Mathematics Tripos Part III
Information Theory
Michaelmas 2026

Part III Information Theory: Course Notes

Lampros Gavalakis *

September 23, 2026

Notes for the material covered in the 2026 *Information Theory* course in Part III of the Cambridge University Mathematics Tripos. These are essentially the notes for the course given in Michaelmas 2025 by Ioannis Kontoyiannis, but some of the material is different (Sections 10 and 11). The notes are meant to be supplementary material, while **the course syllabus is the material covered in the lectures.** All examples sheets are included in the last section.

The material in these notes can be found, perhaps in somewhat different form, in the wonderful textbook by Cover & Thomas [6], the papers [14, 15, 20, 16], and the book [19]. The texts [7, 17, 8] may also be useful. Further references to specific results are given in the text.

The few places where measure-theoretic arguments are required are marked with a '*'. Feel free to ignore the corresponding proofs, although the arguments themselves should be fairly easy to follow on an intuitive level.

Comments and corrections are welcome.

*Statistical Laboratory, Department of Pure Mathematics & Mathematical Statistics, University of Cambridge. Email: lg560@cam.ac.uk. Web: dpmms.cam.ac.uk/lg560

Contents

1 What is entropy? 1
- 1.1 The Asymptotic Equipartition Property (AEP) . . . 2
- 1.2 Fixed-rate lossless data compression . . . 5

2 What is relative entropy? 7
- 2.1 Asymptotically optimal hypothesis testing . . . 8
- 2.2 Relative entropy and optimal hypothesis testing . . . 10

3 Properties of entropy and relative entropy 14
- 3.1 Entropy . . . 15
- 3.2 Relative entropy . . . 18

4 Poisson approximation 21
- 4.1 Poisson approximation via entropy . . . 21
- 4.2 What is the Poisson distribution? . . . 23

5 What is mutual information? 26
- 5.1 Basic properties . . . 27
- 5.2 Synergy and redundancy . . . 28

6 Entropy rate 30

7 * The Shannon-McMillan-Breiman theorem 32

8 Types and large deviations 35
- 8.1 The method of types . . . 35
- 8.2 Sanov's theorem . . . 37
- 8.3 The Gibbs conditioning principle . . . 41
- 8.4 Error probability in fixed-rate data compression . . . 43

9 Variable-rate lossless data compression 46
- 9.1 The codes-distributions correspondence . . . 47
- 9.2 Shannon codes and their properties . . . 49

10 Entropy and additive combinatorics: Sumset bounds 51
- 10.1 Simple sumset entropy bounds . . . 51
- 10.2 The doubling-difference inequality for entropy . . . 53
- 10.3 Small Ruzsa distance . . . 55

11 The Entropy Power Inequality 57

12 Examples papers 62

Bibliography 73

1 What is entropy?

In this section we argue that the entropy is the natural, inevitable answer to the question, *how much can a typical message be compressed?*

Let us start with an elementary computation in the simplest possible setting. Consider a binary string, or "message," $x_1^n := (x_1, x_2, \ldots, x_n) \in \{0,1\}^n$, consisting of n bits x_i. If there was only one possible such message, then we would know it and surely there would be no need to ever describe it, let alone to compress it. To model the uncertainty in the message x_1^n let us suppose it is a realisation of the binary random variables (RVs) $X_1^n := (X_1, X_2, \ldots, X_n)$, and let us further assume that the X_i are independent and identically distributed (IID), with common distribution $X_i \sim \text{Bern}(p)$, that is, each X_i has a Bernoulli distribution with parameter p,

$$X_i = \begin{cases} 1, & \text{with probability } p, \\ 0, & \text{with probability } 1-p, \end{cases}$$

for some fixed $p \in (0,1)$.

How do we go about compressing such a message? Or even, what does it mean to compress it? Perhaps the first, most natural idea, is to try and give shorter descriptions to common, high-probability messages, and longer descriptions to rare, low-probability messages. Which leads to the question, what is the probability distribution of these messages like, for large n?

This is something we can compute. Let P denote the common probability mass function (PMF) of the X_i, $P(1) = 1 - P(0) = p$, and write P^n for the joint PMF of the vector X_1^n. Then, the probability $\mathbb{P}(X_1^n = x_1^n) = P^n(x_1^n)$ of a given message x_1^n can be written,

$$P^n(x_1^n) = \prod_{i=1}^{n} P(x_i) = p^k (1-p)^{n-k},$$

where k is the number of 1s in x_1^n. But the law of large numbers tells us that, for large n, the proportion of 1s in a random x_1^n is $k/n \approx p$, with high probability, so,

