course a fundamental problem in choosing our set B_n, if we try to say minimise our Type I error, by making B_n small, then we are making B_n^c big, thus potentially increasing our Type II error probability.

Definition 2.2 (Relative Entropy) The relative entropy between two PMFs P and Q on the same discrete alphabet $\mathscr{A}$ is

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

The question we aim to answer now is: if we require $e_2^{(n)} \longrightarrow 0$, how small can we make $e_1^{(n)}$? It turns out that the best we can do is $e_1^{(n)} \approx 2^{-nD(P\|Q)}$. Thus showing once again how the larger the relative entropy between the two random variables, the better we will be able to tell them apart. The following

Theorem should be seen as a version of the AEP for relative entropies. I encourage the reader to think deeply as to why the proof Stein's Lemma is morally the same as the proof of the AEP.

Theorem 2.3 (Stein's Lemma ♩) Let P and Q be two distributions on a finite alphabet $\mathscr{A}$, with $0 < \underbrace{D(P\|Q)}_{D} < \infty$. Then

- For all $\epsilon > 0$, there are decision regions B_n^* such that $e_1^{(n)} \leq 2^{-n(D-\epsilon)}$ for all n, and $e_2^{(n)} \longrightarrow 0$.
- If $\{B_n\}$ is a family of decision regions with $e_2^{(n)} \longrightarrow 0$, then for all $\epsilon > 0$, eventually

$$e_1^{(n)} \geq 2^{-n\left(D+\epsilon+\frac{1}{n}\right)}$$

Main idea: We will essentially replicate the proof of the AEP. In the proof of the AEP, we talked about a special region, which was defined as the region of probability-typical strings, because this region was connected to entropy. This region was defined in terms of an upper and a lower bound on the probability of a string, and this gave us the desired size and rate bounds. In this case, we will talk about the region of log-likelihood-typical strings, because log-likelihood is connected to relative entropy. Appropriate upper and lower bounds on the definition of this set of strings will also give us the bounds we want.

Proof. We first form our decision region. To do this, we note that we have the following convergence in probability:

$$\frac{1}{n}\log\left(\frac{P^n(X_1^n)}{Q^n(X_1^n)}\right) = \frac{1}{n}\log\left(\prod_{i=1}^{n}\frac{P(X_i)}{Q(X_i)}\right) = \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{P(X_i)}{Q(X_i)}\right) \longrightarrow \mathbf{E}_P\left[\log\frac{P(X)}{Q(X)}\right] =: D(P\|Q)$$

Formally, for any $\epsilon > 0$, we have that

$$\mathbf{P}_P\left(\left|\frac{1}{n}\log\frac{P^n(X_1^n)}{Q^n(X_1^n)} - D(P\|Q)\right| > \epsilon\right) \longrightarrow 0$$

This motivates the creation of our decision region, to be the set of "log-likelihood-ratio" strings,

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

By the above computation, (i.e: by the WLLN), we have that our Type II error probability, $e_2^{(n)} = \mathbf{P}_P(\mathscr{A}^n \setminus B_n^*) \longrightarrow 0$, thus we have satisfied our asymptotically negligible error probability. We now show the desired bound on the decay of our Type I error probability:

$$\begin{aligned} e_1^{(n)} &:= \mathbf{P}_Q(X_1^n \in B_n^*(\epsilon)) \\ &:= Q^n(B_n^*(\varepsilon)) \\ &= \sum_{x_1^n \in B_n^*(\epsilon)} Q^n(x_1^n) \\ &\overset{(!)}{\leq} \sum_{x_1^n \in B_n^*(\epsilon)} 2^{-n(D-\epsilon)} P^n(x_1^n) \leq 2^{-n(D-\epsilon)} \end{aligned}$$

Where the only non-trivial step, comes from the fact that by definition of $B_n^*(\epsilon)$, we have that whenever $x_1^n \in B_n^*(\epsilon)$, then $Q^n(x_1^n) \leq 2^{n(D-\epsilon)} P^n(x_1^n)$. For the converse, we begin by assuming that a decision region B_n has that $e_2^{(n)} \longrightarrow 0$, in other words,

$$\mathbf{P}_P(B_n^c) =: P^n(B_n^c) \longrightarrow 0$$

Since by construction of B_n^* we also have that $P^n(B_n^{*c}) \longrightarrow 0$, we obtain that

$$\begin{aligned} P^n(B_n \cap B_n^*) &= 1 - P^n(B_n^c \cup B_n^{*c}) \\ &\overset{(1)}{\geq} 1 - P^n(B_n^c) - P^n(B_n^{*c}) \longrightarrow 1 \end{aligned}$$

Where the highlighted step comes from a trivial union bound. In particular, since it converges to 1, we must have that eventually, $P^n(B_n \cap B_n^*) \geq \frac{1}{2}$, which now allows us to compute the desired bound on $e_1^{(n)}$:

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

Where step ☺is motivated because we want to talk about $e_1^{(n)} := Q^n(B_n)$, and step (!) comes from the construction of $B_n^*(\epsilon)$, which gives us the upper bound on the likelihood ratio. Rearranging gives the desired result. ♡

Remark 2.4 As usual, here we state that we can make our error probability as good as approximately 2^{-nD}, and then we are showing that actually you can't do any better than that. There are some other comments to make

- Why is D assumed to be in $(0, \infty)$? - If $D = 0$ then they are the same distribution, so nothing to check. If $D = \infty$, then they have disjoint support, that is to say for example, there is some $x_0 \in \mathscr{A}$ with $0 = P(x_0) \neq Q(x_0)$. Thus if we observe some x_0 come up, we determine the source to be Q.

- This Lemma states that if one of the error probabilities is assumed to vanish, then the best decay of the other error probability can vanish is exponential with rate D. Thus the larger D, intuitively reflected by both distributions being "further away", the faster the decay and the easier it is to tell both distributions apart.

- The decision regions B_n^* provided in Stein's Lemma are asymptotically optimal, not necessarily optimal for finite n. The exact form of those test are given by the Neymann-Pearson Lemma.

Theorem 2.5 (Neymann-Pearson Lemma ♪) For a hypothesis test between two distributions P and Q, based on n data samples, then for $T \geq 0$, the decision regions

$$B_{\mathsf{NP}}(T) = \left\{ x_1^n \in \mathscr{A}^n : \frac{P^n(x_1^n)}{Q^n(x_1^n)} \geq T \right\}$$

are are optimal, that is to say, if any other decision region B_n has that $e_2^{(n)}(B_n) \leq e_2^{(n)}(B_{\mathsf{NP}}(T))$, then it must be that $e_1^{(n)}(B_n) \geq e_1^{(n)}(B_{\mathsf{NP}}(T))$.

Main idea: The proof follows immediately after considering the quantity

$$\left[\mathbf{1}(B_{\mathsf{NP}})(x_1^n) - \mathbf{1}(B_n)(x_1^n)\right]\left[P^n(x_1^n) - TQ^n(x_1^n)\right]$$

Proof. For $x_1^n \in \mathscr{A}^n$, consider the quantity

$$\left[\mathbf{1}(B_{\mathsf{NP}})(x_1^n) - \mathbf{1}(B_n)(x_1^n)\right]\left[P^n(x_1^n) - TQ^n(x_1^n)\right] \overset{(!)}{\geq} 0$$

The key to the proof is that this quantity is greater than or equal to zero. To see this, we note that if the right hand side were to be strictly less than zero, then $x_1^n \notin B_{\mathsf{NP}}(T)$, and so the left hand size is less than or equal to zero. If the right hand side is however greater than or equal to zero, then the left hand side is also greater than or equal to zero. With this in mind, we may multiply the expression out, and if we sum over all $x_1^n \in \mathscr{A}^n$, the indicators will restrict each sum to happen on the specified regions which means that we have

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

Rewriting this in terms of $e_1^{(n)}$ and $e_2^{(n)}$ we have that

$$T\left(e_1^{(n)}(B_n) - e_1^{(n)}(B_{\text{NP}})\right) - \left(e_2^{(n)}(B_{\text{NP}}) - e_2^{(n)}(B_n)\right) \geq 0$$

And now we may carry out a similar observation as before. If the region B_n has a performance at least as good as B_{NP}, i.e: $e_1^{(n)}(B_n) \leq e_1^{(n)}(B_{\text{NP}})$, so that the first summand is negative, it must also be that the second summand is negative, which means that $e_2^{(n)}(B_{\text{NP}}) \leq e_2^{(n)}(B_n)$, establishing the claim. ♡

As it turns out, the Neymann-Pearson Lemma also has an information theoretic interpretation. To define it we first need to take a look at the empirical distribution function of our sample data $X_1, \cdots, X_n$

$$\widehat{P_{X_1^n}}(a) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i = a)$$

This function simply counts the proportion of our sample data that attains this value a. Of course this is a random quantity as currently formulated, but we can also use this idea to induce a probability distribution on $\mathscr{A}$ given some observation $x_1^n \in \mathscr{A}^n$. Simply by replacing the X_i with x_i. We are now ready to state the alternative interpretation

Proposition 2.6 (Information Theoretic Neymann-Pearson Lemma ♪) The Neymann-Pearson decision region B_{NP} is equivalent to

$$B_{\text{NP}} = \left\{x_1^n \in \mathscr{A}^n : D(\widehat{P_n} \| Q) > D(\widehat{P_n} \| P) + T'\right\}$$

where T' corresponds to $\frac{1}{n}\log T$ with T as in the Neymann-Pearson Lemma

Main idea: Stare at the definition of B_{NP}, stare at this claim. We have to show that we can express log-likelihood-ratio as a difference of "distances". More precisely, the idea is to show that the log-likelihood is large if $\widehat{P}_{x_1^n}$ is closer to P than to Q:

$$\log \frac{P^n(x_1^n)}{Q^n(x_1^n)} = n\left[D(\widehat{P}_{x_1^n} \| Q) - D(\widehat{P}_{x_1^n} \| P)\right]$$

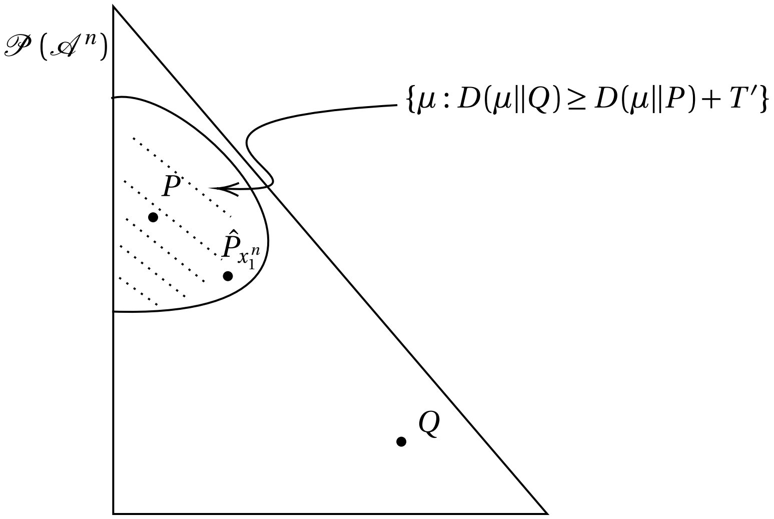

$\mathscr{P}(\mathscr{A}^n)$
$\{\mu : D(\mu\|Q) \geq D(\mu\|P) + T'\}$
P
$\widehat{P}_{x_1^n}$
Q

Figure 2: Log-likelihood ratio admits a geometric interpretation: normalised log-likelihood ratio is how much closer the empirical distribution is from P than from Q.

Proof. We have the following computation

$$\begin{aligned}\log \frac{P^n(x_1^n)}{Q^n(x_1^n)} &= \log \prod_{i=1}^n \frac{P(x_i)}{Q(x_i)} \\ &= \sum_{i=1}^n \log \frac{P(x_i)}{Q(x_i)} \\ &= \sum_{i=1}^n \sum_{a \in \mathscr{A}^n} \mathbf{1}(\{x_i\})(a) \log \frac{P(a)}{Q(a)} \\ &= n \sum_{a \in \mathscr{A}^n} \left(\frac{1}{n} \sum_{i=1}^n \mathbf{1}(\{x_1^n\})(a) \right) \log \frac{P(a)\widehat{P}_{x_1^n}(a)}{Q(a)\widehat{P}_{x_1^n}(a)} \\ &= n \left[D(\widehat{P}_{x_1^n}\|Q) - D(\widehat{P}_{x_1^n}\|P) \right]\end{aligned}$$

Now we simply note that

$$\frac{P(x_1^n)}{Q(x_1^n)} \geq T \iff \frac{1}{n} \log \frac{P(x_1^n)}{Q(x_1^n)} \geq \frac{1}{n} \log T$$

which finishes the claim. ♡

This gives a very elegant interpretation: our decision region of acceptance of P is the set of datapoints whose empirical distribution $\widehat{P_n}$ is closer to P that to Q by some slack tolerance. We finish this section by talking about some elementary bounds on the entropy.

Theorem 2.7 (Bounds on Entropy ♪) Let X be a random variable taking values in $\mathscr{A}$. Then

1. $0 \le H(X) \le \log(|\mathscr{A}|)$. With equality at 0 if and only if X is constant, and equality at $\log(|\mathscr{A}|)$ if and only if X is uniform on A.
2. If P and Q are two PMFs on $\mathscr{A}$ then

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

   with equality if and only if P and Q are the same distribution.

Main idea: It's better to prove first the second part of the claim. This follows immediately by the log-sum inequality. To prove the first claim, use $Q = \text{Unif}(A)$ and compute $D(P\|Q)$.

Proof. We begin by proving the second part of the claim, but by the log-sum inequality:

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

And we have equality if and only if for each a, $P(a) = Q(a)$.
We now prove the second part, we consider the distribution $Q = \text{Unif}(A)$, i.e: $Q(a) = 1/|A|$ for all a. Then

$$D(P\|Q) = \sum_{a\in\mathscr{A}^n} P(a)\log\frac{P(a)}{1/|A|} = \log|A| - H(X)$$

Then we get that

$$H(X) = \log|A| - D(P\|Q) \le \log|A|$$

which occurs with equality if and only if P and Q are the same distribution, i.e: P is uniform. ♡

This solidifies our intuition of using entropy as a measure of how much we can compress a random variable. Indeed, we saw that using fixed-rate codes, the best rate we could achieve in terms of bits needed per symbol was $H(X)$, so if X is deterministic, then there is no need to compress anything, and if X is uniformly random, then we cannot take advantage of any structure in X to compress the information, and we need the maximum amount of $\log(|\mathscr{A}|)$ bits to express our message. In terms of the relative entropy, this result solidifies our intuition as to why it is a (pseudo) distance between probability distributions.

3 Properties of Entropy and relative entropy

3.0.1 Properties of Entropy

In this section we learn:

- Definition of joint and conditional entropies.
- The chain rule for joint (and conditional) entropy (example sheet).
- Giving information cannot increase entropy.
- Data processing inequalities, for entropy and relative entropy.
- Fano's inequality.

We have so far talked about $H(X)$ for a single discrete random variable X, but there is nothing morally stopping us from thinking of a random vector $(X_1, \cdots, X_n)$ of discrete random variables as a single discrete random variables taking values in a Cartesian product of countable sets.

Definition 3.1 Let (X, Y) be a discrete random vector taking values in $\mathscr{A} \times \mathscr{B}$ with joint PMF $P_{XY}(x, y)$. The joint entropy of X and Y is defined as

$$H(X, Y) = -\sum_{(x,y)\in\mathscr{A}\times\mathscr{B}} P_{XY}(x, y) \log P_{XY}(x, y) = \mathbf{E}_{XY}[-\log P_{XY}(X, Y)]$$

Above I denote by $\mathbf{E}_{XY}$ an expectation taken with respect to the probability measure P_{XY} of course. Let us examine some preliminary ideas relating to joint entropies:

Proposition 3.2 (Joint entropy of independent random variables, ♫) Let X, Y be as above. Then if X and Y are independent, then

$$H(X, Y) = H(X) + H(Y)$$

Main idea: Use definition of independence.

Proof. Elementary calculation

$$H(X, Y) = \mathbf{E}[-\log P_{XY}(X, Y)] = \mathbf{E}[-\log P_X(X)] + \mathbf{E}[-\log P_Y(Y)]$$

♡

If we don't have independence however, we cannot write the result above, but we can achieve something similar.

Definition 3.3 (Conditional entropy) Let X and Y be random variables. The conditional entropy $H(Y\mid X)$ is defined as

$$H(Y\mid X)=\mathbf{E}_{XY}[-\log P_{Y|X}(Y\mid X)]$$

Remark 3.4 The conditional entropy is not an entropy as we have defined it above so far, but can be interpreted as an average of entropies. Indeed:

$$\begin{aligned}H(X\mid Y)&=-\sum_{(x,y)\in\mathscr{A}\times\mathscr{B}}P_{XY}(x,y)\log P_{X|Y}(x\mid y)\\&=-\sum_{(x,y)\in\mathscr{A}\times\mathscr{B}}P_{X|Y}(x\mid y)P_Y(y)\log P_{X|Y}(x\mid y)\\&=\sum_{y\in\mathscr{B}}P_Y(y)\underbrace{\mathbf{E}_{X|Y=y}[-\log P_{X|Y=y}]}_{H(X|Y=y)}\end{aligned}$$

And $H(Y\mid X=x)$ does make sense as an entropy! Because $P_{Y|X=x}$ is a probability measure on $\mathscr{B}$.

The point is that this definition leads to a nice interpretation

Theorem 3.5 (Chain rule for entropy, ♫) Let X and Y be two random variables. Then

$$H(X,Y)=H(X)+H(Y\mid X)$$

Main idea: Use definition of conditional probability.

Proof.

$$\begin{aligned}H(X,Y)&=\mathbf{E}[-\log P_{XY}(X,Y)]\\&=\mathbf{E}[-\log P_X(X)]+\underbrace{\mathbf{E}[-\log P_{Y|X}(Y\mid X)]}_{H(Y|X)}\end{aligned}$$

and now noting that all these expectations are with respect to P_{XY}, we simply observe that

$$-\mathbf{E}[\log P_X(X)]=\sum_x\log P_X(x)\underbrace{\sum_y P_{XY}(x,y)}_{P_X(x)}=H(X)$$

♡

The chain rule for entropy may be generalised as follows

Theorem 3.6 (Chain rule for entropy II, ♫) Let X_1^n be a finite collection of discrete random variables. Then

$$H(X_1^n) = \sum_{i=1}^{n} H(X_i \mid X_1^{i-1})$$

Main idea: Induct the elementary version.

Proof. We prove by induction on n. The first non-trivial case was proven in Theorem Chain rule for entropy, now assume the claim holds for a message of length $n-1$. Then we have that

$$\begin{aligned} H(X_1^n) &= \mathbf{E}[-\log P_{X_1^{n-1}}(X_1^{n-1})] + \mathbf{E}[-\log P_{X_n|X_1^{n-1}}(X_n \mid X_1^{n-1})] \\ &= H(X_1^{n-1}) + H(X_n \mid X_1^{n-1}) \\ &\overset{(!)}{=} \sum_{i=1}^{n-1} H(X_i \mid X_1^{i-1}) + H(X_n \mid X_1^{n-1}) = \sum_{i=1}^{n} H(X_i \mid X_1^{i-1}) \end{aligned}$$

Where in step (!) we used the inductive hypothesis. ♡

We saw that strictly speaking $H(X \mid Y)$ is not an entropy per se, but it has natural links with the entropy we have seen so far. Indeed: conditioning on some random variable can be interpreted as observing information. Intuitively, if we add more information to what we already know, we could not become less certain. This is formalised by the following result:

Theorem 3.7 (Giving information never increases entropy, ♩) Let X and Y be two random variables. Then

$$H(Y \mid X) \le H(Y)$$

with equality if and only if X and Y are independent.

Main idea: $H(Y) - H(Y \mid X)$ can be shown to be a $D(P_{XY} \| P_X P_Y)$.

Proof.

$$\begin{aligned} H(Y) - H(Y \mid X) &= \mathbf{E} \log \frac{P_{Y|X}(Y \mid X)}{P_Y(Y)} \\ &= \mathbf{E} \log \frac{P_{Y|X} P_X(X)}{P_Y(Y) P_X(X)} = D(P_{YX} \| P_X P_Y) \ge 0. \end{aligned}$$

> With equality if and only if the relative entropy is zero which we know is equivalent to saying that P_{YX} and $P_Y P_X$ are the same distribution, i.e: the random variables are independent. ♡

Corollary 3.8 (Subadditivity of entropy) For any finite collection X_1^n of discrete random variables, we have that $H(X_1^n) \leq \sum_{i=1}^n H(X_i)$

> *Proof.* Immediate from the (generalised) chain rule as well as conditioning doesn't increase entropy. ♡

Theorem 3.9 (Bounds on conditional entropy) For X and Y random variables, we have that

$$0 \leq H(Y \mid X) \leq \log|\mathscr{B}|$$

Where $\mathscr{B}$ is the alphabet in which Y takes values.

Main idea: Write $H(Y \mid X)$ as an expectation of an entropy.

> *Proof.* This follows easily from the observation that
>
> $$H(Y \mid X) = \mathbf{E}_X[H(Y \mid X = x)]$$
>
> and $H(Y \mid X = x) \leq \log|\mathscr{B}|$ as we know, and is non-negative. Monotonicity of expectation finishes the claim. ♡

We now present two important results regarding conditional entropies: the data processing inequality and Fano's inequality.

Theorem 3.10 (Data processing inequality for entropy, ♩) Let X be a discrete random variable on the alphabet $\mathscr{A}$, and $f : \mathscr{A} \longrightarrow \mathbf{R}$ be a function. Then

- $H(f(X)) \leq H(X)$, with equality if and only if f is injective.
- $H(f(X) \mid X) = 0$

Main idea: Entropy is preserved under injective maps, noting that $X \mapsto (X, f(X))$ is injective allows us to apply the chain rule in two different ways.

Proof. For the first part, we note that the map $\Phi(X) = (X, f(X))$ is trivially injective, and we know that for an injective map Φ, $H(\Phi(X)) = H(X)$, this is because entropy depends solely on the probabilities assigned to values and not the values that the random variable takes per se. Therefore we conclude by the chain rule that

$$H(X) = H(X, f(X)) = H(X) + H(f(X) \mid X)$$

This can only be possible if $H(f(X) \mid X) = 0$.

For the second part, we note that we could have also concluded, by the symmetric application of the chain rule, that

$$H(X) = H(f(X)) + H(X \mid f(X))$$

since entropy non-negative so automatically we get that

$$H(f(X)) \leq H(X)$$

and we know that equality holds if and only if $H(X \mid f(X)) = 0$, that is to say, X is a deterministic function of $f(X)$, i.e: f is in particular invertible. ♡

Theorem 3.11 (Fano's Inequality, ♪) Let (X, Y) be an arbitrary pair of discrete RVs taking values in finite alphabets $\mathscr{A}$ and $\mathscr{B}$. Let $\widehat{X} = f(Y)$ for some function $f : \mathscr{A} \longrightarrow \mathscr{B}$, and write the probability of error $P_e = \mathbf{P}(\widehat{X} \neq X)$. Then

$$H(X \mid Y) \leq h(P_e) + P_e \log(|\mathscr{A}| - 1)$$

Where $h(p)$ denotes the binary entropy.

The interpretation of this is as follows. Suppose we are trying to estimate a random variable X, but we are only allowed to observe a possibly correlated random variable Y. Fano's inequality says that if the probability of error is small, i.e: we can do a good job at estimating X from Y, then once we condition on Y, there is little information left on X, that is to say $H(X \mid Y)$ is small.
Main idea: Consider $E \sim \text{Ber}(P_e)$. We want to study the quantity $H(X \mid Y)$ so apply chain rule in both directions to $H(X, E \mid Y)$, as well as data processing.

Proof. Let $E \sim \text{Ber}(P_e)$. Clearly $H(E) = h(P_e)$. From the chain rule for conditional entropy we have that

$$H(X, E \mid Y) = H(X \mid Y) + H(E \mid X, Y)$$

Since E is a function of X and Y, by the Data processing inequality for entropy, we have that $H(E \mid X,Y)=0$, so we deduce on the one hand that $H(X,E \mid Y)=H(X \mid Y)$. On the other hand, we can use the chain rule in reverse:

$$H(X,E \mid Y)=H(E \mid Y)+H(X \mid E,Y)\leq H(E)+H(X \mid E,Y)=h(P_e)+H(X \mid E,Y)$$

where in this calculation we have used the fact that conditioning does not increase entropy. All in all we have that

$$H(X \mid Y)\leq h(P_e)+H(X \mid E,Y)$$

But E being bernoulli means it is very easy to compute $H(X \mid E,Y)$. Indeed:

$$H(X \mid E,Y)=P_eH(X \mid E=1,Y)+(1-P_e)H(X \mid E=0,Y)$$

If $E=0$, then it means that $X=f(Y)$ and as such $H(X \mid E=0,Y)=0$, and if $E=1$, then it means that given Y, we can discard one of the values that X could take, so X can take $|\mathscr{A}|-1$ values, and since entropy is upper bounded by the logarithm of the alphabet size, we have that $H(X \mid E=1,Y)\leq \log(|\mathscr{A}|-1)$. This finishes the claim. ♡

3.1 Examples

Let us see a few questions with interesting techniques:

Example 3.12 Let $\mathscr{A}$ be a finite alphabet and let $B\subseteq\mathscr{A}^n$ be a collection of strings of length n. In this question we will find an interesting bound for $|B|$. As we will see later on these notes, for a string $x_1^n\in\mathscr{A}^n$, we define its type $\widehat{P}_{x_1^n}$ to the empirical distribution on $\mathscr{A}$ induced by the string, meaning:

$$\widehat{P}_{x_1^n}(a)=\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{x_i=a\}.$$

Let us now define the following probability measure on B:

$$P_B(a)=\frac{1}{|B|}\sum_{x_1^n\in B}\widehat{P}_{x_1^n}(a).$$

Show that

$$|B|\leq 2^{nH(P_B)}.$$

Proof. The hint is to consider random variables X_1^n and J that are uniform on B and $\{1,\cdots,n\}$

respectively. We see that $P_B(a) = \mathbf{P}(X_J = a)$. Therefore,

$$\begin{aligned}
-H(P_B) &= \sum_{a \in \mathscr{A}} \mathbf{P}(X_J = a) \log(\mathbf{P}[X_J = a]) \\
&= \sum_{a \in \mathscr{A}} \left(\sum_{i=1}^{n} \frac{1}{n} \mathbf{P}[X_i = a] \right) \log \left(\sum_{i=1}^{n} \frac{1}{n} \mathbf{P}[X_i = a] \right) \\
&\leq \sum_{a \in \mathscr{A}} \sum_{i=1}^{n} \frac{1}{n} \mathbf{P}[X_i = a] \log(\mathbf{P}[X_i = a]) \\
&= \frac{1}{n} \sum_{i=1}^{n} -H(X_i)
\end{aligned}$$

and so $nH(P_B) \geq \sum_{i=1}^{n} H(X_i) \geq H(X)$ by subadditivity of entropy. But since X is uniform on B, we have that $H(X) = \log|B|$ and so we get the claim. ♡

3.2 Properties of relative entropy

Returning to the question of hypothesis testing, we saw that if we wish to perform a hypothesis test between two distributions P and Q, and we impose the probability of a Type II error approaching zero asymptotically, then the best decay exponent of the Type I error probability we could have was $D(P\|Q)$. A natural question would be, if we process our observations x_1^n, say by applying some function $f(x_1^n)$, can we make our distributions more statistically distinguishable? The answer to this is no, and its given by the data processing inequality for relative entropy.

Theorem 3.13 (Data processing inequality for relative entropy) Let X and Y be two random variables with distributions P_X and Q_Y taking values in the same alphabet $\mathscr{A}$. Let $f : \mathscr{A} \longrightarrow \mathscr{B}$ be a function to some other alphabet $\mathscr{B}$. Then, writing $P_{f(X)}$ and $Q_{f(Y)}$ for the distributions of $f(X)$ and $f(Y)$ gives that

$$D(P_{f(X)}\|Q_{f(Y)}) \leq D(P_X\|Q_Y)$$

Main idea: Use the fact that for a general function $f : \mathscr{A} \longrightarrow \mathscr{B}$, we can write

$$\mathscr{A} = \bigcup_{y\in f(\mathscr{A})} f^{-1}(\{y\}) = \bigcup_{y\in f(\mathscr{A})} \bigcup_{x\in f^{-1}(\{y\})} \{x\}$$

Proof.

$$\begin{aligned} D(P_X\|Q_X) &= \sum_{y\in f(\mathscr{A})} \sum_{x\in f^{-1}(\{y\})} P_X(x) \log \frac{P_X(x)}{Q_X(x)} \\ &\geq \sum_{y\in f(\mathscr{A})} \left(\sum_{x\in f^{-1}(\{y\})} P_X(x) \right) \log \frac{\left(\sum_{x\in f^{-1}(\{y\})} P_X(x)\right)}{\left(\sum_{x\in f^{-1}(\{y\})} Q_X(x)\right)} \end{aligned}$$

Where we have used log-sum. Now we note that

$$\sum_{x\in f^{-1}(\{y\})} P_X(x) = P_{f(X)}(y)$$

Indeed: the probability that $f(X)$ takes the value y is precisely the probability that X takes **some** value x that gets mapped to y under f. Since the singletons $\{x\}$ are disjoint, we turn the probability of the **union** into the sum of the probabilities. This finishes the proof. ♡

We have spoken of $D(P\|Q)$ so far as an *informal* distance between probability measures. Indeed: we have seen that $P = Q$ if and only if $D(P\|Q) = 0$, but we are going to see how we can also get quantitative bounds on the actual difference of the probabilities. First we need to introduce a genuine distance between probability measures:

Definition 3.14 (Total Variation) Let P and Q be two probability measures on $\mathscr{A}$. The Total Variation distance is defined by

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

Note that sometimes the convention is to write a $1/2$ in front. If you have suffered through mixing times of Markov chains you will know. Recall also that TV distance has a convenient alternative characterisation: letting $B=\{x:P(x)>Q(x)\}$ gives

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

We now have the result that quantifies the relation between relative entropy and an actual distance between probability measures:

Theorem 3.15 (Pinsker's inequality)

$$\|P-Q\|_{\mathsf{TV}}^2 \le (2\log_e 2)D(P\|Q) \equiv 2D_e(P\|Q)$$

Main idea: There are two steps:

1. Show the inequality for the binary case $P\sim\mathsf{Ber}(p)$ and $Q\sim\mathsf{Ber}(q)$. We may assume WLOG that $q\le p$ to do this compute everything explicitly, and letting $\Delta(p,q)=2D_e(P\|Q)-\|P-Q\|_{\mathsf{TV}}^2$, and noting that $\Delta(p,p)=0$ and $\partial_q\Delta(p,q)\le 0$ we get that $\Delta(p,q)\ge 0$.

2. To reduce the general case to the binary case we consider functions of P and Q where $P'=\mathscr{L}(\mathbf{1}(X\in B))$ where $X\sim P$ and similarly for Q', where B is the "special TV set". Then by employing the binary case and some explicit TV computations the claim follows.

Proof. To show the binary case first: let P and Q be the distributions of a $\mathsf{Ber}(p)$ and a $\mathsf{Ber}(q)$ random variable respectively. To compute $D(P\|Q)$, we assume that $p\ge q$. This assumption is justified because even though $D(P\|Q)$ is not symmetric, noting that for the binary case

$$D(\mathsf{Ber}(p)\|\mathsf{Ber}(q)) = p\log\frac{p}{q}+(1-p)\log\frac{1-p}{1-q} = D(\mathsf{Ber}(1-p)\|\mathsf{Ber}(1-q))$$

So if p were not greater than or equal to q, we could just swap $p\longleftrightarrow 1-p$ and $q\longleftrightarrow 1-q$ and preserve $D(P\|Q)$. Now the goal is to show that the quantity

$$\Delta(p,q) = 2D_e(P\|Q) - \|P-Q\|_{\mathsf{TV}}^2 \geq 0$$

We think of p as being fixed, and since $\Delta(p,p)$ is clearly zero, and we have assumed $q \leq p$, it is sufficient for us to show that $\frac{\partial \Delta(p,q)}{\partial q} \leq 0$ (see the graph to convince yourself). Since the binary case is easy, we can explicitly compute $\Delta(p,q)$ and get

$$\Delta(p,q) = 2p \ln \frac{p}{q} + 2(1-p)\ln\frac{1-p}{1-q} - (2(p-q))^2$$

The derivative gives

$$\frac{\partial \Delta(p,q)}{\partial q} = (p-q)\left(4 - \frac{1}{q(1-q)}\right) \geq 0$$

Where this last inequality (and hence the desired result) follows because $f(x) = x(1-x)$ is a downward parabola with height at most $1/4$, so the difference $4 - \frac{1}{q(1-q)} \geq 0$.

Now that we have established the validity of Pinsker's inequality for the binary case, we can reduce the general case to this one by the data processing inequality. Indeed: let P and Q now take values in a general alphabet $\mathscr{A}$, and let $B = \{x : P(x) > Q(x)\}$. Then define P' and Q' be the mass functions of Bernoulli's $P(B)$ and $Q(B)$ respectively. P' and Q' can be thought of as the mass functions of $\mathbf{1}(X \in B)$ and $\mathbf{1}(Y \in B)$ respectively for $X \sim P$ and $Y \sim Q$. So

$$\begin{aligned}
2D_e(P\|Q) &\geq 2D_e(P'\|Q') && \text{(Data processing)}\\
&\geq \left\|P' - Q'\right\|_{\mathsf{TV}}^2 && \text{(Binary case)}\\
&= (2(P(B) - Q(B)))^2 && \text{(Explicit computation of TV for binary RVs)}\\
&= \|P-Q\|_{\mathsf{TV}}^2 && \text{(Alternative expression for TV)}
\end{aligned}$$

♡

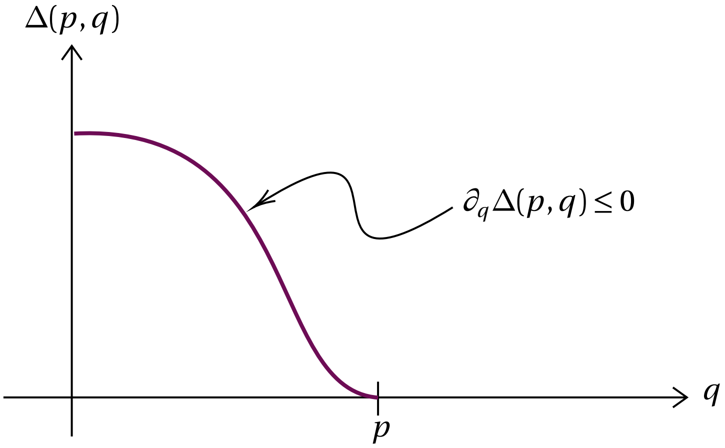


Figure 3: Idea for the proof of the binary case of Pinsker's inequality

4 Poisson approximation

It is a fact in elementary probability that if $(X_i : 1 \leq i \leq n)$ is a collection of iid Bernoulli λ/n, and we let

$$S_n = \sum_{i=1}^{n} X_i$$

I.e: $S_n \sim \text{Bin}(n, \lambda/n)$, then we have that

$$S_n \xrightarrow{d} \text{Poi}(\lambda)$$

It turns out that statements that are both stronger and more general are possible.

Theorem 4.1 (Entropic Poisson approximation) Let $X_1, \cdots, X_n$ be not necessarily independent Bernoulli random variables with $X_i \sim \text{Ber}(p_i)$. Let $\lambda = \sum_{i=1}^{n} p_i$ and $S_n = \sum_{i=1}^{n} X_i$. Then we have that

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

Before we prove the Theorem we disect the conclusion: if we want convergence (for example in TV), we want to make the right hand side small. The first sum can be killed by having the (p_i)s close to being all the same, i.e: not having a single p_i dominating the others. In fact, if we use all $p_i = \lambda/n$, then $\sum_{i=1}^{n} p_i^2 \longrightarrow 0$, and if we moreover assume independence, then the second sum disappears and we recover the usual Poisson convergence result. This Theorem is very nice because it very clearly quantifies the distance in terms of "how identically distributed the X_is are", and "how independent" they are.

Main idea: Express $T_n \sim \text{Poi}(\lambda)$ as a function of Z_1^n where $Z_i \sim \text{Poi}(p_i)$ and then apply data processing.