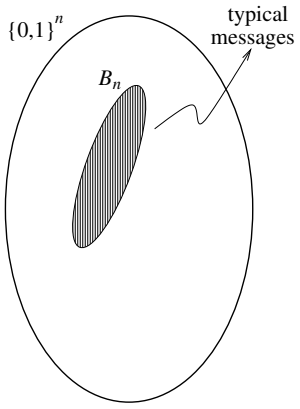
$$P^n(x_1^n) \approx p^{np}(1-p)^{n(1-p)} = 2^{-n[-p\log p - (1-p)\log(1-p)]},$$

where, throughout these notes, $\log \equiv \log_2$ denotes the logarithm taken to base 2. Therefore, writing,

$$h(p) := -p\log p - (1-p)\log(1-p), \qquad p \in (0,1), \tag{1}$$

for the *binary entropy function*, we have that, for a typical long message x_1^n,

$$P^n(x_1^n) \approx 2^{-nh(p)}. \tag{2}$$

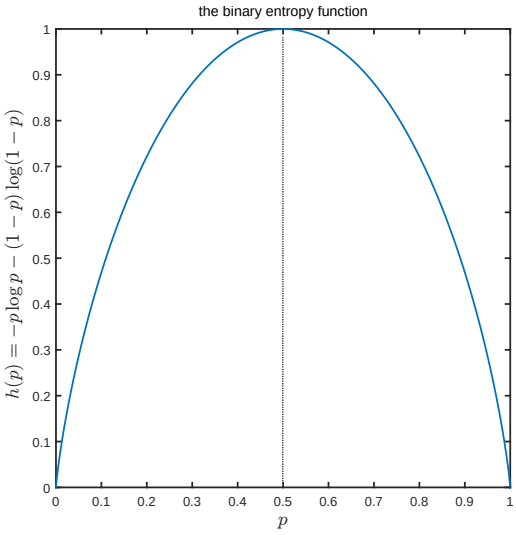


This elementary computation has a remarkable feature: The result of the approximation in (2) is *independent of the message* x_1^n: *Any* message x_1^n we are likely to encounter will have the same probability $\approx 2^{-nh(p)}$. This means that we can think of the set $\{0,1\}^n$ of all possible n-bit messages as consisting of two parts: The *typical* messages, that are approximately uniformly distributed with probability $\approx 2^{-nh(p)}$ each, and the non-typical messages that occur with negligible probability. Moreover, since nearly all the probability is concentrated on a set all of whose elements have probability $\approx 2^{-nh(p)}$, there must be $\approx 2^{nh(p)}$ of them.

Let $B_n \subset \{0,1\}^n$ denote this set of $\approx 2^{nh(p)}$ typical messages. For the purposes of compression, suppose that an encoder and a decoder share B_n ahead of time and agree on an ordering of all its elements as $B_n = \{x_1^n(1), x_1^n(2), \ldots, x_1^n(b)\}$, say, where $b = |B_n| \approx 2^{nh(p)}$ is the cardinality of B_n. Then the encoder can describe any typical $x_1^n(j)$ to the decoder by describing its index j. And since j is an integer between 1 and b, it can be described in its binary representation using,

$$\lceil \log b \rceil = \lceil \log |B_n| \rceil \approx nh(p) \quad \text{bits.}$$

Some important observations are in order. First, the more random our messages, i.e., the closer p is to $1/2$, the higher the entropy $h(p)$, and the larger the number of typical strings. Second, assuming we ignore non-typical strings (since they have vanishingly small probability for large n), the above method achieves a compression "rate" of $h(p)$: Our initial n bits x_1^n are compressed down to $nh(p)$ bits, where $h(p)$ is strictly less than 1 unless our message is a completely random sequence of fair coin flips. In fact, as we will see later on, the entropy is the *best* compression rate that can be ever achieved. And third, the less random our messages, i.e., the closer p is to 0 or 1, the greater the compression we can achieve.



The following two subsections make these ideas precise, in a more general setting.

1.1 The Asymptotic Equipartition Property (AEP)

Suppose $X_1, X_2, \ldots, X_n$ are IID RVs, with values in a finite alphabet $A = \{a_1, a_2, \ldots, a_m\}$ of $m \geq 1$ symbols. Their common distribution is described by their PMF P, so that, $P(x) = \mathbb{P}(X_i = x)$, for $x \in A$, $i = 1, 2, \ldots, n$. With a slight abuse of terminology we will often say 'X has distribution P on A,' denoted $X \sim P$. Also, for simplicity we will write X_i^j for any block of RVs $(X_i, X_{i+1}, \ldots, X_j)$ with $i \leq j$, and similarly we denote by x_i^j a sting of length $j - i + 1$ consisting of letters from the alphabet A, so that $x_i^j = (x_i, x_{i+1}, \ldots, x_j) \in A^{j-i+1}$. Since the X_i are independent, their joint PMF, denoted by P^n, can be written as a product,

$$\mathbb{P}(X_1^n = x_1^n) = \prod_{i=1}^n \mathbb{P}(X_i = x_i) = \prod_{i=1}^n P(x_i) = P^n(x_1^n).$$

We therefore also say that 'the RVs X_1^n have the (product) distribution P^n.'