Proof. We use the fact that if $A \sim \text{Poi}(\alpha)$ and $B \sim \text{Poi}(\beta)$ are independent, then $A+B \sim \text{Poi}(\alpha+\beta)$. Then if we let $T_n \sim \text{Poi}(\lambda)$, we can write

$$T_n \overset{d}{=} \sum_{i=1}^{n} Z_i$$

Where each $Z_i \sim \text{Poi}(p_i)$ is independent. Similarly we have that $S_n = \sum_{i=1}^{n} X_i$, so we have expressed $S_n = f(X_1^n)$ and $T_n = f(Z_1^n)$. By data processing:

$$\begin{aligned}
D(P_{S_n} \| \text{Poi}(\lambda)) \leq D(P_{X_1^n} \| P_{Z_1^n}) & \\
&= \sum_{x \in \mathscr{A}^n} P_{X_1^n}(x) \log \frac{P_{X_1^n}(x)}{P_{Z_1^n}(x)} \\
&= \sum_{x \in \mathscr{A}^n} P_{X_1^n}(x) \log \left(\frac{P_{X_1^n}(x)}{P_{Z_1^n}(x)} \cdot \frac{\prod_{i=1}^n P_{X_i}(x_i)}{\prod_{i=1}^n P_{X_i}(x_i)} \right) \\
&= \sum_{x \in \mathscr{A}^n} P_{X_1^n}(x) \log \left(\frac{P_{X_1^n}(x)}{\prod_{i=1}^n P_{Z_i}(x_i)} \cdot \frac{\prod_{i=1}^n P_{X_i}(x_i)}{\prod_{i=1}^n P_{X_i}(x_i)} \right) \qquad \text{(Each } Z_i \text{ is independent)} \\
&= \sum_{x \in \mathscr{A}^n} P_{X_1^n}(x) \log P_{X_1^n}(x) + \sum_{x \in \mathscr{A}^n} P_{X_1^n}(x) \log \frac{1}{\prod_{i=1}^n P_{X_i}(x_i)} + \sum_{x \in \mathscr{A}^n} P_{X_1^n}(x) \log \left(\prod_{i=1}^n \frac{P_{X_i(x_i)}}{P_{Z_i}(x_i)} \right) \\
&= -H(X_1^n) + \sum_{i=1}^n H(X_i) + \sum_{i=1}^n D(P_{X_i} \| P_{Z_i})
\end{aligned}$$

In this last computation we have used the Marginal trick (A.3). Now we only need to use the Lemma we will prove later, that says that $D(\text{Ber}(p) \| \text{Poi}(p)) \leq p^2$, and since $Z_i \sim \text{Poi}(p_i)$ and $X_i \sim \text{Ber}(p_i)$ we apply this and we are done. ♡

Lemma 4.2

$$D(\text{Ber}(p) \| \text{Poi}(p)) \leq p^2$$

Proof. If we compute the quantity explicitly, we see that

$$D(\text{Ber}(p) \| \text{Poi}(p)) = p \log(e^p) + (1-p) \log\big((1-p)e^p\big)$$

and we now claim that $(1-p)e^p \in [0,1]$ whenever $p \in [0,1]$, which means that the log will be negative and as such the claim will follow. To prove this last step, we simply note that when $p = 0$, $(1-p)e^p = 1$, when $p = 1$, $(1-p)e^p = 0$ and the derivative

$$\frac{d}{dp}(1-p)e^p = -pe^p \leq 0$$

♡

We have now some immediate corollaries to the Entropic Poisson Approximation Theorem

Corollary 4.3 If $X_1, \cdots, X_n$ are independent Bernoulli's p_i, then

$$D_e(P_{S_n} \| \text{Poi}(\lambda)) \leq \sum_{i=1}^{n} p_i^2$$

And moreover, this results strengthen the elementary result we talked about at the start of this section, because if in addition to independence, we assume that $p_i = \lambda/n$, then we have that

$$D_e(P_{S_n} \| \text{Poi}(\lambda)) \leq \frac{\lambda^2}{n}$$

Which even more amazingly gives you explicit bounds on the distance between $P_{S_n}(x)$ and the respective probability of x under Poisson for a given n. Indeed, if P_λ denotes the probability measure of the Poisson with parameter λ. Then

$$|P_{S_n}(x) - P_\lambda(x)| \leq \left\| P_{S_n} - P_\lambda \right\|_{\text{TV}} \leq \frac{\sqrt{2}\lambda}{\sqrt{n}}$$

4.1 Poisson distribution as a maximal entropy result

Here is a motivating question: suppose that we are trying to count the number of events that occur during a given time interval, say $[0,1]$, with the only information known to us its the mean number of events, λ. A natural construction is to partition the interval $[0,1]$ into n bins, say $I_i = \left[\frac{i-1}{n}, \frac{i}{n}\right)$ for $1 \leq i \leq n$. And we let X_i be a random variable that denotes if an event has occurred on the interval I_i. If we assume that the events occur maximally at random throughout $[0,1]$, then $X_i \sim \text{Ber}(\lambda/n)$, and so the number of occurrences is given by

$$S_n = \sum_{i=1}^{n} X_i \sim \text{Bin}(n, \lambda/n) \xrightarrow{d} \text{Poi}(\lambda)$$

We will now show a result that gives an information theoretic interpretation for the Poisson distribution as a "maximally random" distribution. Suppose we are once again interested in counting the number of occurrences of some discrete events that happen independently of each other. If each event, of say n total events occurs with probability p_i, we can simply model whether the event has occurred by $X_i \sim \text{Ber}(p_i)$, and the total count of events is given by $S_n = \sum_{i=1}^{n} X_i$, subject to the assumption that we know the average number of events, $\mathbf{E}[S_n] =: \lambda$, the following Theorem tells us that the maximally random distribution that can be achieved by any such S_n is that of a Poisson random variable.

Theorem 4.4 (Poisson maximum entropy)

$$\sup\left\{H_e(S_n) : S_n = \sum_{i=1}^{n} X_i \text{ for independent } X_i \sim \mathsf{Ber}(p_i), \text{ with } \sum_{i=1}^{n} p_i = \lambda, n \geq 1\right\} = H_e(\mathsf{Poi}(\lambda))$$

We require the following Lemma:

Lemma 4.5 (Binomial as maximum entropy) Let

$$B_n(\lambda) = \left\{\mathscr{L}(S_n) : S_n = \sum_{i=1}^{n} X_i, \{X_i\}_\perp = \mathsf{Ber}(p_i), \lambda = \sum_i p_i\right\}$$

Be the set of distributions that arise as sums of independent Bernoulli p_i's where the p_i's sum to λ. Then

$$\sup\{H(P) : P \in B_n(\lambda)\} = H(\mathsf{Bin}(n, \lambda/n))$$

That is to say, if you have n events, and each occur with a probability p_i, with $\lambda = \sum p_i$ the most random way in which they can occur is described by a Binomial $\mathsf{Bin}(n, \lambda/n)$. With this Lemma we may approach the Poisson maximum entropy Theorem.

Proof. We start by noting that of course, the supremum in the Theorem is just $H^* = \sup_n \sup\{H(P) : P \in B_n(\lambda)\}$, and since $B_n(\lambda) \subseteq B_{n+1}(\lambda)$, indeed, if $\mathscr{L}(S_n) \in B_n(\lambda)$, then $\mathscr{L}(S_n) = \mathscr{L}(S_n + X_{n+1})$ where $X_{n+1} \sim \mathsf{Ber}(0)$. So by this monotonicity, we have that

$$\sup_n \sup\{H(P) : P \in B_n(\lambda)\} = \lim_{n \longrightarrow \infty} \sup\{H(P) : P \in B_n(\lambda)\} \overset{(!)}{=} \lim_{n \longrightarrow \infty} H(P_n)$$

Where $P_n = \mathscr{L}(\mathsf{Bin}(n, \lambda/n))$, and step (!) is due to the Binomial as maximum entropy Lemma. We therefore just need to show that we can in some sense pass the limit inside the entropy symbol. The way we go about this is as follows:

$$\begin{aligned} H(P_n) &= -\sum_{x \in \mathbf{N}} P_n(x) \log(P_n(x)) \\ &= -\sum_{x \in \mathbf{N}} P_n(x) \log\left(\frac{P_n(x)}{Q(x)} Q(x)\right) \\ &= -\sum_{x \in \mathbf{N}} P_n(x) \log(Q(x)) - D(P_n \| Q) \end{aligned}$$

From the entropy convergence of P_n to Q (Theorem 4.1) we already know that $\lim_{n \longrightarrow \infty} D(P_n \| Q) =$

0, so all that's left for us to determine is whether

$$-\lim_{n\longrightarrow\infty}\sum_{x\in\mathbf{N}}P_n(x)\log(Q(x))=H(Q)$$

The way we go about this is by treating this sum as an expectation with respect to the discrete counting measure, and then we will show two things: that $H(Q)$ is finite, and that $P_n(x)\lesssim Q(x)$, then it will follow that $P_n(x)\log Q(x)$ is bounded above (up to constant) by $Q(x)\log Q(x)$ which is integrable (with respect to the counting measure, indeed, it will be equal to the entropy which we will have shown to be finite).

- $H(Q)<\infty$:

$$\begin{aligned}H(Q)&:=\sum_x Q(x)\log\frac{x!}{e^{-\lambda}\lambda^x}\\&=\sum_x Q(x)\lambda-\log\lambda\sum_x xQ(x)+\sum_x Q(x)\log(x!)\end{aligned}$$

  There's no doubt that the first two sums are finite, indeed the first one is finite by normalisation of Q and the second one is finite because it equals the expectation of a Poisson random variable which is just λ. So all we need to show is that the last sum is also finite. To do so we will employ the following bound: $x!\leq x^x=\exp(x\log x)\leq\exp(x^2)$.

$$\sum_x Q(x)\log(x!)\leq\sum_x Q(x)\log(\exp(x^2))=\sum_x Q(x)x^2$$

  which is finite because the Poisson distribution has finite variance.

- $P_n(x)\lesssim Q(x)$:

$$\begin{aligned}P_n(x)&=\binom{n}{x}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x}\\&\overset{(1)}{\leq}\binom{n}{x}\left(\frac{\lambda}{n}\right)^x\\&=\frac{n!\lambda^x}{x!(n-x)!n^x}\\&\overset{(2)}{\leq}\frac{\lambda^x}{x!}\leq\exp(\lambda)Q(x)\end{aligned}$$

  Where step (1) comes from the fact that the thing that disappeared is at most one, and step (2) comes from the fact that $n!\leq\underbrace{n\times n\times\cdots\times n}_{x\text{ times}}\times(n-x)!$.

This finishes the proof. ♡

5 Mutual Information

Suppose X and Y are two random variables, we would like to now study how much information the two random variables share, i.e: how much does knowing about one random variable tell you about the other.

Definition 5.1 (Mutual information) Let X and Y be two random variables. Their mutual information $I(X;Y)$ is given by any of the following equivalent characterisations

$$\begin{aligned} I(X;Y) &= H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) \\ &= H(X) + H(Y) - H(X,Y) \\ &= D(P_{XY} \| P_X P_Y) \end{aligned}$$

It is a straightforward application of the chain rule to verify that these are indeed equivalent definitions, and each tells us slightly different implications to the nature of mutual information. The first one tells us that $I(X;Y)$ is the uncertainty in X minus the uncertainty left in X after observing Y, i.e: precisely how much information Y told you about X. The second characterisation tells you that $I(X;Y)$ is the difference between the sum of the individual uncertainties of X and Y and the uncertainty that they contain as a joint random variable, this definition is perhaps not very enlightening but it tells you that $I(X;Y) = I(Y;X)$. Perhaps more illustrative is definition three, that tells you that $I(X;Y)$ is quantified by "how far away" in the sense of relative entropy, X and Y are from being independent. In light of Stein's Lemma, if we wanted to carry out a test to detect whether X and Y are independent, the best decay of error probability we can hope for is $I(X;Y)$.

Definition 5.2 (Conditional mutual information) Let X, Y, Z be random variables, their conditional mutual information $I(X;Y \mid Z)$ is given by

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

We now present some basic properties of mutual information.

Proposition 5.3 (Basic properties of mutual information) Let X, Y, Z be random variables. Then

- $I(X;Y) \geq 0$ with equality if and only if X and Y are independent.
- $I(X;Y \mid Z) \geq 0$ with equality if and only if $X - Z - Y$ forms a Markov chain.

Main idea: The first bulletpoint is obvious when looking at (an equivalent) definition $I(X;Y) = D(P_{XY}\|P_X P_Y)$. The second bulletpoint comes from the fact that we can express

$$I(X;Y \mid Z) = \sum_z p(z) D\left(P_{XY|\{Z=z\}} \| P_{X|\{Z=z\}} P_{Y|\{Z=z\}}\right)$$

Proof. The first part is clear from the fact that $I(X;Y) = D(P_{XY}\|P_X P_Y)$ and this quantity obviosuly satisfies the desirede property. To show the second part, we simply note that

$$\begin{aligned} I(X;Y \mid Z) &= H(X \mid Z) + H(Y \mid Z) - H(X, Y \mid Z) \\ &= \mathbf{E}_Z \left[\mathbf{E}\left[-\log P_{X|Z}(X \mid z) - \log P_{Y|Z}(Y \mid z) + \log P_{XY|Z}(X, Y \mid z)\right]\right] \\ &= \mathbf{E}_Z \left[D(P_{XY|Z}(X, Y \mid z) \| P_{X|Z}(X \mid z) P_{Y|Z}(Y \mid z))\right] \end{aligned}$$

Since relative entropy is non-negative, non-negativity follows. For the equality claim it is easy to see that we must have that for each z, $D(P_{XY|Z}(X, Y \mid z) \| P_{X|Z}(X \mid z) P_{Y|Z}(Y \mid z)) = 0$, which means that X and Y are conditionally independent given Z. ♡

Proposition 5.4 (Chain rule for mutual information) For a sequence X_1^n of random variables and another random variable Y, we have that

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

Main idea: Definition of mutual information and use the chain rules for entropy and conditional entropy.

Proof.

$$\begin{aligned} I(X_1^n; Y) &= H(X_1^n) - H(X_1^n \mid Y) \\ &= \sum_{i=1}^n H(X_i \mid X_1^{i-1}) - H(X_i \mid X_1^{i-1}, Y) \\ &= \sum_{i=1}^n I(X_i; Y \mid X_1^{i-1}). \end{aligned}$$

♡

Recall that whenever we talk about random variables being conditionally independent, or we say about $X - Y - Z$ forming a Markov chain, what we mean is that

$$\begin{aligned} P_{XZ|Y}(x,z \mid y) &:= \mathbf{P}(X = x, Z = z \mid Y = y) \\ &\overset{!}{=} P(X = x \mid Y = y) P(Z = z \mid Y = y) \\ &=: P_{X|Y}(x \mid y) P_{Z|Y}(z \mid y) \end{aligned}$$

An important class of said Markov chains is when we have $X - Y - F(Y)$. Clearly, given the value of Y, X and $F(Y)$ are independent, and this can be interpreted as having two random variables X, Y (possibly correlated), and then receiving a noisy observation of Y. It is unsurprising then that we have the following result

Theorem 5.5 (Data processing for mutual information) Let $X - Y - Z$ form a Markov chain, then:

- $I(X;Z) \leq I(X;Y)$
- $I(X;Y \mid Z) \leq I(X;Y)$

The first statement should be immediately intuitive, specially when we think of the chain $X - Y - F(Y)$ observing a noisy version of Y will give no more information of X than observing Y directly. The second statement tells us that by observing the value of a random variable Z that is "downstream", one cannot all of a sudden increase the amount of information that observing Y tells us about X, i.e: observing a Z cannot add more dependencies, it can only remove already existing dependencies, hence reducing the information.
Main idea: Use the chain rule to expand $I(X;Y,Z)$ in two different ways.

Proof. By the chain rule, we have that

$$I(Y,Z;X) = I(Y;X) + I(Z;X \mid Y)$$

But also

$$I(Y,Z;X) = I(Z;X) + I(Y;X \mid Z)$$

However, $I(Z;X \mid Y) = 0$ because of the assumption that conditional on Y, Z and X are independent, so we have that

$$I(X;Y) = I(X;Z) + I(X;Y \mid Z)$$

and since mutual information is non-negative, from here both inequalities follow. In fact, we get conditions for the equalities:

- $I(X;Y) = I(X;Z)$ if and only if X and Y are conditionally independent given Z.

- $I(X;Y) = I(X;Y \mid Z)$ if X and Z are independent.

♡

5.1 Synergy

Note that $I(X;Y_1,Y_2)$ can be greater than/equal/less than $I(X;Y_1)+I(X;Y_2)$. As a toy example, suppose that X represents some "health condition" which can be "at risk" or "not at risk", and Y_1 represents whether a person exercises regularly, and Y_2 represents whether the person has a balanced diet. Clearly knowing Y_1 will give some information about X, and Y_2 will also give some information about X, but the information that you gain from knowing both at once is intuitively larger than the sum of the informations you gain from knowing each independently, this is because knowing only Y_1 or Y_2 *actually doesn't tell you that much about the health of the patient*: it could be the case that the patient exercises 3 times a week but also drank 40 liters of vodka last year [1], or it could be that the patient eats a healthy diet, but hasn't exercised since middle school. In this case, (Y_1, Y_2) provide synergystic information about X.

Definition 5.6 (Synergy) We define the synergy of (Y_1, Y_2) about X as:

$$S(X;Y_1,Y_2) = I(X;Y_1,Y_2) - [I(X;Y_1) + I(X;Y_2)]$$

Remark 5.7 We can use the chain rule to re-express synergy as

$$S(X;Y_1,Y_2) = I(X;Y_2 \mid Y_1) - I(X;Y_2)$$

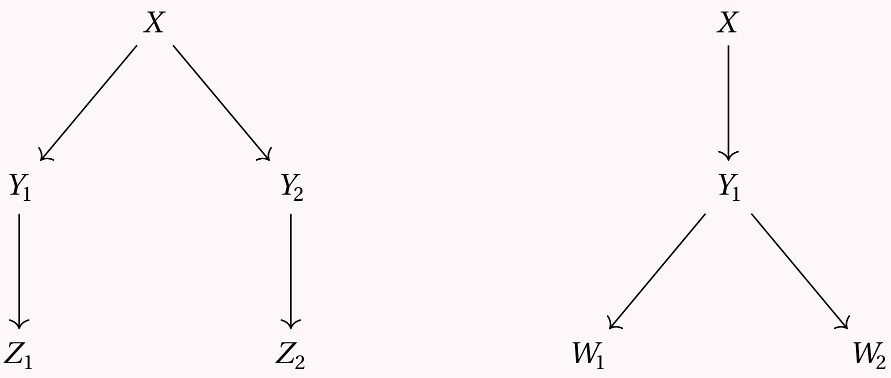
And asking whether the right hand side is positive or negative admits the following interpretation: Does knowing Y_1 make Y_2 more or less useful in determining X.

Definition 5.8 If $S(X;Y_1,Y_2)$ is:

- Positive: we say (Y_1, Y_2) are synergystic. (They say different things about X and the pieces of information complement each other. They are independent and build on each other)
- Zero: we say (Y_1, Y_2) are orthogonal. (They say different things about X but the two pieces of information they give don't complement each other. They are just independent)
- Negative: we say (Y_1, Y_2) are redundant. (They say roughly the same things about X)

[1] If you think this is too specific for me to have made it up, you are right

Example 5.9 (Are independent samples better?) Suppose we are allowed to observe two equally noisy versions of some signal X, where the level of noise if fixed, i.e: this corresponds to observing some $Y = f(X)$. What is better: to observe two independent noisy versions of X, say Y_1 and Y_2, or two correlated ones?