Following the intuition gained by the computation above, let us look at the "random probability" $P^n(X_1^n)$ of a random string X_1^n:

$$P^n(X_1^n) = \prod_{i=1}^n P(X_i).$$

Taking logarithms and dividing by n,

$$\frac{1}{n} \log P^n(X_1^n) = \frac{1}{n} \sum_{i=1}^n \log P(X_i),$$

and using the weak law of large numbers (WLLN) as $n \to \infty$ for the IID RVs $Y_i := -\log P(X_i)$, $i \geq 1$, yields:

Theorem 1.1. *Let $\{X_n\}$ be IID RVs with common distribution P on the finite alphabet A. Then,*

$$-\frac{1}{n}\log P^n(X_1^n) \to H, \qquad \textit{in probability},$$

as $n \to \infty$, where $H = H(X_i)$ is the common entropy *of the RVs X_i:*

Definition 1.2 (Entropy)**.** *For a discrete RV $X \sim P$ on finite or countably infinite alphabet A, the* entropy of X *is:*

$$H(X) = H(P) = -\sum_{x\in A} P(x)\log P(x) = \mathbb{E}[-\log P(X)]. \tag{3}$$

Remarks.

1. In case some of the probabilities $P(a)$ in the above definition are zero, we take $0\log 0 = 0$ by convention, as suggested by continuity: $x\log x \to 0$ as $x \downarrow 0$. The more rigorously-minded reader who might find this somewhat *ad hoc* convention unsatisfying can observe that we may directly define $H(X)$ as the expectation $\mathbb{E}[-\log P(X)]$ which, if interpreted as the Lebesgue integral of the function $-\log P(\cdot)$ with respect to the probability measure that places masses $P(a)$ at the points a of the probability space A equipped with the power set σ-algebra, leads to the same result.

2. Although the entropy is really a functional of the probability distribution P and not of the RV X, we use the notation $H(P)$ and $H(X)$ interchangeably. In fact, $H(X)$ only depends on the probabilities $P(x)$ not on the values $x \in A$ themselves. Therefore $H(f(X)) = H(X)$ for any 1-1 function f on A.

3. Since all the terms in the definition (3) of $H(X)$ are nonnegative, the sum always exists and takes a unique value in $[0, +\infty]$, even when A is countably infinite. Moreover, $H(X) \geq 0$ and can only be zero if all the terms are zero, i.e., if $P(x)$ is either 0 or 1 for all $x \in A$, which can only happen if X is constant (or 'deterministic') in that it is equal to a specific $x^* \in A$ with probability 1.

4. Despite the fact that the result of Theorem 1.1 is a simple application of the WLLN, it has far reaching consequences and generalisations, several of which will be explored in the course of the discussion in these notes. A first such consequence is Shannon's celebrated *Asymptotic Equipartition Property* (AEP) stated next and exploited further in Section 1.2. The general form of Theorem 1.1 – known as the Shannon-McMillan-Breiman theorem – is proved in Section 7.

Corollary 1.3 (AEP)**.** *Let $\{X_n\}$ be IID RVs on the finite alphabet A, with common distribution P and common entropy $H = H(X_i)$.*

($\Rightarrow$) *For any $\epsilon > 0$, the set of typical strings $B_n^* \subset A^n$,*

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

satisfies:

$$|B_n^*| \leq 2^{n(H+\epsilon)}, \qquad \textit{for all } n \geq 1, \tag{4}$$

and,

$$\mathbb{P}(X_1^n \in B_n^*) \to 1, \qquad \textit{as } n \to \infty. \tag{5}$$

($\Leftarrow$) *Let $\{B_n\}$ be any sequence of subsets of A^n, $n \geq 1$. If $\mathbb{P}(X_1^n \in B_n) \to 1$, as $n \to \infty$, then, for any $\epsilon > 0$,*

$$|B_n| \geq (1-\epsilon)2^{n(H-\epsilon)}, \qquad \textit{eventually.} \tag{6}$$

Before giving the proof, we make some observations.

Remarks.

5. In the information-theoretic literature it is common to state results in a direct-and-converse format as above, although the actual statements are not, strictly speaking, each other's converses. The intention is to indicate that a specific type of object exists that can achieve a certain quantitative performance (the direct $\Rightarrow$ part), and that no other object of this type can substantially outperform it (the converse $\Leftarrow$ part). Theorem 1.6 below is a typical example of this formulation in the context of data compression.