For X being our variable of interest, let $P_{Y|X}$ and $Q_{Z|Y}$ be the collections of conditional PMFs. In the first scenario of independent observations, we have that Y_1 and Y_2 are conditionally independent given X, i.e: $P_{Y_1,Y_2|X} = P_{Y_1|X}P_{Y_2|X}$. And both distributed according to $P_{Y|X}(\cdot \mid X)$ And let Z_1 and Z_2 be distributed according to $Q_{Z|Y}(\cdot \mid Y_1)$ and $Q_{Z|Y}(\cdot \mid Y_2)$ respectively. In the second scenario of correlated observations, we have that $Y \sim P_{Y|X}(\cdot \mid X)$, and then W_1 and W_2 are conditionally independent given Y and are distributed according to $Q_{Z|Y}(\cdot \mid Y)$.

It is easy to check that the joint distributions of (X, Z_1) and (X, W_1) are the same, indeed:

$$\begin{aligned}
\mathbf{P}(X = x, Z_1 = z) &= \mathbf{P}(Z_1 = z \mid Y = y)\mathbf{P}(Y = y \mid X = x)\mathbf{P}(X = x) \\
&= Q_{Z|Y}(z, y)P_{Y|X}(x, y)P_X(x) \\
&= \mathbf{P}(W_1 = z \mid Y = y)\mathbf{P}(Y = y \mid X = x)\mathbf{P}(X = x) \\
&= \mathbf{P}(X = x, W_1 = z)
\end{aligned}$$

However, the joint distributions (X, Z_1, Z_2) and (X, W_1, W_2) are different. To answer the question as to which observation we should pick, we have the following result.

Proposition 5.10 In the same setting as the above example, if the synergy $S(X; W_1, W_2)$ is positive, then correlated observations are better, in the sense that

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

Main idea: Start by noting that $S(X; W_1, W_2) > 0$ if and only if $I(X; W_2 \mid W_1) > I(X; W_2)$ and due to the setup, since $(X, W_2) \overset{(d)}{=} (X, Z_2)$ that is equal to $I(X; Z_2)$, then use data processing and $Z_1 - X - Z_2$

form a Markov chain.

Proof. Recall that since $S(X;Y_1,Y_2) := I(X;Y_1,Y_2) - I(X;Y_1) - I(X;Y_2)$, we can use the chain rule and express this simply as

$$S(X;Y_1,Y_2) = I(X;Y_2 \mid Y_1) - I(X;Y_2)$$

So if $S(X;W_1,W_2) > 0$, then we have that $I(X;W_2 \mid W_1) > I(X;W_2) = I(X;Z_2)$. But due to the assumption of Z_1 and Z_2 being conditionally independent given X, the data processing property implies that $I(X;Z_2) \geq I(X;Z_2 \mid Z_1)$, so putting it all together:

$$I(X;W_2 \mid W_1) > I(X;Z_2 \mid Z_1)$$

Now we can add $I(X;Z_1) = I(X;W_1)$ to both sides of the equation and deduce that

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

♡

6 Entropy and additive combinatorics

In this section we will show some entropy bounds that are inspired from ideas in additive combinatorics. First we note that if A and B are subsets of the integers, then it is obvious that

$$\max\{|A|,|B|\} \le |A+B| \le |A||B|$$

Proposition 6.1 (Simple sumset entropy bounds) If X and Y are independent integer-valued RVs, then

$$\max\{H(X),H(Y)\} \le H(X+Y) \le H(X)+H(Y)$$

Proof. It's a rather simple calculation

$$\begin{aligned} H(X)+H(Y) &\overset{(1)}{=} H(X,Y) \\ &\overset{(2)}{=} H(X,X+Y) \\ &\overset{(3)}{=} H(X+Y)+H(X \mid X+Y) \\ &\overset{(4)}{\le} H(X+Y)+H(X) \end{aligned}$$

And so $H(Y) \le H(X+Y)$. By symmetry the lower bound follows. What happened here was, in step (1) we used independence of X and Y, in step (2) we used the fact that the map $(x,y) \mapsto (x,x+y)$ is $1-1$. Step (3) follows from the chain rule, and step (4) comes from Conditioning Reduces Entropy. Incidentally this also shows the upper bound, indeed, looking at equation (3) and using the fact that $H(X \mid X+Y) \ge 0$ also shows the upper bound. ♡

Another inequality from additive combinatorics is Rusza's triangle inequality, that says

$$|A-C| \le \frac{|A-B||B-C|}{|B|}$$

This also admits an entropic version:

Theorem 6.2 (Rusza triangle ienquality for entropy) If X,Y,Z are independent integer-valued Random Variables, we have that

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

Main idea: First show that $I(X;(X-Y,Y-Z)) \ge I(X;X-Z)$. Showing inequalities for mutual information is best done with the data processing property. Then re-express each side of the inequality in terms of entropies.

Proof. Note that since $X-Z=(X-Y)+(Y-Z)$, given $(X-Y, Y-Z)$, $X-Z$ is a constant and so X and $X-Z$ are conditionally independent given $(X-Y, Y-Z)$. Therefore by the data processing inequality for mutual information, we have that

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

However, note that

$$\begin{aligned} I(X;X-Z) &:= H(X-Z) - H(X-Z \mid X) \\ &\overset{(1)}{=} H(X-Z) - H(Z \mid X) \\ &\overset{(2)}{=} H(X-Z) - H(Z) \end{aligned}$$

Where (1) comes from the fact that given $X=x$, the map $x-z \mapsto z$ is injective. Step (2) comes from the independence of X and Z. We can now also inspect

$$\begin{aligned} I(X;(X-Y, Y-Z)) &:= H(X) + H(X-Y, Y-Z) - H(X, X-Y, Y-Z) \\ &\overset{(1)}{=} H(X) + H(X-Y, Y-Z) - H(X, Y, , Z) \\ &\overset{(2)}{=} H(X-Y, Y-Z) - H(Y) - H(Z) \end{aligned}$$

Where step (1) comes from the fact that the map $(x, x-y, y-z) \mapsto (x, y, z)$ is injective, and step (2) comes from the fact that X, Y, Z are independent. Combining everything we have that

$$\begin{aligned} H(X-Z) &\leq H(X-Y, Y-Z) - H(Y) \\ &\leq H(X-Y) + H(Y-Z) - H(Y) \end{aligned}$$

Where in the last step we have used the subadditivity of entropy proven in the Theorem "Simple Sumset Entropy Bounds". ♡

The next and final inequality of this chapter is based on the following observation. If X_1, X_2 are IID RV.s, then we know from Simple Sumset Bounds, that $H(X_1) \leq H(X_1 + X_2)$ and similarly, by swapping $X_2 \mapsto -X_2$, we also have that $H(X_1) \leq H(X_1 - X_2)$. The question we ask now is, which - the sum or the difference of two IID RVs - increases the entropy more? The next Theorem asserts that the ratio of the differences lies between $1/2$ and 2.

Theorem 6.3 (Doubling-difference inequality) If X_1 and X_2 are IID integer-valued RVs, 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 the following lemma

Lemma 6.4 Let X, Y, Z be independent, integer valued, then

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

(This intuitively makes sense, because since Z is independent of everything else, the information X says about $X+Y$ is going to be greater than or equal than than the information that X tells you about $X+Y+Z$, since adding Z adds more randomness to the sum, and X can't tell anything about that randomness) As an immediate corollary:

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

Main idea: To show the first inequality, use data processing on a smartly selected Markov Chain and then use the chain rule as well as independence. To show the second result express both sides of the mutual information inequality in terms of entropies.

Proof. Since $X+Y+Z=(X+Y)+Z$, if we are given the value of the vector $(X+Y, Z)$, then $X+Y+Z$ is constant and so $X-(X+Y,Z)-(X+Y+Z)$ forms a Markov chain, so we have the following:

$$\begin{aligned} I(X+Y+Z;X) &\overset{(1)}{\leq} I(X+Y,Z;X) \\ &\overset{(2)}{=} I(X+Y;X)+I(Z;X \mid X+Y) \\ &\overset{(3)}{=} I(X+Y;X) \end{aligned}$$

Where step (1) comes from the data processing inequality and the observation that $X-(X+Y,Z)-(X+Y+Z)$ forms a Markov chain. Step (2) comes from the chain rule, and step (3) comes from the fact that Z and X are independent, so no information about Z is gained upon observing X, given the value of $X+Y$. To show the second claim, we simply note that

$$\begin{aligned} I(X+Y+Z;X) &:= H(X+Y+Z) - H(X+Y+Z \mid X) \\ &\overset{(1)}{=} H(X+Y+Z) - H(Y+Z \mid X) \\ &\overset{(2)}{=} H(X+Y+Z) - H(Y+Z) \end{aligned}$$

Where step (1) comes from the fact that given that $X=x$, the map $X+Y+Z \mapsto Y+Z$ is injective, and step (2) comes from the fact that X is independent to $Y+Z$. Now finally we take a look at the other side of the main inequality:

$$\begin{aligned} I(X+Y;X) &:= H(X+Y) - H(X+Y \mid X) \\ &\overset{(1)}{=} H(X+Y) - H(Y \mid X) \\ &\overset{(2)}{=} H(X+Y) - H(Y) \end{aligned}$$

Where step (1) comes as always, because given $X=x$, the map $X+Y \mapsto Y$ is injective, and step (2) comes from the fact that X and Y are independent. Now combining everything gives the second inequality. ♡

7 Entropy rates of a stochastic process

If we have a sequence of n random variables, a natural question is to ask how does the entropy of the sequence grow with n? This growth rate is what we call the entropy rate:

Definition 7.1 (Entropy rate) The entropy rate of a stochastic process $\mathbf{X} = (X_n : n \geq 0)$ is given by

$$H(\mathbf{X}) = \lim_{n \longrightarrow \infty} \frac{1}{n} H(X_1, \cdots, X_n)$$

Example 7.2 (Memoryless source) Let $\mathbf{X} = (X_n : n \geq 0)$ be a memoryless source. Then we have that $H(X_1, \cdots, X_n) = \sum_{i=1}^{n} H(X_i)$ by independence, but each of these summands is the same, say $H(X)$ because of identical distributions. Therefore $H(\mathbf{X}) = H(X)$

Example 7.3 (Ergodic Markov chain) Recall that a Markov chain $\mathbf{X} = (X_n)$ is ergodic if its irreducible and aperiodic. Let P be the transition matrix of the chain: $P(x, y) = \mathbf{P}(X_{n+1} = y \mid X_n = x)$. Suppose $X_0 \sim P_{X_0}$ is the initial distribution of the chain, and let π be the unique stationary distribution. Then we can compute

$$\begin{aligned} H(X_1^n) &= H(X_1) + \sum_{i=2}^{n} H(X_i \mid X_1^{i-1}) && \text{(Chain rule)} \\ &= H(X_1) + \sum_{i=2}^{n} H(X_i \mid X_{i-1}) && \text{(Markov property)} \\ &= H(X_1) - H(X_{n+1} \mid X_n) + \sum_{i=1}^{n} H(X_{i+1} \mid X_i) && \text{(Relabelling indices)} \end{aligned}$$

Now note that since $H(X \mid Y)$ is a continuous function of the PMFs of X and Y, we have that $H(X_n \mid X_{n-1}) \longrightarrow H(\bar{X}_1 \mid \bar{X}_0)$ where $\bar{X}_0 \sim \pi$ and $\bar{X}_1$ has the conditional distribution $\mathbf{P}(\bar{X}_1 = y \mid \bar{X}_0 = x) = P(x, y)$. Therefore we have that

$$\frac{H(X_1) - H(X_{n+1} \mid X_n)}{n} + \frac{1}{n} \sum_{i=1}^{n} H(X_{i+1} \mid X_i) \longrightarrow H(\bar{X}_1 \mid \bar{X}_0)$$

Due to the fact that the first term clearly goes to zero, and the second term converges by the Cesáro Lemma.

A general class of sources for which we can tell the entropy rate is that of Stationary Processes.

Definition 7.4 (Stationary process) A stochastic process is said to be stationary if the joint distribution of any subset of the sequence of random variables is invariant with respect to shifts in the time index, i.e:

$$\begin{aligned}&\mathbf{P}(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n)\\ &= \mathbf{P}(X_{1+l} = x_1, X_{2+l} = x_2, \cdots, X_{n+l} = x_n)\end{aligned}$$

For every n and every shift index $l \geq 0$.

Theorem 7.5 (Stationary sources have entropy rate) Let $\mathbf{X}$ be a stationary source on $\mathscr{A}$. Then the entropy rate exists and is equal to

$$H(\mathbf{X}) = \lim_{n \longrightarrow \infty} H(X_n \mid X_1^{n-1})$$

Main idea: Use Cesáro Lemma to show that the entropy rate converges if and only if $a_n = H(X_n \mid X_1^{n-1})$ converges, and then show this is a non-increasing sequence bounded below to show it converges.

Proof. We first note that by the chain rule:

$$\frac{1}{n} H(X_1^n) = \frac{1}{n} \sum_{i=1}^{n} H(X_i \mid X_1^{i-1})$$

So in light of the Cesáro Lemma, if we understand the convergence of $a_n = H(X_n \mid X_1^{n-1})$, then we will be done. First note obviously a_n is bounded below by zero, and intuitively, for progressively larger n, we are conditioning on more and more random variables, so somehow it makes sense that a_n would decrease. This is indeed what happens:

$$\begin{aligned}a_{n+1} = H(X_{n+1} \mid X_1^n) \leq H(X_{n+1} \mid X_2^n) \quad &\text{(Conditioning decreases } H\text{)}\\ = H(X_n \mid X_1^{n-1}) \quad &\text{(Stationarity)}\\ = a_n\end{aligned}$$

Thus (a_n) is non-increasing and bounded below, so it converges. ♡

We have shown that for a stationary source $\mathbf{X}$, its entropy rate is

$$H(\mathbf{X}) = \lim_{n \longrightarrow \infty} H(X_n \mid X_1^{n-1})$$

There turns out to be another interesting characterisation of this rate, which should seem true at an intuitive level due to stationarity. Suppose now that our source is indexed both in the present and past: $\mathbf{X} = (X_n : n \in \mathbf{Z})$. Then

Theorem 7.6 (Entropy rate and infinite past) For a stationary source $\mathbf{X}$, we have that

$$H(\mathbf{X}) = H(X_0 \mid X_{-\infty}^{-1})$$

The proof is skipped as it uses some big measure theoretic guns.

7.1 Shannon-McMillan-Breiman Theorem

We now present a general statement of the Asymptotic Equipartition Property for discrete stationary sources.

Definition 7.7 (Ergodic source) Let $\mathbf{X} = (X_n : n \in \mathbf{Z})$ on a finite alphabet $\mathscr{A}$ be stationary. Let $T : \mathscr{A}^{\mathbf{Z}} \longrightarrow \mathscr{A}^{\mathbf{Z}}$ be the left shift operator:

$$(T\mathbf{x})_n = x_{n+1}$$

The source $\mathbf{X}$ is said to be ergodic if all T-invariant measurable sets are trivial, i.e: if $T^{-1}B = B$ implies that $\mathbf{P}(\mathbf{X} \in B) = 0$ or 1.

Morally, a source is ergodic if if satisfies a general form of a Law of Large Numbers. This is made precise by the following Theorem:

Theorem 7.8 (Birkhoff's Ergodic Theorem) Let $\mathbf{X} = (X_n : n \in \mathbf{Z})$ be a stationary and ergodic source on a finite alphabet $\mathscr{A}$. Then for any measurable function $f : \mathscr{A}^{\mathbf{Z}} \longrightarrow \mathbf{R}$ with $\mathbf{E}[|f(\mathbf{X})|] < \infty$, one has that almost surely:

$$\frac{1}{n}\sum_{i=1}^{n} f(T^i\mathbf{X}) \longrightarrow \mathbf{E}[f(\mathbf{X})]$$

With this Theorem, of which we omit the proof, we are ready to state and prove the general version of the AEP:

Theorem 7.9 (Shannon-McMillan-Breiman) Let $\mathbf{X}$ be a stationary and ergodic source on a finite alphabet $\mathscr{A}$ with entropy rate $H = H(\mathbf{X})$, then

$$\frac{1}{n}\log P_n(X_1^n) \longrightarrow H \quad a.s$$

8 Types and large deviations

We now study a powerful tool in large deviation theory to calculate the possibility of rare events. In the AEP, we focused our attention on the small set of typical sequences, and this gave us some interesting results. Now we will focus on sets of sequences that have the same empirical distribution, and with this restriction, obtain bounds on the number of sequences in with a particular empirical distribution, and the probability of each of these sequences.

8.1 Method of types

For a fixed finite alphabet $\mathscr{A}$, let $\mathscr{P}(\mathscr{A})$ denote the set of all PMFs of $\mathscr{A}$, i.e:

$$\mathscr{P}(\mathscr{A}) = \left\{ P \in [0,1]^{\mathscr{A}} : \sum_{x \in \mathscr{A}} P(x) = 1 \right\}$$

The set $\mathscr{P}(\mathscr{A})$ can be identified with the probability simplex.

Definition 8.1 (Type) For a message X_1^n, we define its type to be its empirical PMF on the alphabet $\mathscr{A}$:

$$\widehat{P}_n(a) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(X_i = a)$$

Remark 8.2 Note that the type is indeed a PMF on the alphabet $\mathscr{A}$, indeed:

$$\sum_{a \in \mathscr{A}} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(x_i = a) = \frac{1}{n} \sum_{i=1}^{n} \underbrace{\sum_{a \in \mathscr{A}} \mathbf{1}\{x_i = a\}}_{=1} = 1.$$

Proposition 8.3 (Probability of an IID string, ♩) For any $x_1^n \in \mathscr{A}^n$ with type $\widehat{P}_n$ and any PMF Q, we have that

$$Q^n(x_1^n) = 2^{-n\left[H(\widehat{P}_n) + D(\widehat{P}_n \| Q)\right]}$$

Moreover, if $\widehat{P}_n = Q$, then clearly

$$Q^n(x_1^n) = 2^{-nH(Q)}$$

Main idea: It is clear that $\log Q^n(x_1^n) = \sum_{i=1}^n \log Q(x_i)$. The trick is that it is now possible to introduce out of this air the empirical PMF, by notnig that for any function H, $H(x_i) = \sum_{a \in \mathscr{A}} \mathbf{1}(a = x_i) H(a)$, and so, by multiplying and dividing by n, we can turn this into something that looks like an empirical PMF.

Remark 8.4 The reason why this Proposition is called Probability of an IID string, is that we are essentially saying that if $\{X_i\} \sim Q$ are IID, then $\mathbf{P}[X_1^n = x_1^n] = 2^{-n[H(\widehat{P}_n)+D(\widehat{P}_n\|Q)]}$. This is remarkable as it says that the probability of an IID string depends **only** through the type of the string!

Proof. We start by writing

$$\begin{aligned}
\log Q^n(x_1^n) &= \log \prod_{i=1}^n Q(x_i) \\
&= \sum_{i=1}^n \log Q(x_i) \\
&= \sum_{i=1}^n \sum_{a\in\mathscr{A}} \mathbf{1}(\{x_i\})(a) \log Q(a) \\
&= n \sum_{a\in\mathscr{A}} \underbrace{\frac{1}{n}\sum_{i=1}^n \mathbf{1}(\{x_i\})(a)}_{\widehat{P}_n(a)} \log\left(\frac{Q(a)\widehat{P}_n(a)}{\widehat{P}_n(a)}\right) \\
&= -n\left[H(\widehat{P}_n) + D(\widehat{P}_n\|Q)\right]
\end{aligned}$$

♡

Definition 8.5 (n-types) For a finite alphabet $\mathscr{A}$, the set of n-types, $\mathscr{P}_n(\mathscr{A})$ is defined to be

$$\mathscr{P}_n(\mathscr{A}) = \{P \in \mathscr{P}(\mathscr{A}) : nP(a) \in \mathbf{Z},\ \forall a \in \mathscr{A}\}$$

i.e: the set of empirical distributions with denominator n. Of course the type of a string x_1^n is an n-type.

Remark 8.6 (Size of $\mathscr{P}_n$) I.e: the set of n-types consists of PMFs on $\mathscr{A}$ that look like

$$P = \left(\tfrac{k_1}{n},\quad \tfrac{k_2}{n},\quad \cdots\quad, \tfrac{k_m}{n}\right)$$

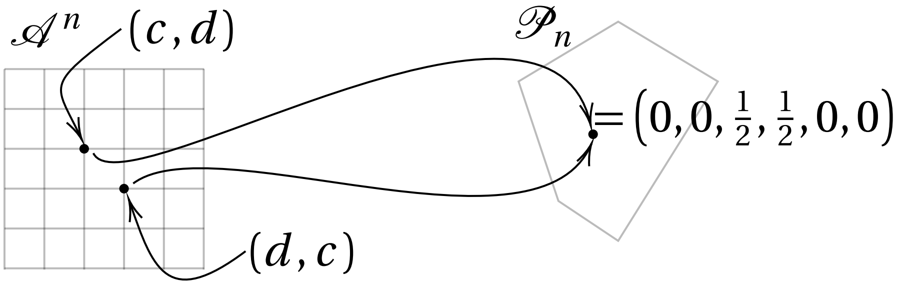
Where $m = |\mathscr{A}|$. Therefore a straightforward observation is that since for a given n-type P,since $k_i \in \{0,\cdots,n\}$, and there are m choices to make, we have that

$$|\mathscr{P}_n(\mathscr{A})| \leq (n+1)^m$$

Of course this is a brutal bound because we have to account for the fact that $k_1 + \cdots + k_m = n$.

Definition 8.7 (Type class) Given an n-type P, the type class $T(P)$ is the set of strings whose empirical distribution is P:

$$T(P) = \left\{ x_1^n \in \mathscr{A}^n : \widehat{P}_{x_1^n} = P \right\}$$



Since each string x_1^n on $\mathscr{A}^n$ gives rise to n-type, namely $\widehat{P}_{x_1^n}$, and each $T(P)$ is a subset of $\mathscr{A}^n$, it follows that

$$\mathscr{A}^n = \bigcup_{P \in \mathscr{P}_n(\mathscr{A})} T(P)$$

Moreover, we can give an explicit expression for the size $|T(P)|$, indeed, if $\mathscr{A} = \{a_1, a_2, \cdots, a_m\}$, then to describe a string $x_1^n \in T(P)$, we need to arrange $nP(a_1)$ lots of a_1, $nP(a_2)$ lots of a_2, $\cdots$, and $nP(a_m)$ lots of a_n, note that these quantities are indeed integers because P is an n-type. I.e: we have to sort n elements (the individual bits of the string) into m groups (the letters of the alphabet), the i^{th} of which has size $nP(a_i)$, this is described by the multinomial coefficient:

$$|T(P)| = \binom{n}{nP(a_1)\, nP(a_2) \cdots nP(a_m)} = \frac{n!}{\prod_{i=1}^m (nP(a_i))!}$$

Now we attempt to give some information-theoretic bounds on the size of $T(P)$.

Proposition 8.8 (Size of type class ♩) For a type $P \in \mathscr{P}_n$ over an alphabet of size m, we have that

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

To prove this we will need the following result:

Lemma 8.9 (P maximises $P^n(T(P))$, ♩) Let P be an n-type, then the most likely type class under P^n is $T(P)$, i.e:

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

Main idea: The key is that for a string $x_1^n \in T(P)$, its probability under P^n is very simply $\prod_{a\in\mathscr{A}} P(a)^{nP(a)}$, because each character a occurs $nP(a)$ times, and since the string is IID, we don't care about the order. This allows us to compute then the likelihood ratio between $P^n(T(P))$ and $P^n(T(P'))$

Proof. Let P and P' be two n-types, and write $\mathscr{A} = \{a_1, \cdots, a_m\}$. Consider an $x_1^n \in T(P)$, we have that

$$P^n(x_1^n) = P(x_1)P(x_2)\cdots P(x_n)$$

and since each $x_i \in \{a_1, \cdots, a_m\}$, we have that

$$P(x_1)P(x_2)\cdots P(x_n) = P(a_1)^{\#a_1} P(a_2)^{\#a_2} \cdots P(a_m)^{\#a_m}$$

where $\#a_i$ indicates the number of occurences of a_i in x_1^n. However the number of occurrences of each a_i in x_1^n is precisely $nP(a_i)$, this is because $x_1^n \in T(P)$. Therefore:

$$\begin{aligned}
P^n(T(P)) &= \sum_{x_1^n \in T(P)} P^n(x_1^n) \\
&= \sum_{x_1^n \in T(P)} \prod_{i=1}^{n} P(x_i) \\
&= \sum_{x_1^n \in T(P)} \prod_{j=1}^{m} P(a_j)^{nP(a_j)} \\
&= |T(P)| \prod_{j=1}^{m} P(a_j)^{nP(a_j)}
\end{aligned}$$

Therefore we have that the likelihood ratio

$$\frac{P^n(T(P))}{P^n(T(P'))} = \frac{|T(P)| \prod_{j=1}^m P(a_j)^{nP(a_j)}}{|T(P')| \prod_{j=1}^m P(a_j)^{nP'(a_j)}} = \prod_{j=1}^{n} \frac{(nP'(a_j))!}{(nP(a_j))!} P(a_j)^{n[P(a_j)-P'(a_j)]}$$

Where in the last equality, we have used the expression of the size of $T(P)$ and $T(P')$ we did before using multinomial coefficients. Now we use the inequality

$$k!/\ell! \geq \ell^{k-l}$$

Then we have that

$$\begin{aligned}\frac{P^n(T(P))}{P^n(T(P'))} &\geq \prod_{j=1}^{m} (nP(a_j))^{n[P'(a_j)-P(a_j)]} P(a_n)^{n[P(a_j)-P'(a_j)]} \\ &= \prod_{j=1}^{m} n^{n[P'(a_j)-P(a_j)]} \\ &= n^{n\sum_{j=1}^{m}[P'(a_j)-P(a_j)]} = n^0 = 1\end{aligned}$$

This establishes the proof. ♡

Now we are ready to prove Proposition 8.8.

Main idea: The upper bound is the standard counting argument, with the addition of the Probability of IID Strings Lemma, which tells us that if $x_1^n \in T(P)$, then $P^n(x_1^n) = 2^{-nH(P)}$. The lower bound comes from the observation we did a while ago that

$$\mathscr{A}^n = \bigcup_{Q\in\mathscr{P}_n} T(Q)$$

and using the Lemma which states that $P = \text{argmax}_{Q\in\mathscr{P}_n} P^n(T(Q))$

Proof of Proposition 8.8. Using Proposition 8.3, we have that

$$1 \geq P^n(T(P)) = \sum_{x_1^n \in T(P)} P^n(x_1^n) = \sum_{x_1^n} 2^{-nH(P)} = |T(P)| 2^{-nH(P)}$$

Which gives the upper bound. For the lower bound, we have that

$$\begin{aligned}1 = P^n(\mathscr{A}^n) = P^n\left(\bigcup_{P'\in\mathscr{P}_n} T(P')\right) &\leq \sum_{P'\in\mathscr{P}_n} P^n(T(P')) \\ &\leq |\mathscr{P}_n| \max_{P'\in\mathscr{P}_n} P^n(T(P')) \\ &= |\mathscr{P}_n| P^n(T(P)) \\ &\leq (n+1)^m |T(P)| 2^{-nH(P)}.\end{aligned}$$

Where in the last inequality we have used the observation that $|\mathscr{P}^n| \leq (n+1)^{|\mathscr{A}|}$ as well as Proposition 8.3. ♡

The size of type class proposition and Proposition 8.3 give a non-asymptotic version of the AEP: they explain that the strings in the set $T(P)$ (which correspond to the typical strings, because they have relative frequency equal to P) all have a uniform probability of 2^{-nH} and that there are approximately

2^{nH} of them. As an immediate corollary we have that

Corollary 8.10 (Probability of type class) For any type $P \in \mathscr{P}_n$ over an alphabet of size m, and any PMF Q on the alphabet, we have that

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

Main idea: Since all IID strings have probability determined uniquely by its type: $Q^n(T(P)) = |T(P)| 2^{-n[H(P)+D(P\|Q)]}$. Now using the bounds on $|T(P)|$ from Proposition Size of Type Class finishes the claim.

Before moving on, let us summarise the basic theorems concering types we have just proven. This suite of results *is* the method of types, and we will use it to obtain some results in Large Deviation Theory

$\lvert\mathscr{P}_n\rvert \le (n+1)^{\lvert\mathscr{A}\rvert}$	Bound on n types.
$Q^n(x_1^n) = 2^{-n(H(\widehat{P}_{x_1^n})+D(\widehat{P}_{x_1^n}\Vert Q))}$	Probability of IID strings.
$\lvert T(P)\rvert \approx 2^{nH(P)}$	Size of Type Class.
$Q^n(T(P)) \approx 2^{-nD(P\Vert Q)}$	Probability of Type Class.

8.2 Large Deviations

Example 8.11 Let $X_1, \cdots, X_n \overset{iid}{\sim} Q$ with $\mu = \mathbf{E}[X]$ its common mean. Then the probability that its empirical mean $S_n = \frac{1}{n}\sum_{i=1}^n X_i$ is ϵ further to the right of μ is of course going to decay to zero by the WLLN. But to obtain precise asymptotics one could bound the probability above by

$$\begin{aligned}\mathbf{P}(S_n \ge \mu+\epsilon) &= \mathbf{P}\left(e^{\lambda n S_n} \ge e^{n\lambda(\mu+\epsilon)}\right)\\ &\le e^{-n\lambda(\mu+\epsilon)}\mathbf{E}[e^{\lambda\sum X_i}]\\ &= \exp(-n[\lambda(\mu+\epsilon)-\Lambda(\lambda)])\end{aligned}$$

Where $\Lambda(t) = \ln\mathbf{E}\left[e^{tX}\right]$ is the log-moment generating function. We thus rewrite

$$\mathbf{P}(S_n \ge \mu+\epsilon) \le \exp(-n\Lambda^*(\mu+\epsilon))$$

Where $\Lambda^*(t) = \sup_{\lambda\ge 0}(\lambda t - \Lambda(\lambda))$. This gives one way of looking at the rate at which the probability decays. Another way we could have approached this is as follows:

The empirical averages S_n can be written as an expectation under a random empirical distribution, indeed:

$$S_n = \frac{1}{n}\sum_{i=1}^n f(X_i) = \sum_{a\in\mathscr{A}}\frac{1}{n}\sum_{i=1}^n \mathbf{1}(X_i = a)f(a) = \sum_{a\in\mathscr{A}}\widehat{P}_{X_1^n}(a)f(a)$$

Where $\widehat{P}_{X_1^n}$ is the random empirical measure on $\mathscr{A}$ induced by X_1^n. Therefore S_n being at least $\mu+\epsilon$ is an equivalent condition to the empirical distribution $P = \widehat{P}_{X_1^n}$ of X_1^n being such that $\mathbf{E}_{X\sim P}[f(X)]$ is at least $\mu+\epsilon$. So that if we let $E = \{P \in \mathscr{P} : \mathbf{E}_{X\sim P}[f(X)] \ge \mu+\epsilon\}$, then

$$\mathbf{P}[S_n \ge \mu+\epsilon] = Q^n(\{x_1^n : \widehat{P}_{x_1^n} \in E\})$$

8.2.1 Sanov's Theorem

From the discussion above, we saw that asking if for $X_1, \cdots, X_n \sim Q$ are IID, their empirical mean S_n exceeds $\mu+\epsilon$, is the same question as asking if an IID string X_1^n generated according to Q^n gives rise to an empirical distribution $\widehat{P}_{X_1^n}$ that belongs to a special set E of probability distributions. Sanov's Theorem asks this question in slightly more generality, attemptig to bound the probability that a random string $X_1^n \sim Q^n$ gives rise to an empirical PMF $\widehat{P}_{X_1^n}$ that belongs to a general class of PMFs. It will also give a very nice geometric picture of this question.

Remark 8.12 (A question on notation) Whenever I write seemingly nonsense things like $Q^n(\widehat{P}_n \in E)$, what I really mean is - given that Q^n induces a probability measure on $\mathscr{A}^n$ - chosen a random string $x_1^n \in \mathscr{A}^n$ according to the distribution Q^n, what is the probability that the empirical measure this string induces, $\widehat{P}_n \equiv \widehat{P}_{x_1^n}$ belongs to E. Alternatively, we can think just of X_1^n being a random variable with $\mathbf{P}[X_1^n \in B] = Q^n(B)$, and ask, what is the probability that the random empirical measure induced by X_1^n is in E. Hence:

$$Q^n(\widehat{P}_n \in E) := Q^n\{x_1^n : \widehat{P}_{x_1^n} \in E\} = \mathbf{P}\{\omega : \widehat{P}_{X_1^n(\omega)} \in E\}$$

Theorem 8.13 (Sanov's Theorem ☕) Let $X_1, \cdots, X_n \overset{iid}{\sim} Q$ where Q is of full support on a finite alphabet $\mathscr{A}$ of size m. Then for any $E \subseteq \mathscr{P}$

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

Moreover, if $E = \overline{E^\circ}$, then

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

Where P^* is the PMF that attains the infimum $\inf_{P \in E} D(P\|Q)$.

Main idea: To show the upper bound, we will employ the method of types, along the trivial expression

$$Q^n(\widehat{P}_n \in E) = \sum_{P \in E \cap \mathscr{P}_n} Q^n(T(P))$$

in particular, we will use the Probability of Type Class Lemma. Since Q is of full support and E is nice, $\inf_{P \in E} D(P\|Q)$ is attained by some P^*. This and the upper bound gives that

$$\lim_{n \longrightarrow \infty} -\frac{1}{n} \log Q^n(\widehat{P}_n \in E) \geq D(P^*\|Q)$$

To show the opposite inequality, consider a sequence $\{\mu_n\}$ with $\mu_n \in \mathscr{P}_n \cap E$, and $\mu_n \longrightarrow P^*$, and from the sum above, we see that $Q^n(\widehat{P} \in E) \geq Q^n(T(\mu_n))$. Now use the Probability of Type Class again and then send $n \longrightarrow \infty$.

Proof. To establish the upper bound, we first note that

$$Q^n(\widehat{P}_n \in E) = \sum_{P \in E \cap \mathscr{P}_n} Q^n(T(P)) \qquad (\star)$$

Indeed, $Q^n(\widehat{P}_n \in E)$ is the probability that a string in $\mathscr{A}^n$ gives rise to a type $\widehat{P}_n$ that belongs to E. Therefore to compute this probability we sum over all the probabilities of the type classes of each n-type in E. Therefore

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

Where in the second inequality we have used that $|\mathscr{P}_n| \leq (n+1)^m$ and $Q^n(T(P)) \leq 2^{-nD(P\|Q)}$ which was obtained in Corollary 8.10. To establish the second claim, we note that thinking of a PMF P as a vector on the probability simplex, we have that $D(P\|Q)$ is continuous in P since we have assumed Q is of full support. Therefore since E is closed, we have that the infimum is attained by some $P^* \in E$ which we fix throughout the proof. From this we can already see from $(\star\star)$ that

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

Now observe that the set $\bigcup_n \mathscr{P}_n$ of n-types is dense in $\mathscr{P}$ (the probability simplex), and hence also dense in E. Thus we can pick a sequence $\{\mu_n\}$ with $\mu_n \in \mathscr{P}_n \cap E$ (which is non-empty because E being the closure of its interior means its interior is non-empty, and since being dense means that every open set will be eventually intersected by $\mathscr{P}_n$, we can guarantee that this μ_n can exist ☕) such that $\mu_n \longrightarrow P^*$ and therefore $D(\mu_n\|Q) \longrightarrow D(P^*\|Q)$ as $n \longrightarrow \infty$.

From $(\star)$ and the fact that for all n, $P_n \in E$, we have that $Q^n(\widehat{P}_n \in E) \geq Q^n(T(P_n))$, So in particular, we also have that $Q^n(\widehat{P}_n \in E) \geq Q^n(T(\mu_n))$. Moreover, using again Corollary 8.10 we can further bound $Q^n(T(\mu_n)) \geq (n+1)^{-m} 2^{-nD(\mu_n\|Q)}$ This implies that

$$\limsup_{n \longrightarrow \infty} -\frac{1}{n} \log Q^n(\widehat{P}_n \in E) \leq \limsup_{n \longrightarrow \infty} D(\mu_n\|Q) = D(P^*\|Q)$$

♡

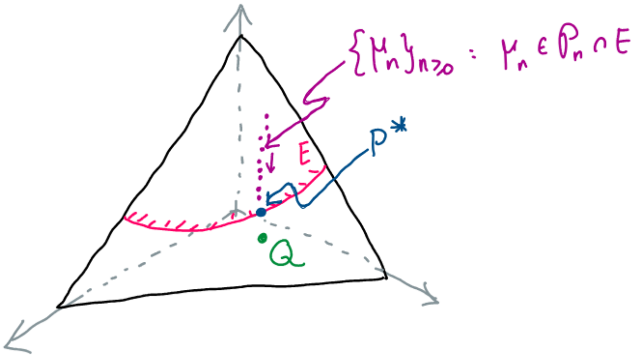


Figure 4: Proof by picture of the second part of Sanov's Theorem

Remark 8.14 (Topological Remark) Let us explain a bit more in detail why the requirement of E being the closure of its interior. Since $\bigcup_n \mathscr{P}_n$ is dense in $\mathscr{P}$, it means that for any PMF $P \in \mathscr{P}$, given any error tolerance $\epsilon > 0$, there will always be a PMF $\mu_n \in \mathscr{P}_n$ large enough so that $|\mu_n - P| < \epsilon$ (Here remember we are treting PMFs as vectors in $\mathbf{R}^n$). However, the PMF P in question is the minimiser of an infimum, so to guarantee the existence of the minimiser, which we labelled P^*, we needed E to be closed. Finally, note that we used the bound

$$Q^n(\widehat{P}_n \in E) \geq Q^n(T(\mu_n))$$

This came from the expression

$$Q^n(\widehat{P}_n \in E) = \sum_{P \in E \cap \mathscr{P}_n} Q(T(P))$$

So we had to guarantee that we could choose our μ_n to also belong to E. Let us show this. Let $\epsilon > 0$, we are going to show there exists some $\mu \in E \cap \mathscr{P}_n$ for some n large enough such that $|\mu - P^*| < \epsilon$. Recall the definition of closure, for a set S in a topological space X,

$$\text{cl}(S) = \left\{ x \in X : \text{ for } U \ni x \text{ open}, U \cap S \neq \varnothing \right\}$$

Therefore, since

$$E = \text{cl}(E^\circ)$$

and $P^* \in E$, we have that $\mathbb{B}_\epsilon(P^*) \cap E^\circ \neq \varnothing$. And since both $\mathbb{B}_\epsilon(P^*)$ and E° are open, then so is their intersection. Now by $\bigcup_n \mathscr{P}_n$ being dense, we have that for n large enough, we will intersect the open set $\mathbb{B}_\epsilon(P^*) \cap E^\circ$. Which will give us our desired μ.