6. From Theorem 1.1 we get a mathematical interpretation to the entropy. It says that the probability of a random string X_1^n typically decays exponentially with n, with exponent given by the entropy: For large n, $P^n(X_1^n) \approx 2^{-nH}$, with high probability. The AEP offers a different, more 'operational' interpretation. It states that the smallest set of strings that can carry 'almost all' the probability of P^n, contains $\approx 2^{nH}$ strings.

7. From the AEP we also see that, the higher the entropy, the greater the number of typical strings, and the more unpredictable the possible values of X_1^n. In this sense, we naturally think of "high entropy" RVs as being "more random" and "less predictable."

PROOF. For the direct part ($\Rightarrow$), note that (5) is simply a re-statement of the convergence in probability in Theorem 1.1: For $\epsilon > 0$ arbitrary:

$$\mathbb{P}(X_1^n \notin B_n^*) = \mathbb{P}\Big(\Big| -\frac{1}{n}\log P^n(X_1^n) - H\Big| > \epsilon\Big) \to 0, \qquad \text{as } n \to \infty.$$

Also, by the definition of B_n^*,

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

since all $x_1^n \in B_n^*$ have $P^n(x_1^n) \geq 2^{-n(H+\epsilon)}$, which gives (4).

For the converse ($\Leftarrow$), since $\mathbb{P}(X_1^n \in B_n) = P^n(B_n) \to 1$ by assumption, and $\mathbb{P}(X_1^n \in B_n^*) = P^n(B_n^*) \to 1$ by (5), we have that,

$$\begin{aligned} P^n(B_n \cap B_n^*) &= P^n(B_n) + P^n(B_n^*) - P^n(B_n \cup B_n^*) \\ &\geq P^n(B_n) + P^n(B_n^*) - 1, \end{aligned} \tag{7}$$

also tends to 1, as $n \to \infty$. Therefore, for $\epsilon > 0$ arbitrary, we must have $P^n(B_n \cap B_n^*) \geq 1 - \epsilon$, eventually, so that,

$$1 - \epsilon \leq P^n(B_n \cap B_n^*) = \sum_{x_1^n \in B_n \cap B_n^*} P^n(x_1^n),$$

eventually. And since all the x_1^n that are being summed over belong to B_n^*, their probabilities are all $\leq 2^{-n(H-\epsilon)}$, hence,

$$1 - \epsilon \leq |B_n \cap B_n^*| \times 2^{-n(H-\epsilon)} \leq |B_n| \times 2^{-n(H-\epsilon)},$$

and rearranging, gives (6). □

Notation. In the proof we wrote $P^n(B_n)$ for the probability $\mathbb{P}(X_1^n \in B_n)$, implicitly identifying the PMF P^n with the discrete probability measure induced by the RVs X_1^n on A^n. This convention will remain in effect throughout.

1.2 Fixed-rate lossless data compression

Let $\boldsymbol{X} = \{X_n\}$ be a *memoryless source* with distribution P, i.e., a sequence of IID RVs, with common PMF P on the same alphabet A.

Definition 1.4 (Fixed-rate codes)**.** *A* fixed-rate lossless compression code *for the source* $\boldsymbol{X}$ *consists of a sequence of* codebooks $\{B_n\}$*, where each* B_n *is a set of source strings of length* n*,* $B_n \subset A^n$*,* $n \geq 1$*.*

The way the code works is simple. Suppose we want to compress a specific source string, or message, x_1^n. We assume that both the encoder and decoder have access to the codebooks and that they have agreed to sort each B_n in a mutually agreed upon way, e.g., lexicographically. Then, to send x_1^n, the encoder first checks whether $x_1^n \in B_n$. If it is, she describes to the decoder by first sending a "1," followed by the index of x_1^n in B_n. This takes,

$$1 + \lceil \log |B_n| \rceil \quad \text{bits.}$$

Otherwise, she declares an "error" and sends a "0" followed by x_1^n uncompressed; this takes,

$$1 + \lceil \log |A^n| \rceil = 1 + \lceil n \log |A| \rceil \quad \text{bits.}$$

Definition 1.5 (Rate and error probability)**.** *For each* $n \geq 1$*: The* rate *of a fixed-rate code* $\{B_n\}$ *for source* $\boldsymbol{X}$ *is,*

$$R_n = \frac{1}{n}\big(1 + \lceil \log |B_n| \rceil\big) \approx \frac{1}{n} \log |B_n| \quad \textit{bits/symbol,}$$

and its probability of error *is:*

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

Question: What is the best compression rate that can be achieved, if the probability of error is required to be negligible?

Answer: The entropy. The following result is a re-statement of the AEP in Corollary 1.3.