Example 8.15 (Example 8.11 continued) Recall that for the event $B = \{S_n \geq \mu+\epsilon\}$, the Chernoff bound told us that

$$Q^n(B) \leq \exp(-n\Lambda^*(\mu+\epsilon))$$

Note that $E = \{P \in \mathscr{P} : \mathbf{E}_P[X] \geq \mu+\epsilon\}$ has that $E = \overline{E^\circ}$ so by Sanov's Theorem, $Q^n(B) \leq (n+1)^m e^{-nD_e(P^*\|Q)}$ and moreover $D_e(P^*\|Q)$ is the asymptotically right exponent. We must then have that $\Lambda^*(\mu+\epsilon) \leq D_e(P^*\|Q)$.

In fact these two quantities are equal:

Proposition 8.16 (Decay exponent and Sanov's Theorem, ☕) Let $X_1^n \overset{iid}{\sim} Q$ on an alphabet $\mathscr{A}$. For events be of the form

$$B = \left\{ \frac{1}{n}\sum_{i=1}^{n} f(X_i) \geq \mu+\epsilon \right\}.$$

we have that

$$\Lambda^*(\mu+\epsilon) = \inf_{P\in E} D_e(P\|Q) =: D_e(P^*\|Q)$$

Where recall that

$$E = \left\{P \in \mathscr{P} : \mathbf{E}_P[f(X)] \geq \mu+\epsilon\right\}$$

Main idea: Compute the supremum in the definition of $\Lambda^*(\mu+\epsilon)$. To do so, you will show that $\Lambda'(0) = \mu$, and $\lim_{x\longrightarrow\infty}\Lambda'(x) = f_{\text{max}} > \mu+\epsilon$. Moreover, show $\Lambda''(x) \geq 0$. Therefore, we have that $\Lambda(\lambda^*) = \mu+\epsilon$ for some λ^*. Now consider the PMF

$$P_\lambda(a) = \frac{\exp(\lambda f(a))}{\mathbf{E}[\exp(\lambda f(X_1))]} Q(a)$$

From this you will show that $P_{\lambda^*} \in E$, and as such $D_e(P^*\|Q) \leq D_e(P_{\lambda^*}\|Q)$ and by computing this quantity you will reach that $D_e(P_{\lambda^*}\|Q) = \Lambda^*(\mu+\epsilon)$ as required.

Proof. The only thing remaining to be shown is that $\Lambda^*(\mu+\epsilon) \geq D_e(P^*\|Q)$. We first compute the supremum in the definition of $\Lambda^*(x)$. Define the PMF

$$P_\lambda(a) = \frac{\exp(\lambda f(a))}{\mathbf{E}[\exp(\lambda f(X_1))]} Q(a)$$

Then

$$\Lambda'(\lambda) = \frac{\mathbf{E}[f(X_1)e^{\lambda f(X_1)}]}{\mathbf{E}[e^{\lambda f(X_1)}]} = \mathbf{E}_{Y\sim P_\lambda}[f(Y)]$$

By taking another derivative it's pretty easy to see that

$$\Lambda''(\lambda) = \mathrm{Var}_{Y\sim P_\lambda}(f(Y)) \geq 0$$

Therefore $\Lambda'(\lambda)$ is non decreasing from from $\Lambda'(0) = \mu$ to

$$\lim_{\lambda\longrightarrow\infty} \Lambda'(\lambda) = \lim_{\lambda\longrightarrow\infty} \sum_x \frac{Q(x)f(x)}{\sum_y Q(y)\exp\{\lambda[f(y)-f(x)]\}} \longrightarrow f_{\mathsf{max}}$$

Because whenever $x \neq \mathrm{argmax}_x f(x)$, the denominator will contain an exponential with a positive exponent and so the limit will kill the entire fraction. Therefore, since $\mu < \mu+\epsilon < f_{\mathsf{max}}$, there must be some $\lambda^* > 0$ for which $\Lambda'(\lambda^*) = \mu+\epsilon$, as Λ' is continuous. Recalling that

$$\Lambda^*(x) := \sup_{\lambda\geq 0}\{\lambda x - \Lambda(x)\}$$

We see that we can compute this supremum by taking derivatives and setting to zero, i.e: $\Lambda'(x) = \lambda$, and since $\Lambda'' \geq 0$, this will be indeed a maximum of $\lambda x - \Lambda(x)$. Thus we see that $\Lambda^*(\mu+\epsilon) = \lambda^*$. Now recall that $\mathbf{E}_{Y\sim P_{\lambda^*}}[f(Y)] = \Lambda'(\lambda^*) = \mu+\epsilon$, so $P_{\lambda^*} \in E$, and so

$$\begin{aligned}
D_e(P^*\|Q) &\leq D_e(P_{\lambda^*}\|Q) \\
&= \mathbf{E}_{Y\sim P_{\lambda^*}}\left[\log\frac{P_{\lambda^*}(Y)}{Q(Y)}\right] \\
&= \mathbf{E}_{Y\sim P_{\lambda^*}}\left[\log\frac{e^{\lambda^* f(Y)}}{\mathbf{E}[e^{\lambda^* f(X_1)}]}\right] \\
&= \lambda^*\mathbf{E}_{Y\sim P_{\lambda^*}}[f(Y)] - \Lambda(\lambda^*) \\
&= \Lambda^*(\mu+\epsilon)
\end{aligned}$$

as required. ♡

Incidentally we have also found the PMF P^* that minimises relative entropy.

Corollary 8.17 Let X_1^n be IID with common PMF Q on $\mathscr{A}$. For events B such as the one in Proposition 8.16, we have that the PMF P^* that minimises relative entropy, $P^* = \inf_{P\in E} D(P\|Q)$, is equal to

$$P^*(a) = \frac{\exp(\lambda^* f(a))}{\mathbf{E}[\exp(\lambda^* f(X_1))]}Q(a)$$

Where $\lambda^* > 0$ is chosen so that $\mathbf{E}_{Y\sim P_{\lambda^*}}[f(Y)] = \mu+\epsilon$.

Proof. From Example 8.15 we know that

$$\Lambda^*(\mu+\epsilon) \leq D(P^*\|Q)$$

But in Proposition 8.16, we have found that

$$\Lambda^*(\mu+\epsilon) = D(P_{\lambda^*}\|Q)$$

which means that $D(P_{\lambda^*}\|Q) \leq D(P^*\|Q)$. But since by assumption $\mathbf{E}_{Y\sim P_{\lambda^*}}[f(Y)] = \mu+\epsilon$, it follows that $P_{\lambda^*} \in E$ and so by definition of P^*, it must be that $D(P_{\lambda^*}\|Q) \geq D(P^*\|Q)$, and as such both relative entropies are equal. However, since $D(P\|Q)$ is strictly convex, and E is nonempty, convex and closed, we have that the minimiser is unique. (See Proposition A.6) ♡

Example 8.18 (An example on using Sanov's Theorem) Here's a neat example of how to use Sanov's Theorem to make a very clever counting argument. Let $\mathscr{A} = \{1,2,\cdots,m\}$ and $g : \mathscr{A} \longrightarrow \mathbf{R}$ be a function. Fix some $\alpha \in \mathbf{R}$. Show that

$$\#\left\{x_1^n : \frac{1}{n}\sum_{i=1}^n g(x_i) \geq \alpha\right\} \approx 2^{nH^*} \qquad H^* = \sup_{P:\sum_j P(j)g(j)\geq\alpha} H(P)$$

Solution. The clever ideas are as follows: first, the constraint we have on our string can be expressed in terms of an expectation with respect to an empirical measure. Indeed:

$$\begin{aligned}
\mathbf{E}_{\widehat{P}_{x_1^n}}[g] &:= \sum_x g(x)\widehat{P}_{x_1^n}(x) \\
&= \sum_x g(x)\frac{1}{n}\sum_{i=1}^n \mathbf{1}(x_i = x) \\
&= \sum_{i=1}^n \frac{1}{n}\sum_x g(x)\mathbf{1}(x_i = x) \\
&= \sum_{i=1}^n \frac{1}{n} g(x_i)
\end{aligned}$$

So if we let

$$E = \{P \in \mathscr{P} : \mathbf{E}_P[g] \geq \alpha\}$$

Then it is clear by the above calculation that the set of strings we are interested in, is precisely the set $\{x_1^n : \widehat{P}_{x_1^n} \in E\}$. To count the number of strings in this set, we use the second clever idea, which is to introduce uniform random variables. In particular, by setting $X_1,\cdots,X_n \sim \text{Unif}\{1,\cdots,m\}$, we have that

$$\mathbf{P}[\widehat{P}_{X_1^n} \in E] = \frac{\#\{x_1^n : \widehat{P}_{x_1^n} \in E\}}{m^n}$$

Because the numerator is the number of favorable strings and the denominator is the total number of strings. Therefore, all we need to do is estimate the probability $\mathbf{P}[\widehat{P}_{X_1^n} \in E] = Q^n(\{x_1^n : \widehat{P}_{x_1^n} \in E\})$, and since it can be checked that E is the closure of its interior, we can employ Sanov's Theorem,

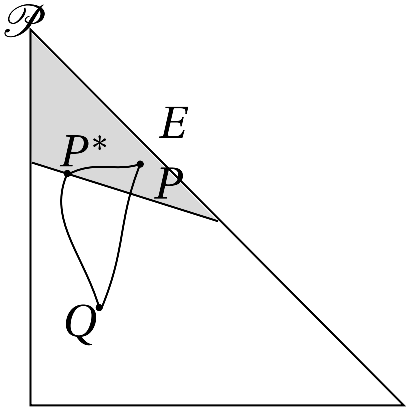


Figure 5: Illustration of Pythagorean Theorem

to assert that $Q(E) \approx 2^{-n \inf_{P\in E} D(P\|Q)}$. It follows that we are just left to compute this infimum. But since Q is uniform, this is a really easy task because

$$D(P\|Q) = \sum_{x=1}^{m} P(x) \log \frac{P(x)}{Q(x)} = -H(P) + \log(m)$$

It now follows that $Q^n(E) \approx 2^{-n(\inf_{P\in E} -H(P)\log(m))} = m^{-n} 2^{nH^*}$ Which finishes our claim.

♡

Theorem 8.19 (Pythagorean identity for relative entropy) Suppose $E \subseteq \mathscr{P}$ is closed and convex, take $Q \notin E$ and assume Q has full support, and let P^* achieve the minimum in Sanov's Theorem, i.e:

$$D(P^*\|Q) = \inf_{P\in E} D(P\|Q)$$

Then:

$$D(P\|Q) \geq D(P\|P^*) + D(P^*\|Q) \quad \text{for all } P \in E$$

Remark 8.20 (Remark and waffle) Just for completeness, we recall that by the strict convexity of $D(P\|Q)$ as well as its continuity in P, and the fact that E is convex and closed, then P^* exists and is unique. Before proving it, w note that the Pythagorean identity quantifies the following statement. If $D(P^*\|Q)$ is close to $D(P\|Q)$, then P is necessarily close to P^*, i.e: *if a PMF is close to minimising the "distance", then it must be close to the minimiser.*

Main idea: Consider the function $D(\lambda) = D(P_\lambda\|Q)$, where $P_\lambda = \lambda P + (1-\lambda)P^*$. Since P_λ is in E by convexity, and P^* is the minimiser, it follows that $D(P_\lambda\|Q) \geq D(P^*\|Q) = D(P_0\|Q)$, and so $\partial_\lambda D(\lambda) \geq 0$. Now compute this derivative and everything comes out nicely.

Proof. We show the result for D_e, and then just multiply by $\log_2 e$ throughout. Let $P \in E$ be arbitrary, and consider the convex combination

$$P_\lambda = \lambda P + (1-\lambda)P^*$$

By convexity of E, $P_\lambda \in E$ for all $\lambda \in [0,1]$, and by definition of P^*,

$$D(P_\lambda \| Q) \geq D(P^* \| Q) = D(P_0 \| Q) \geq 0.$$

Therefore the derivative of $D(P_\lambda \| Q)$ at zero, if it exists, must be non-negative (indeed think of the defining limit of the derivative at zero plus). I.e:

$$\begin{aligned}
0 \leq \frac{d}{d\lambda} D_e(P_\lambda \| Q)\Big|_{\lambda=0^+} &= \frac{d}{d\lambda} \sum_{a \in \mathscr{A}} P_\lambda(a) \log_e \frac{P_\lambda(a)}{Q(a)}\Big|_{\lambda=0^+} \\
&= \sum_{a \in \mathscr{A}} [P(a) - P^*(a)] \log_e \frac{P_\lambda(a)}{Q(a)}\Big|_{\lambda=0^+} + \underbrace{\sum_{a \in \mathscr{A}} [P(a) - P^*(a)]}_{0} \\
&= \sum_{a \in \mathscr{A}} P(a) \log_e \left(\frac{P^*(a)}{Q(a)} \frac{P(a)}{P(a)} \right) - \sum_{a \in \mathscr{A}} P^*(a) \log_e \frac{P^*(a)}{Q(a)} \\
&= D_e(P\|Q) - D_e(P\|P^*) - D_e(P^*\|Q)
\end{aligned}$$

♡

8.3 The Gibb's conditioning principle

Sanov's Theorem told us the following: if $E \subseteq \mathscr{P}$ encodes a rare event, then if $X_1, \cdots, X_n \sim Q$ are IID, the probability that their empirical measure $\widehat{P}_n$ belonged to E was approximately for large n, $2^{-nD(P^*\|Q)}$, where P^* minimised the relative entropy between Q and E. Now we will use some stronger argument to show that actually the probability of E is essentially the same as the probability of the type P^*, and that all other types that are not close to P^* contribute a negligible amount of mass to E. Thus explaining that if a rare observation is seen, then among the rare observations, the least rare is the most likely to be observed: rare things happen in the least rare way possible.

Theorem 8.21 (Gibb's conditioning principle) Let $\widehat{P}_n$ be the type of IID observations $X_1, \cdots, X_n$ with PMF Q which has full support on A. Suppose that $E \subseteq \mathscr{P}$ is closed, convex, and has non-empty interior, and does not contain Q. Then

$$\mathbf{E}\big[\widehat{P}_n(a) \mid \widehat{P}_n \in E\big] = \mathbf{P}(X_1 = a \mid \widehat{P}_n \in E) \longrightarrow P^*(a)$$

Where

$$P^* = \text{argmin}_{P \in E} D(P \| Q)$$

Main idea: The first equality is just by definition of the empirical PMF. To show that $\mathbf{E}[\widehat{P}_n(a) \mid \{\widehat{P}_n \in E\}] \longrightarrow P^*(a)$, we aim to show that conditional on $\{\widehat{P}_n \in E\}$, $\widehat{P}_n(a) \longrightarrow P^*(a)$ in probability. To do this, we show that for any δ, $Q^n(D(\widehat{P}_n \| P^*) > 2\delta) \longrightarrow 0$, and then use Pinsker's inequality. To show this convergence, define appropriate relative entropy neighborhoods and use the Pythagorean inequality, which remember, said that if $D(P\|Q) \leq D(P^*\|Q) + 2\delta$, then $D(P\|P^*) \leq 2\delta$.

Proof. It is easy to see that

$$\begin{aligned}
\mathbf{E}\big[\widehat{P}_n(a) \mid \widehat{P}_n \in E\big] &:= \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}[\mathbf{1}(X_i = a) \mid \widehat{P}_n \in E] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbf{P}(X_i = a \mid \widehat{P}_n \in E) \\
&= \mathbf{P}(X_1 = a \mid \widehat{P}_n \in E)
\end{aligned}$$

Where the last step comes from the fact that each X_i still has the same distribution. And if $X = Y$ in distribution, then for events A, B, we also have that $\mathbf{P}[X \in A \mid B] = \mathbf{P}[Y \in A \mid B]$ by simply expanding the definition of conditional probability). Now to show that $\mathbf{P}(X_1 = a \mid \widehat{P}_n \in E) \longrightarrow P^*(a)$, we will in fact show that as distributions, given that $\{\widehat{P}_n \in E\}$ holds, $\widehat{P}_n$ and P^* are close. We will do this by showing they are close in the sense of relative entropy, which by Pinsker's inequality will mean they are close in the sense of Total Variation, and from there the claim will follow.