Theorem 1.6 (Fixed-rate coding theorem)**.** *Suppose* $\boldsymbol{X} = \{X_n\}$ *is a memoryless source with distribution* P *and entropy* $H = H(X_i)$*.*

($\Rightarrow$) *For any* $\epsilon > 0$*, there is a fixed-rate code* $\{B_n^*\}$ *with vanishing probability of error* $P_e^{(n)} \to 0$ *as* $n \to \infty$ *and with rate,*

$$R_n \leq H + \epsilon + \frac{2}{n} \quad \textit{bits/symbol,} \qquad \textit{for all } n \geq 1.$$

($\Leftarrow$) *Let* $\{B_n\}$ *be any fixed-rate code with vanishing error probability* $P_e^{(n)} \to 0$ *as* $n \to \infty$*. Then for any* $\epsilon > 0$ *its rate must satisfy,*

$$R_n > H - \epsilon \quad \textit{bits/symbol, eventually.}$$

Proof. Take B_n^* to be the sets of typical strings as in Corollary 1.3. Then, by (5), $P_e^{(n)} = 1 - \mathbb{P}(X_1^n \in B_n^*) \to 0$ as $n \to \infty$ and by (4),

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

This gives the direct part. For the converse, take $\epsilon > 0$ arbitrary and assume without loss of generality that $\epsilon < 1/2$. Since $P_e^{(n)} \to 0$, we have $\mathbb{P}(X_1^n \in B_n) = 1 - P_e^{(n)} \to 1$, as $n \to \infty$, so that, by (6),

$$R_n \geq \frac{1}{n} \log |B_n| + \frac{1}{n} \geq \frac{1}{n} \log(1 - \epsilon) + (H - \epsilon) + \frac{1}{n} = H - \epsilon + \frac{1}{n} \log(2(1 - \epsilon)) > H - \epsilon,$$

eventually. □

So, what is entropy? Theorem 1.6 says that, with vanishing error probability, we can achieve compression rates as close to the entropy H as we wish, but not smaller than H. This answers our question in the beginning of this section: The entropy is the best achievable compression rate. In this sense, we generally think of $H(X)$ as the *average amount of information contained in* X, and we measure it in bits.

Another useful interpretation is that, instead of thinking of the RVs X_1^n as a collection of random samples generated sequentially from the distribution P, we can also – to a first approximation – think of X_1^n as a single uniformly drawn sample from the set of all typical strings of length n. This point of view is both elegant and very useful, as we will see later on.

Although our original intention was to perform data compression by giving shorter descriptions to high-probability strings, our analysis led us to a somewhat different strategy: Give shorter descriptions to the *typical strings*. In fact, the highest-probability strings are generally not typical. For example, if the $\{X_n\}$ are IID Bern(3/4), then the highest probability strings are those where the proportion of 1s is close to 1, whereas for typical strings it is close to 3/4. Nevertheless, we can recover our original intuition for data compression, as explained in Exercise 12.2 (c).

What if we want to know something more specific about the probability of error? That is, if we select a target rate $R > H$, among all codes with $R_n \to R$ as $n \to \infty$, how small can the error probability $P_e^{(n)}$ be made? The answer, given in Section 8.4, is that $P_e^{(n)}$ can be exponentially small, $P_e^{(n)} \approx 2^{-nD^*}$, and the best (i.e., the largest) possible exponent D^* can be explicitly computed.

Finally, it is worth noting that the results of this section also hold for IID sources on a countably infinite alphabet, as long as the entropy is finite. It is interesting that even in this case, the typical set is finite and sufficient for optimal compression.

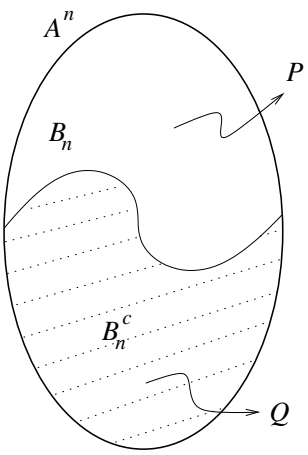
We will next look at a related problem in statistics, and we will revisit the problem of data compression in Section 9 from a more practical point of view.

2 What is relative entropy?

In this section we argue that relative entropy is the natural answer to the question of how "statistically distinguishable" two probability distributions are in a hypothesis test.

Consider a "simple vs. simple" hypothesis test: Given data x_1^n produced by a memoryless source, and given two candidate distributions P and Q on the same finite alphabet $A = \{a_1, a_2, \ldots, a_m\}$, we wish to decide whether the data comes from P or Q.

Recall that, earlier, the data compression problem was reduced to dividing the set A^n of all possible sequences into two parts: The codebook B_n, which consists of those strings that we think of as "typical" or "more likely," and its complement, B_n^c, which we ignore.