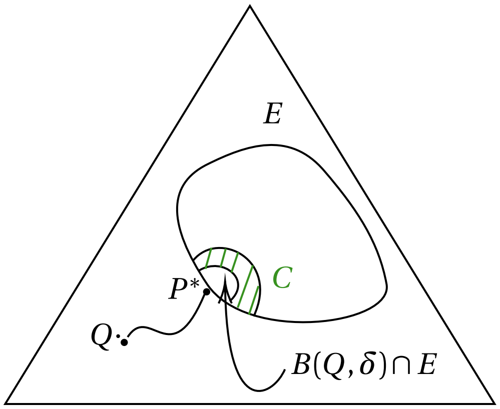
Define the relative entropy neighborhood

$$B(Q, \delta) := \{P \in \mathscr{P} : D(P\|Q) \leq D(P^*\|Q) + \delta\}$$

And define the subsets

$$C = B(Q, 2\delta) \cap E \qquad D = E \setminus C$$

The picture to have in mind is something like this:



It is clear that the goal is to show that for $\delta \ll 1$, we have that

$$Q^n(\widehat{P}_n \in C \mid \widehat{P}_n \in E) \approx 1$$

Which is equivalent to showing that

$$Q^n(\widehat{P}_n \in D \mid \widehat{P}_n \in E) \approx 0$$

We begin by noting that

$$\begin{aligned} Q^n(\widehat{P}_n \in D \mid \widehat{P}_n \in E) &:= \frac{Q^n(P^n \in D \cap E)}{Q^n(\widehat{P}_n \in E)} \\ &\leq \frac{Q^n(P^n \in D)}{Q^n(\widehat{P}_n \in E)} \\ &\leq \frac{Q^n(P^n \in D)}{Q^n(\widehat{P}_n \in C)} \end{aligned}$$

By definition of D being the set of PMFs in E whose relative entropy distance from Q is further than $2\delta + D(P^*\|Q)$, we have that, in light of Sanov's Theorem:

$$Q^n(\widehat{P}_n \in D) \leq (n+1)^m 2^{-n \inf_{P\in D} D(P\|Q)} \leq (n+1)^m 2^{-n[D(P^*\|Q)+2\delta]}$$

(Indeed, $\inf_{P\in D} D(P\|Q) \leq 2\delta + D(P^*\|Q)$ and so taking minus signs and taking powers gives the last inequality above) To lower bound the denominator, we note that since $\bigcup_{n\geq 1}\mathscr{P}_n$ is dense in $\mathscr{P}$, every open set in $\mathscr{P}$ will eventually be intersected by $\mathscr{P}_n$, and since E has nonempty interior, then $B(Q,\delta)\cap E\cap\mathscr{P}_n$ will eventually be non-empty.Let P_n be one such PMF. Then

$$\begin{aligned} Q^n(\widehat{P}_n\in C) &\overset{(1)}{\geq} Q^n(\widehat{P}_n\in B(Q,\delta)\cap E)) \\ &\overset{(2)}{\geq} Q^n(T(P_n)) \\ &\overset{(3)}{\geq} (n+1)^{-m}2^{-nD(P_n\|Q)} \\ &\overset{(4)}{\geq} (n+1)^{-m}2-n^{[D(P^*\|Q)+\delta]} \end{aligned}$$

where (1) comes from the fact that $C := B(Q,2\delta)\cap E\supset B(Q,\delta)\cap E$, (2) comes from the fact that if a string $x_1^n\in T(P_n)$, then since $P_n\in B(Q,\delta)\cap E\cap\mathscr{P}_n$, then the type of this string $\widehat{P}_n$ also belongs to $B(Q,\delta)\cap E$. Step (3) comes from the fact that the probability of a type class can be bounded below by Corollary 8.10. Step (4) holds because $P_n\in B(Q,\delta)$, so its distance to Q is at most $D(P^*\|Q)+\delta$. Combining with the upper bound we obtained on $Q^n(\widehat{P}_n\in D)$ gives that:

$$Q^n(\widehat{P_n}\in D\mid\widehat{P}_n\in E)\leq(n+1)^{2m}2^{-n\delta}\longrightarrow 0\qquad(n\longrightarrow\infty)$$

Therefore, by definition of the set C, we have that with very high probability,

$$D(\widehat{P}_n\|Q)\leq D(P^*\|Q)+2\delta.$$

Note that for any $P\in C$, we have that

$$\begin{aligned} D(P^*\|Q)+2\delta &\geq D(P\|Q) \\ &\geq D(P\|P^*)+D(P^*\|Q) \end{aligned}$$

where the first inequality comes from the definition of being in C, and the second inequality is the Pythagorean inequality. I.e: if $P\in C$, then $D(P\|P^*)\leq 2\delta$. So

$$Q^n\left(D(\widehat{P}_n\|P^*)\leq 2\delta|\widehat{P}_n\in E\right)\geq Q^n\left(\widehat{P}_n\in C|\widehat{P}_n\in E\right)\longrightarrow 1$$

Therefore by Pinsker's Inequality, if for some $\epsilon>0$ we have that $\left\|\widehat{P}_n-P^*\right\|_{\mathsf{TV}}>\epsilon$ then it must also be the case that $D(\widehat{P}_n\|P^*)\gtrsim\epsilon^2$ but since δ was arbitrary, combining with the above gives that for any $\epsilon>0$,

$$Q^n\left(\left\|\widehat{P}_n-P^*\right\|_{\mathsf{TV}}>\epsilon|\widehat{P}_n\in E\right)\longrightarrow 0$$

which in particular implies that for any $a\in A$,

$$Q^n(|\widehat{P}_n(a)-P^*(a)|>\epsilon|\widehat{P}_n\in E)\longrightarrow 0$$

which means that conditional on $\{\widehat{P}_n\in E\}$, the random variables $\widehat{P}_n(a)$ converge to $P^*(a)$ in probability. Since they are bounded by 1, the bounded convergence theorem implies they also converge in L^1, and so we get the desired result, namely that

$$\mathbf{E}[\widehat{P}_n(a) \,|\, \{\widehat{P}_n \in E\}] \longrightarrow P^*(a)$$

♡

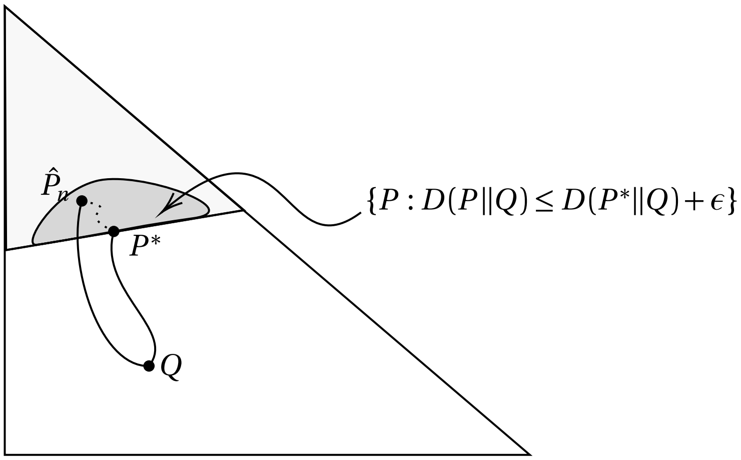


Figure 6: An illustration of what happens in Gibbs' Theorem: if you fix an ϵ tolerance about P^* in the $D(P\|Q)$ sense, it will happen with high probability that conditional on $\widehat{P}_n \in E$, for n large enough, the empirical measure will fall within the specified tolerances.

Example 8.22 (On using Gibbs' conditioning principle) Let us showcase how to use this in a practical example. Suppose $X_1, \cdots, X_n$ are distributed according to a Bin(m, q) where $q \in (0,1)$. For some $\alpha \in (mq, m)$, we have that

$$\lim_{n \longrightarrow \infty} \mathbf{P}\left[X_1 = k \left| \frac{1}{n} \sum_{i=1}^{n} X_i \geq \alpha \right. \right] = P^*[k]$$

Where $P^* \sim$ Bin(m, λ) for some special value of λ which we will also determine.

Solution. First we make the remark that we really need the assumption that $\alpha > mq$, so that $\alpha = \mu + \epsilon$ and this is a rare event in the form of the Gibb's conditioning principle. By Gibb's conditioning principle, we have that

$$P^*[k] \propto \exp(\eta k) Q(k)$$

where η has to be chosen so that

$$\mathbf{E}_{P^*}[X_1] = \mu + \epsilon = \alpha$$

If we expand the expression above for $P^*[k]$, we obtain that

$$\begin{aligned}P^*[k] &\propto \exp(\eta k)\binom{m}{k}q^k(1-q)^{m-k}\\ &= (1-q)^m\binom{m}{k}\left(\frac{\exp(\eta)q}{1-q}\right)^k\\ &\propto \binom{m}{k}\left(\frac{\exp(\eta)q}{1-q}\right)^k\\ &:= \binom{m}{k}\left(\frac{\lambda}{1-\lambda}\right)^k\end{aligned}$$

Where λ is chosen so that

$$\frac{\exp(\eta)q}{1-q} = \frac{\lambda}{1-\lambda}$$

Now we note that with this choice of λ, we obtained that $P^*[k] \propto \binom{m}{k}\lambda^k(1-\lambda)^k$, which is in the form of $P^* \sim \text{Bin}(m,\lambda)$. All that's left to do is choose η so that $\mathbf{E}_{P*}[X_1] = \alpha$, but since P^* is binomial, this expectation is nothing but $m\lambda$, and so we have that $\lambda = \frac{\alpha}{m}$. Now one can choose η appropriately so that everything checks out. ♡

Example 8.23 (On using Gibbs' conditioning principle for rare events) Suppose a dice is rolled 10^{10} times and a proportion of 6's was observed $\geq 25\%$ of the time. Given that this event has occurred, what is the average result of the dice?

Solution. Let $X_1, X_2, \cdots, X_n$ be IID Unif$\{1,2,3,4,5,6\}$. We begin by defining our rare event in terms of empirical measures:

$$E = \left\{ P \in \mathscr{P} : \mathbf{E}_P[\mathbf{1}(X_1 = 6)] \geq \alpha = \frac{1}{4} \right\}$$

Then by Gibbs' conditioning principle, we are interested in computing $\mathbf{E}_{P*}[X]$, where

$$P^*[\cdot] = \mathbf{P}\big[X_1 \in \cdot \mid \widehat{P}_{X_1^n} \in E\big]$$

Since we have rolled the dice a very large number of times, we have that approximately:

$$P^*[k] \propto \exp\big(\eta\,\mathbf{1}(k=6)\big)$$

Therefore $P^*[k] \approx (q,q,q,q,q,1-5q)$. And to find q, we have to impose the constraint on η, namely that

$$\frac{1}{4} = \alpha = \mathbf{E}_{P*}[f(X_1)] = 1 - 5q$$

Therefore $q = \frac{3}{20}$, and now one can calculate that $\mathbf{E}_{P*}[X] = 3.75$. ♡

Remark 8.24 Let us appreciate for a second Gibbs' conditioning principle. If one wanted to simulate this event, you would be in deep shit, because if you simulated 10^{10} dice throws, it would almost never be the case that you would observe a proportion of 6's greater than $1/4$. Therefore by simulation it would be a completely unfeasible task to simulate $\mathbf{E}_{P*}[X]$.

8.4 Error probability in fixed-rate data compression

Recall in the earlier sections our discussion about fixed-rate data compression: we had a source $\mathbf{X} = \{X_n\}$ and we established that data compression using codebooks, which was nothing else than using a collection $\{B_n\}$ of strings of length n to compress our data, could achieve a compression rate of H whilst having the error probability go to zero, but any codebook that has an error probability going to zero, that is to say, the probability that a string x_1^n does not belong to the codebook, goes to zero, cannot have a better rate than H, thus in some sense, H was the optimal rate one could achieve. We now pose the following question, suppose a codebook with a given rate $R > H$ has decaying probability of error $P_e^{(n)}$. How fast can this probability of error decay? It turns out that at best the decay will be approximately of

$$P_e^{(n)} \approx 2^{-nD^*} \quad \text{where} \quad D^* = \inf_{H(P)\geq R} D(P\|Q)$$

Theorem 8.25 (Error exponent for fixed-rate compression) Let $\mathbf{X} = \{X_n\}$ be a memoryless source with entropy H and distribution Q which as full support on a finite alphabet $\mathscr{A}$. Let R be any rate with $H < R < \log|\mathscr{A}|$ (This latter inequality is just a pedantic remark, because if we take the rate to be $\log|\mathscr{A}|$ then there is no compression at all and so we have 0 probability of error). Then

- There is a fixed-rate code $\{B_n^*\}$ with asymptotic rate no more than R bits/symbol,i.e:

  $$\limsup_{n\longrightarrow\infty} \frac{1}{n}\left(1 + \lceil \log|B_n^*| \rceil\right) = \limsup_{n\longrightarrow\infty} \frac{1}{n} \log|B_n^*| \leq R$$

  whose probability of error decays at least as fast than exponentially with exponent D^*, i.e:

  $$\limsup_{n\longrightarrow\infty} \frac{1}{n} \log P_e^{(n)} \leq -D^*$$

- Conversely, if $\{B_n\}$ is any fixed-rate code with asymptotic rate no more than R bits/symbol, i.e:

  $$\limsup_{n\longrightarrow\infty} \frac{1}{n}\left(1 + \lceil \log|B_n| \rceil\right) = \limsup_{n\longrightarrow\infty} \frac{1}{n} \log|B_n| \leq R$$

  Then its probability of error cannot decay faster than exponentially exponentially with exponent D^*, i.e:

  $$\liminf_{n\longrightarrow\infty} \frac{1}{n} \log P_e^{(n)} \geq -D^*$$

  Where $D^* = \inf_{H(P)\geq R} D(P\|Q)$.

Proof. For the direct part, we have to construct said codebook, and our candidate will be

$$B_n^* = \bigcup_{P\in\mathscr{P}_n : H(P)<R} T(P)$$

i.e: all strings whose type is a PMF with entropy less than R. Understanding the asymptotic compression rate of this codebook is quite straightforward, simply:

$$\begin{aligned}
|B_n^*| &\overset{(1)}{\leq} |\mathscr{P}_n| \max_{P\in\mathscr{P}_n : H(P)<R} |T(P)| \\
&\overset{(2)}{\leq} (n+1)^m \max_{P\in\mathscr{P}_n : H(P)<R} |T(P)| \\
&\overset{(3)}{\leq} (n+1)^m \max_{P\in\mathscr{P}_n : H(P)<R} 2^{nH(P)} \\
&\overset{(4)}{\leq} (n+1)^m 2^{nR}
\end{aligned}$$

Where (1) comes from a simple bound on the size of a union of sets, (2) comes from the fact that to determine an n-type we need to make choose among $n+1$ values for each of the m elements of the alphabet, (3) comes from the bound in $|T(P)| \leq 2^{nH(P)}$ from Proposition size of Type Class (Proposition 8.8) and the final step, (4) comes from the fact that the PMFs we are taking the maximum over all have entropies less than R. Therefore it now follows immediately that

$$\limsup_{n\longrightarrow\infty} \frac{1}{n} \log |B_n^*| \leq R.$$

We have now constructed a codebook with the desired rate. We must now show that the decay of the error probability of this codebook is as fast as exponential with exponent D^*. To do this we explicitly compute the error probability:

$$\begin{aligned}
P_e^{(n)} &:= \mathbf{P}(X_1^n \notin B_n^*) \\
&= \mathbf{P}\left(X_1^n \in \bigcup_{P\in\mathscr{P}_n : H(P)\geq R} T(P) \right) \\
&= \sum_{P\in\mathscr{P}_n : H(P)\geq R} Q^n(T(P)) \\
&\overset{(a)}{=} \sum_{P\in\mathscr{P}_n : H(P)\geq R} 2^{-nD(P\|Q)} \\
&\overset{(b)}{\leq} |\mathscr{P}_n| 2^{-nD^*} \leq (n+1)^m 2^{-nD^*}
\end{aligned}$$

The first three steps are elementary, and step (a) is due to the bound $Q^n(T(P)) \leq 2^{-nD(P\|Q)}$ found in Proposition 8.10, and (b) comes from a simple size bound along the fact that any $P \in \mathscr{P}_n$ with $H(P) \geq R$ will have that $D(P\|Q) \geq \inf_{P\in\mathscr{P}_n : H(P)\geq R} D(P\|Q) := D^*$. Putting this all together shows that

$$\limsup_{n\longrightarrow\infty} \frac{1}{n} \log P_e^{(n)} \leq -D^*.$$

To show the converse statement, we will show that for any codebook $\{B_n\}$ with rate no more than

R and any $\epsilon > 0$, one has that:

$$\liminf_{n\longrightarrow\infty} \frac{1}{n}\log P_e^{(n)} \geq -(D^* + 2\epsilon)$$

Hence, let $\epsilon > 0$ be given, we will make the following rather long-winded but necessary observation:

- Stare at the definition $D^* := \inf_{H(P)\geq R} D(P\|Q)$. Since relative entropy is continuous in P, if we slightly shrink the set that we are taking the infimum over, then this infimum will only increase slightly. In precise terms, for the ϵ we have been given, there will be a small enough δ such that

  $$\inf_{H(P)\geq R+\delta} D(P\|Q) \leq D^* + \epsilon. \qquad (\star)$$

  Now note that by taking δ even smaller if needed so that $R + \delta/2$ is still less than $\log|\mathscr{A}|$, the set $\{P \in \mathscr{P} : H(P) \geq R + \delta/2\}$ has nonempty interior, so since the n-types $\{\mathscr{P}_n : n \geq 1\}$ are dense in $\mathscr{P}$, for all n large enough, there will be some $P_n \in \mathscr{P}_n$ such that $H(P_n) \geq R + \delta/2$ and by taking n even larger if needed, we can get close enough to the P that minimises the infimum in $(\star)$ so that

  $$D(P_n\|Q) \leq D^* + 2\epsilon$$

  Now we claim that B_n is necessarily much smaller than the type class $T(P_n)$, this will come in handy later. Let $r_n \downarrow 0$ be a sequence of real numbers tending to zero with

  $$\frac{1}{n}\log|B_n| \leq R + r_n \qquad \text{for all } n \geq 1$$

  (This is allowed because of the assumption on the rate of B_n) Then we have the following:

  $$\begin{aligned}\frac{|B_n|}{|T(P_n)|} &\overset{(1)}{\leq} \frac{2^{n(R+r_n)}}{(n+1)^{-m}2^{nH(P_n)}} \\ &= (n+1)^m 2^{n[R-H(P_n)+r_n]} \\ &\overset{(3)}{\leq} (n+1)^m 2^{n[-\delta/2+r_n]}\end{aligned}$$

  Where (1) comes from the lower bound found in Proposition size of Type Class (Proposition 8.8), the second step is trivial, and step (3) comes from the construction of P_n. It is therefore clear that the right hand size goes to zero as $n \longrightarrow \infty$, and so in particular, for all n large enough,

  $$\frac{|B_n|}{|T(P_n)|} \leq \frac{1}{2}. \qquad (\star\star)$$

We are now armed to lower bound $P_e^{(n)}$ as required. Let $x_1^n \in A^n$ be any string in $T(P_n)$, then

$$\begin{aligned}
P_e^{(n)} &\overset{(1)}{=} \mathbf{P}(X_1^n \in B_n^c) \\
&\overset{(2)}{\geq} \mathbf{P}(X_1^n \in T(P_n) \cap B_n^c) \\
&\overset{(3)}{=} |T(P_n) \cap B_n^c| Q^n(x_1^n) \\
&\overset{(4)}{=} \frac{|T(P_n) \cap B_n^c|}{|T(P_n)|} Q^n(T(P_n)) \\
&\overset{(5)}{\geq} \left(1 - \frac{|T(P_n) \cap B_n|}{|T(P_n)|}\right)(n+1)^{-m} 2^{-nD(P_n\|Q)} \\
&\overset{(6)}{\geq} \frac{1}{2(n+1)^m} 2^{-n(D^*+2\epsilon)}
\end{aligned}$$

Where (1) is by definition of the error probability, (2) is trivial, (3) comes from the fact that all strings in $T(P_n)$ have the same frequency of symbols of $\mathscr{A}$, and since they are IID, they all in particular have the same probability. Step (4) is again because all the strings in $T(P_n)$ have the same probability, so the probability of one string is the sum of the probabilities of all strings divided by the number of strings. Step (5) is because of the lower bound found in Corollary of Type Class proposition (Proposition 8.10) as well as the fact that the proportion of strings in $T(P_n)$ that are in B_n^c is one minus the proportion of strings that are in B_n, and finally step (6) is due to the bound we have established in ($\star\star$) as well as the bound we have established on $D(P_n\|Q)$. This now gives the desired lower bound on the decay exponent of the error probability, namely that

$$\liminf_{n\longrightarrow\infty} \frac{1}{n} \log P_e^{(n)} \geq -(D^* + 2\epsilon)$$

Since ϵ was arbitrary the proof is finished. ♡

9 Variable rate Lossless data compression

In preceding discussions we studied the fundamental information-theoretic properties that fixed-rate lossless data compression had. One of the practical problems however, is that the codebook method required an exponential search when encoding and decoding. We now study a more practical method of data compression:

Definition 9.1 (Variable-rate code) A variable-rate lossless compression code with blocklength n on an alphabet $\mathscr{A}$ is a pair (C_n, L_n) where the encoder

$$C_n : \mathscr{A}^n \longrightarrow \{0,1\}^*$$

Is a one-to-one function that maps source strings x_1^n to variable-length binary codewords $C_n(x_1^n)$, and the length-function L_n is defined as

$$L_n(x_1^n) = [\text{length of } C_n(x_1^n)]\,\text{bits}$$

For this method to be a sensible way of communicating messages, we must have the following condition:

Definition 9.2 (Prefix-free code) The code (C_n, L_n) is a prefix-free code if no code $C_n(x_1^n)$ is a prefix of another code $C_n(y_1^n)$ whenever $x_1^n \neq y_1^n$

Symbol	Codeword
A	0
B	10
C	110
D	111

Table 1: Prefix-Free Codewords

Symbol	Codeword
A	0
B	01
C	001
D	0111

Table 2: Non-Prefix-Free Codewords

Naturally in the context of a non-prefix-free codeword, if the decoder were to receive the message 001, it would not be possible to detect whether the message was AB or C. Naturally, from a practical standpoint, we wish to create a code that assigns to each message a codeword of minimal length, however, the following result expresses a fundamental combinatorial limit on possible prefix-free codeword lengths, indicating that it is not possible for a code to assign only short codes and still be prefix-free.