The situation is very similar here. Any hypothesis test, that is, any method of looking at the data and declaring whether it came from P or Q, conceptually boils down to selecting a *decision region* $B_n \subset A^n$, so that, if the data string $x_1^n \in B_n$ we declare it came from P, otherwise, that it came from Q. In this setting, the goal is to pick a test or, equivalently, to select a decision region B_n, so that the probability of making the wrong decision is as small as possible. In fact, here there are two different probabilities of error,

$$\begin{aligned} e_1^{(n)} &= \mathbb{P}(\text{declare } P \mid X_1^n \sim Q^n), \\ \text{and} \quad e_2^{(n)} &= \mathbb{P}(\text{declare } Q \mid X_1^n \sim P^n), \end{aligned}$$

or, more formally,

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

These expressions make clear the fact that there is a tension between making both probabilities small: For small $e_1^{(n)}$ we need B_n to be small, and for small $e_2^{(n)}$ we need its complement to be small.

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

Answer: It can be made exponentially small, $e_1^{(n)} \approx 2^{-nD}$, and the best (i.e., the largest possible) exponent is the relative entropy between P and Q:

Definition 2.1 (Relative entropy)**.** *The* relative entropy *between two PMFs P, Q on the same discrete (finite or countably infinite) alphabet A is:*

$$D(P\|Q) = \sum_{x \in A} P(x) \log \frac{P(x)}{Q(x)}. \tag{8}$$

Remarks.

1. As with the entropy, we make the convention that $0 \log 0 = 0 \log \frac{0}{0} = 0$.
2. In the case of infinite alphabets, some care is required in order to make sense of the definition of $D(P\|Q)$. For the series in (8) to be well-defined, either we must assume that the elements of $A = \{a_1, a_2, \ldots\}$ are explicitly ordered so that,

$$D(P\|Q) = \sum_{i=1}^{\infty} P(a_i) \log \frac{P(a_i)}{Q(a_i)},$$

or the series must converge absolutely,

$$\sum_{x \in A} P(x) \left| \log \frac{P(x)}{Q(x)} \right| < \infty, \tag{9}$$

so that $D(P\|Q) < \infty$ and the order of summation is immaterial. Throughout these notes, whenever we consider a relative entropy over an infinite alphabet, it will always be assumed that we are in one of these two cases.

3. A more principled way to define the relative entropy – which is indeed the general definition of relative entropy used when considering RVs with values in general, not necessarily discrete spaces – between any two probability measures μ, ν on the same probability space, is as $D(\mu\|\nu) := \int f \log f \, d\nu$, when their Radon-Nikodým derivative $f := d\mu/d\nu$ exists, and $D(\mu\|\nu) := +\infty$, otherwise. It is simple to check that this agrees with (8) when P, Q are discrete.

As we show in Section 2.2, $D(P\|Q)$ is always nonnegative, and it is only equal to zero if $P = Q$. Although it is not strictly speaking a metric (for example, $D(P\|Q) \neq D(Q\|P)$ in most cases), as we will see $D(P\|Q)$ does admin a natural interpretation as a "distance" measure between the distributions P and Q.

2.1 Asymptotically optimal hypothesis testing

Theorem 2.2 (Stein's lemma)**.** *Let P and Q be two distributions on the finite alphabet A, such that the relative entropy $D = D(P\|Q)$ is neither 0 nor $+\infty$. Suppose $\boldsymbol{X} = \{X_n\}$ is a memoryless source on the alphabet A, distributed either according to P or Q.*

($\Rightarrow$) *For any $\epsilon > 0$, there is a hypothesis test with decision regions $\{B_n^*\}$ such that,*

$$e_1^{(n)} \leq 2^{-n(D-\epsilon)}, \qquad \text{for all } n \geq 1,$$

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

($\Leftarrow$) *Any hypothesis test with decision regions $\{B_n\}$ for which $e_2^{(n)} \to 0$ as $n \to \infty$, has, for any $\epsilon > 0$,*

$$e_1^{(n)} \geq 2^{-n(D+\epsilon+\frac{1}{n})}, \qquad \text{eventually.} \tag{10}$$

Here is the main idea in the direct part ($\Rightarrow$) of the theorem. Suppose the source is indeed distributed according to P. Let us look at how much more (or less) likely a random string X_1^n is under P versus under Q,

$$\frac{P^n(X_1^n)}{Q^n(X_1^n)} = \prod_{i=1}^n \frac{P(X_i)}{Q(X_i)},$$

or, taking logarithms and dividing by n,

$$\frac{1}{n} \log \frac{P^n(X_1^n)}{Q^n(X_1^n)} = \frac{1}{n} \sum_{i=1}^n \log \frac{P(X_i)}{Q(X_i)},$$

which is the empirical average of the IID RVs, $Z_i = \log[P(X_i)/Q(X_i)]$, $i \geq 1$. Therefore, by the WLLN these averages converge in probability to,

$$\mathbb{E}(Z_1) = \mathbb{E}\left[\log \frac{P(X_1)}{Q(X_1)}\right] = \sum_{x \in A} P(x) \log \frac{P(x)}{Q(x)},$$

i.e., as $n \to \infty$,

$$\frac{1}{n} \log \frac{P^n(X_1^n)}{Q^n(X_1^n)} \to D(P\|Q), \quad \text{in probability.} \tag{11}$$

Informally, (11) says that, for large n,

$$\frac{P^n(X_1^n)}{Q^n(X_1^n)} \approx 2^{nD(P\|Q)}, \qquad \text{with high probability,}$$

so the random string X_1^n looks exponentially more likely under its true distribution P compared to the alternative Q.

Formally, (11) says that, for any $\epsilon > 0$,

$$\mathbb{P}\left\{\left|\frac{1}{n}\log\frac{P^n(X_1^n)}{Q^n(X_1^n)} - D(P\|Q)\right| > \epsilon\right\} \to 0, \quad \text{as } n \to \infty.$$

In other words, if we define a set of "likelihood-ratio typical" strings,

$$B_n^* = \left\{x_1^n \in A^n \ : \ 2^{n(D-\epsilon)} \le \frac{P^n(x_1^n)}{Q^n(x_1^n)} \le 2^{n(D+\epsilon)}\right\}, \tag{12}$$

where $D = D(P\|Q)$, then, for any $\epsilon > 0$,

$$P^n(B_n^{*c}) \to 0, \quad \text{as } n \to \infty. \tag{13}$$

Indeed, this B_n^* is the decision region in the direct part of the theorem.

Proof of Stein's lemma. For the direct part, consider the decision regions defined in (12). The result of the WLLN in (13) says exactly that $e_2^{(n)} = P^n(B_n^{*c}) \to 0$, as $n \to \infty$. To bound $e_1^{(n)}$, recall that $P^n(x_1^n)/Q^n(x_1^n) \ge 2^{n(D-\epsilon)}$ for each $x_1^n \in B_n^*$ so that,

$$1 \ge P^n(B_n^*) = \sum_{x_1^n \in B_n^*} P^n(x_1^n) = \sum_{x_1^n \in B_n^*} \frac{P^n(x_1^n)}{Q^n(x_1^n)} Q^n(x_1^n) \ge \sum_{x_1^n \in B_n^*} 2^{n(D-\epsilon)} Q^n(x_1^n),$$

from which we obtain that,

$$1 \ge 2^{n(D-\epsilon)} \sum_{x_1^n \in B_n^*} Q^n(x_1^n) = 2^{n(D-\epsilon)} Q^n(B_n^*) = 2^{n(D-\epsilon)} e_1^{(n)},$$

and which implies, $e_1^{(n)} \le 2^{-n(D-\epsilon)}$, as required.

For the converse, suppose $e_2^{(n)} = P^n(B_n^c) \to 0$ as $n \to \infty$. Since $P^n(B_n^{*c})$ also goes to zero by the direct part, we have that, as $n \to \infty$,

$$P^n(B_n \cap B_n^*) \ge 1 - P^n(B_n^c) - P_n(B_n^{*c}) \to 1,$$

so $P^n(B_n \cap B_n^*) \ge 1/2$, eventually. Therefore,

$$\frac{1}{2} \le P^n(B_n \cap B_n^*) = \sum_{x_1^n \in B_n \cap B_n^*} P^n(x_1^n) = \sum_{x_1^n \in B_n \cap B_n^*} Q^n(x_1^n) \frac{P^n(x_1^n)}{Q^n(x_1^n)} \le 2^{n(D+\epsilon)} \sum_{x_1^n \in B_n \cap B_n^*} Q^n(x_1^n),$$

where in the last step we used the fact that all the x_1^n in the sum belong to B_n^* so the ratio in the sum is always $\le 2^{n(D+\epsilon)}$. So, eventually,

$$\frac{1}{2} \le 2^{n(D+\epsilon)} Q^n(B_n \cap B_n^*) \le 2^{n(D+\epsilon)} Q^n(B_n) = 2^{n(D+\epsilon)} e_1^{(n)},$$

and rearranging gives (10). □

Remarks.

1. The assumption that $D \neq 0$ is trivial, since otherwise P and Q would be the same (see Theorem 2.5 below) and any test would always give the right answer. The assumption that D is finite is also quite reasonable; it means that there are no letters in the alphabet that have positive P-probability but zero Q-probability. If that were the case, then it would be fairly easy to check if the data came from P or Q, simply by checking whether any one of those letters appear in x_1^n or not.

2. In Stein's lemma the two error probabilities are treated in an asymmetric fashion. For one we compute the precise exponential rate at which it goes to zero, while for the other we simply assume it vanishes, possibly at an arbitrarily slow rate. This type of asymmetry is quite natural in many applications, where making an error in one direction is much more problematic than in the other. For example, it is much more important that a doctor *not* miss the presence of cancer, say, in a patient's medical tests, than detecting cancer where there is none.