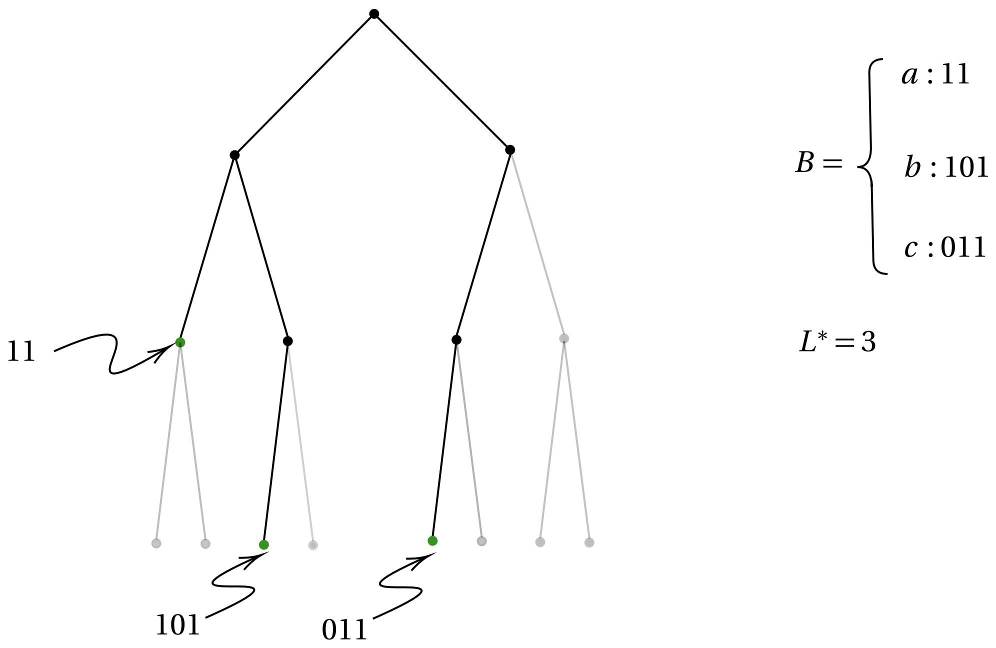

$B = \begin{cases} a : 11 \\ b : 101 \\ c : 011 \end{cases}$
$L^* = 3$
11
101
011

Figure 7: The diagram that says it all: Kraft's Inequality

Theorem 9.3 (Kraft's Inequality, ♫) Let $L_n : \mathscr{A}^n \longrightarrow \mathbf{N}$ be a length function.

- If L_n satisfies Kraft's inequality, i.e:

$$\sum_{x_1 \in \mathscr{A}^n} 2^{-L_n(x_1^n)} \leq 1,$$

  there is a prefix-free code C_n on A^n with length function L_n.

- If (C_n, L_n) is a prefix-free code, then L_n must satisfy Kraft's inequality.

Main idea: This all steps from the correspondence of codes and nodes in a binary tree. The prefix-free condition guarantees that no codeword is a descendant of any other codeword. When trying to check Kraft's inequality, the fact that the subtrees rooted at the codewords are disjoint give the inequality. When trying to construct the codeword, Kraft's inequality guarantees there are enough leaves.

> *Proof.* We first prove the converse. Let L^* be the maximum code-word length, and consider the complete binary tree of length L^*. Every codeword can be thought of as a node in the tree, so let the i^{th} code be represented by a node v_i. Let A_i represent the set of leaves of the subtree rooted at v_i. If a codeword v_i has length L_i, then it has $2^{L^*-L_i}$ leaves at depth L^* as its descendants. Since we are assuming the prefix-free condition, it follows that no codeword is a direct descendant of any other codeword in the tree, as that would contradict the prefix-free condition. Therefore the collection of leaves at the bottom of the tree descending from each codeword are disjoint, i.e:

$A_i \cap A_j = \varnothing$ for $i \neq j$. Therefore we have the following:

$$\begin{aligned} 2^{L^*} &\overset{(1)}{\geq} \left|\bigcup_i A_i\right| \\ &\overset{(2)}{\geq} \sum_i |A_i| \\ &\overset{(3)}{=} \sum_i 2^{L_*-L_i} \end{aligned}$$

Step (1) comes from the fact that there are a total of 2^{L^*} leaves at the bottom of the tree, so the collection of all the leaves out of each codeword cannot exceed this number (the reason why it is not an equal sign is because we can have sections of the tree that contain no codewords, see figure). Step (2) comes from the fact that the A_i's are disjoint, and step (3) comes from the calculation we did on the size of each A_i. For the direct part, given a collection of codeword lengths $\{L_n(x_1^n) : x_1^n \in \mathscr{A}^n\}$, i.e: we have codeword lengths

$$\ell_1 \leq \ell_2 \leq \cdots \leq \ell_n = L^*$$

We choose the codewords as follows:

- Choose any node from the full tree at depth ℓ_1. Since we are building a prefix-free code, all its descendants become unusable. After doing this we have discarded $2^{L^*-\ell_1}$ nodes at depth L^*.
- Choose a surviving node at depth ℓ^2, which removes $2^{L^*-\ell_2}$ leaf nodes at depth L^*.
- The question is whether when we have reached the final iteration of the algorithm we have ran out of leaves at the bottom of the tree. When we reach the last choice of our algorithm, i.e: choosing the last codeword, which has length $\ell_n = L^*$, we will need to pick a leaf at the bottom of the tree, and there will be remaining

  $$2^{L^*} - \sum_{i=1}^{n-1} 2^{L^*-\ell_i} = 2^{L^*} - \left(\sum_{i=1}^{n} 2^{L^*-\ell_i} - 2^{L^*-L^*}\right) \geq 1$$

  leaves remaining, i.e: we will not have ran out of options, which means we can construct our code.

♡

9.1 Code-distribution correspondence

Naturally, we wish to consider the problem of assigning the optimally short codewords to our messages, and for this we would like to take advantage of the probability distribution of our underlying messages, assigning shorter codelengths to the more likely messages, as an illustration, if we have a trivial set of messages, where there are only two messages that we have to send, then we would only need 1 bit of length for our codewords. The following result establishes the fact that there is a way, given any PMF on our set of messages, to create a prefix-free code that incorporates this idea of assigning shorter lengths to more probable messages. The following theorem also tells us that we are also limited in how well we can compress messages, and this limitation will also come in shape of PMF.

Theorem 9.4 (Code-distribution correspondence, ♫) We have the following:

- For any distribution P_n on $\mathscr{A}^n$, there is a prefix-free code (C_n^*, L_n^*) with

$$L_n^*(x_1^n) < -\log P_n(x_1^n) + 1 \text{ bits, for all } x_1^n \in \mathscr{A}^n.$$

- For any prefix-free code (C_n, L_n) there is a distribution Q_n on $\mathscr{A}^n$ such that

$$L_n(x_1^n) \geq -\log Q_n(x_1^n) \text{ bits, for all } x_1^n \in \mathscr{A}^n.$$

Main idea: The proof is really easy. From the statement of the direct part, it is no surprise that we need to show that there eixsts a prefix-free code with length function $L_n^*(x_1^n) = \lceil -\log P_n(x_1^n) \rceil$. To show this existence, simply verify Kraft's inequality. Conversely, we are told that Kraft's inequality is satisfied, i.e:

$$Z := \sum_{x_1^n} 2^{-L(x_1^n)} \leq 1$$

This looks an awful lot like a normalisation constant, and so we may define

$$Q^n(x_1^n) = \frac{2^{-L_n(x_1^n)}}{Z}$$

and then simply solve for $L_n(x_1^n)$.

Proof. For the direct part, we note that

$$\sum_{x_1^n \in \mathscr{A}^n} 2^{-\lceil -\log P_n(x_1^n) \rceil} \leq \sum_{x_1^n \in \mathscr{A}^n} 2^{\log P_n(x_1^n)} = 1$$

So Kraft's inequality is satisfied for the length function $L_n^*(x_1^n) = \lceil -\log P_n(x_1^n) \rceil < -\log P_n(x_1^n) + 1$.

For the converse part, if we assume (C_n, L_n) is prefix-free, then we must have that

$$\sum_{x_1^n \in \mathscr{A}^n} 2^{-L_n(x_1^n)} \leq 1$$

We may define a PMF on $\mathscr{A}^n$ by

$$Q_n(x_1^n) = \frac{2^{-L_n(x_1^n)}}{Z}$$

and so

$$-\log Q_n(x_1^n) = L_n(x_1^n) + \log Z \leq L_n(x_1^n)$$

Where this last inequality comes from the fact that $Z \leq 1$ so $\log Z \leq 0$. ♡

We reach the final great result on this topic, which establishes fundamental bounds on the expected length of a code. It should be of no surprise that entropy plays a key role in the limitations of how well we can compress our data, i.e: how short we can make our messages on average whilst still being prefix-free.

Theorem 9.5 (Entropy bounds, ♫) Suppose a message X_1^n has a distribution P_n on our alphabet $\mathscr{A}^n$, then:

- There is a prefix-free code (C_n^*, L_n^*) that achieves an expected description length:

$$\mathbf{E}[L_n^*(X_1^n)] < H(X_1^n) + 1$$

- For any prefix-free code (C_n, L_n) on $\mathscr{A}^n$,

$$\mathbf{E}[L_n(X_1^n)] \geq H(X_1^n)$$

Main idea: We just use the Code-Distribution correspondence. For the first part it is clear what to do, just take expectations to both sides, for the converse, just use the trick

$$\mathbf{E}_P[-\log Q(X)] = \mathbf{E}_P\left[-\log\left(\frac{P(X)Q(X)}{P(X)}\right)\right]$$

Proof. For the direct part, we know from the code-distribution correspondence, that if P_n is the distribution of X_1^n, then $L_n^*(x_1^n) \leq -\log P_n(x_1^n) + 1$, and so

$$\mathbf{E}[L_n^*(X_1^n)] \leq \mathbf{E}[-\log P_n(X_1^n) + 1] = H(X_1^n) + 1$$

For the converse, we use again the code-distribution correspondence to see that there is a distribution Q_n for which $L_n(x_1^n) \geq -\log Q_n(x_1^n)$ bits, and so

$$\begin{aligned}
\mathbf{E}[L_n(X_1^n)] &\geq \mathbf{E}[-\log Q_n(X_1^n)] \\
&= \mathbf{E}\left[\log\left(\frac{1}{P_n(X_1^n)}\frac{P_n(X_1^n)}{Q_n(X_1^n)}\right)\right] \\
&= \mathbf{E}\left[\log\frac{1}{P_n(X_1^n)}\right] + \mathbf{E}\left[\log\frac{P_n(X_1^n)}{Q_n(X_1^n)}\right] \\
&= H(X_1^n) + D(P_n\|Q_n) \geq H(X_1^n)
\end{aligned}$$

♡

Remark 9.6 A nice way of expressing this result is that

$$H(X_1^n) \leq \inf_{(C_n,L_n)} \mathbf{E}[L_n(X_1^n)] < H(X_1^n) + 1$$

where the infimum is taken over prefix-free codes.

We conclude with a final nail in the coffin that really drives the point home that entropy is the best rate one can achieve for variable-rate compression. The proof follows immediately from the remark above.

Corollary 9.7 (Compression rate) Let $\mathbf{X} = \{X_n\}$ be a stationary source. The entropy rate $H = H(\mathbf{X})$ (which exists by Example 7.5) is the best asymptotically achievable compression rate among all variable-rate prefix-free codes:

$$\lim_{n\longrightarrow\infty} \inf_{(C_n,L_n)} \frac{1}{n}\mathbf{E}[L_n(X_1^n)] = H$$

Where the infimum is taken over prefix-free codes.

9.2 Shannon codes and properties

We end this course with a brief discussion of Shannon codes and their properties.

Definition 9.8 (Shannon code) Given a distribution Q^n on $\mathscr{A}^n$, the code with length function

$$L_n(x_1^n) = \lceil -\log Q^n(x_1^n) \rceil$$

is called the Shannon code

Remark 9.9 Note that this code indeed exists, because the length function satisfies Kraft's inequality as shown in the Code-Distribution correspondence Theorem, and moreover, it is not unique, as it is easy to see (by for example looking at an example of a code-tree) that for a given length function $l_1 \leq \cdots \leq l_k$, there is more than one code attaining this lengths (one can think of swapping branches of the tree for example).

Definition 9.10 (Ideal Shannon codelengths) For a distribution Q^n on $\mathscr{A}^n$, we call the function

$$L_n(x_1^n) = -\log Q^n(x_1^n)$$

the ideal Shannon Codelength.

Remark 9.11 This name comes from the fact that as n becomes large, taking the integer part or leaving the decimal part has no effect on the asymptotics.

Theorem 9.12 (Competitive Optimality of Shannon Codes) Let $X_1^n \sim P^n$ be a message on $\mathscr{A}^n$, then for any other distribution Q^n on $\mathscr{A}^n$, we have that

$$\mathbf{P}[-\log Q^n(X_1^n) \leq -\log P^n(X_1^n) - K] \leq 2^{-K}$$

Remark 9.13 (Waffle) Recall that any variable-rate compression code can be identified with a probability measure Q^n on $\mathscr{A}^n$. What this Theorem tells you is that if you are trying to compress any file (that is large enough so that the right hand side of the inequality is not negative), the probability that any algorithm beats the Shannon code by a merely 20 bits is less than one in a million.

Main idea: Markov's Inequality.

Proof.

$$\begin{aligned}\mathbf{P}[-\log Q^n(X_1^n) \leq -\log P^n(X_1^n) - K] &= \mathbf{P}\left[\log \frac{Q^n(X_1^n)}{P^n(X_1)} \geq K\right] \\ &= \mathbf{P}\left[\frac{Q^n(X_1^n)}{P^n(X_1^n)} \geq 2^K\right] \\ &\leq 2^{-K} \sum_{x_1^n} P^n(x_1^n) \frac{Q(x_1^n)}{P(x_1^n)} \\ &= 2^{-K}\end{aligned}$$

♡

A Some nice results

Theorem A.1 (Jensen's Inequality) Let X be a random variable taking values in $\mathscr{A} \subseteq \mathbf{R}$ and let $f : \mathscr{A} \longrightarrow \mathbf{R}$ be a convex function. Then

$$\mathbf{E}[f(X)] \geq f(\mathbf{E}[X])$$

Corollary A.2 (Log-sum inequality) Let (a_i) and (b_i) be a collection of real, non-negative numbers. Then we have that

$$\sum_i a_i \log \frac{a_i}{b_i} \geq \left(\sum_i a_i \right) \log \frac{\left(\sum_i a_i\right)}{\left(\sum_i b_i\right)}$$

Proposition A.3 (Marginal trick) Let (X_1^n) have a joint mass function $P_{X_1^n}$ on $\mathscr{A}^n$. Let $f : \mathscr{A} \longrightarrow \mathbf{R}$ is a function, then

$$\sum_{x \in \mathscr{A}^n} P_{X_1^n}(x) f(x_i) = \mathbf{E}_{P_{X_i}}[f(X_i)] \equiv \sum_{x \in A} P_{X_i}(x) f(x)$$

Proof. The proof is, as the name suggests, a very simple trick to get the marginal distribution. Its a really simple computation, but I write it down for future reference.

$$\begin{aligned}
\sum_{x \in \mathscr{A}^n} P_{X_1^n}(x) f(x_i) &= \sum_{x_1 \in A} \sum_{x_2 \in A} \cdots \sum_{x_i \in A} f(x_i) \cdots \sum_{x_n \in A} P_{X_1^n}(x_1, x_2, \cdots, x_n) \\
&= \sum_{x_i \in A} f(x_i) \underbrace{\left(\sum_{j \neq i} \sum_{x_j \in A} P_{X_1^n}(x_1, \cdots, x_i, \cdots, x_n) \right)}_{\text{the marginal distribution}} \\
&= \sum_{x_i \in A} f(x_i) P_{X_i}(x_i)
\end{aligned}$$

♡

Proposition A.4 (Cesáro Lemma) If $a_n \longrightarrow a$ and $b_n = \frac{1}{n} \sum_{i=1}^n a_i$, then $b_n \longrightarrow a$.

Proposition A.5 (Factorial inequality) Let $k, \ell \in \mathbf{N}$, then $k!/\ell! \geq \ell^{k-l}$.

Proposition A.6 (Uniqueness of minima in strictly convex function) Let $f : E \longrightarrow \mathbf{R}$ be a strictly convex function on a convex domain E. Then if f attains a minimum at $x \in E$, then this minimum is unique.

Proof. Suppose that $y \in E$ is another minimiser, i.e: $f(x) = f(y)$. Then by taking a convex combination of x and y, we see that

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda) f(y) = f(y)$$

and so we have found a point $z = \lambda x + (1-\lambda)y$ (which belongs to E by the convexity of E) with $f(z) < f(y)$, which contradicts the minimality of y. ♡