3. Although we will not pursue this direction further, we remark that it is possible to make both error probabilities exponentially small simultaneously.

4. Stein's lemma states that, if one of the error probabilities is assumed to vanish, then the fastest exponential rate at which the other one can vanish is $D(P\|Q)$. One useful and satisfying interpretation of this fact is that, the larger the "distance" $D(P\|Q)$ between the two distributions, the smaller the error probability, and hence the easier it is to "tell them apart" from looking at the data. In this sense, $D(P\|Q)$ can be seen as natural "distance" between probability distributions, and a reasonable measure of their "distinguishability." We shall have more to say about this in the next section.

5. Although we showed that the decision regions B_n^* defined in (12) provide asymptotically optimal error probabilities, they may not be optimal for finite n, in the sense that there may be tests with error probabilities that are both smaller than those achieved by B_n^*. The exact form of the actual optimal regions is described in Theorem 2.3.

2.2 Relative entropy and optimal hypothesis testing

Theorem 2.3 (Neyman-Pearson lemma)**.** *For a hypothesis test between two distributions P and Q, based on n data samples, consider the* likelihood ratio decision region*,*

$$B_{\rm NP} = \left\{ x_1^n \in A^n \ : \ \frac{P^n(x_1^n)}{Q^n(x_1^n)} \geq T \right\}, \tag{14}$$

for some threshold $T > 0$, with error probabilities,

$$e_{1,\rm NP}^{(n)} = Q^n(B_{\rm NP}) \quad \textit{and} \quad e_{2,\rm NP}^{(n)} = P^n(B_{\rm NP}^c).$$

The decision region $B_{\rm NP}$ is optimal:

For any other decision region $B_n \subset A^n$, if one of its probabilities of error, $e_2^{(n)} = P^n(B_n^c)$, is as good as that of $B_{\rm NP}$,

$$e_2^{(n)} \leq e_{2,\rm NP}^{(n)},$$

then the other one, $e_1^{(n)} = Q^n(B_n)$, is worse:

$$e_1^{(n)} \geq e_{1,\rm NP}^{(n)}.$$

The proof is based on a very simple and elegant argument. For an arbitrary set B, recall the definition of the indicator function $\mathbb{I}_B$:

$$\mathbb{I}_B(x) = \begin{cases} 1, & \text{if } x \in B, \\ 0, & \text{if } x \notin B. \end{cases}$$

PROOF. Observe that the following inequality always holds:

$$\Big[\mathbb{I}_{B_{\rm NP}}(x_1^n) - \mathbb{I}_{B_n}(x_1^n)\Big]\Big[P^n(x_1^n) - TQ^n(x_1^n)\Big] \geq 0. \tag{15}$$

If the first term of the product is zero, then the whole thing is zero. If it is > 0, then x_1^n necessarily belongs to $B_{\rm NP}$, so the second term will be ≥ 0 by the definition of $B_{\rm NP}$. And similarly if the first term is < 0 then the second one is also < 0.

Now, multiplying out the terms in (15) and summing over all $x_1^n \in A^n$, yields,

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

or, equivalently,

$$T\Big[e_1^{(n)} - e_{1,\rm NP}^{(n)}\Big] - \Big[e_{2,\rm NP}^{(n)} - e_2^{(n)}\Big] \geq 0.$$

The result is an immediate consequence of this inequality. □

Before closing our discussion of hypothesis testing, we note that the Neyman-Pearson lemma admits an information-theoretic interpretation. Let $\hat{P}_n$ denote the empirical distribution induced by x_1^n on A^n, so that each $\hat{P}_n(a)$ is simply the frequency with which a appears in x_1^n:

$$\hat{P}_n(a) := \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{\{a\}}(x_i), \qquad a \in A. \tag{16}$$

Proposition 2.4 (Neyman-Pearson and relative entropy)**.** *The Neyman-Pearson decision region $B_{\rm NP}$ can equivalently be expressed as,*

$$B_{\rm NP} = \Big\{x_1^n \in A^n \;:\; D(\hat{P}_n\|Q) \geq D(\hat{P}_n\|P) + T'\Big\}, \tag{17}$$

with $T' = \frac{1}{n}\log T$.

In view of (17), informally speaking, the Neyman-Pearson test says that the optimal hypothesis test between P and Q consists of simply computing the empirical distribution $\hat{P}_n$ of the data and checking whether it is closer to P or to Q, in terms of the relative entropy "distance."

PROOF. We can express the logarithm of the likelihood ratio in (14) as,

$$\begin{aligned}
\log\frac{P^n(x_1^n)}{Q^n(x_1^n)} &= \log\prod_{i=1}^n \frac{P(x_i)}{Q(x_i)} \\
&= \sum_{i=1}^n \log\frac{P(x_i)}{Q(x_i)} \\
&= \sum_{i=1}^n\sum_{a\in A} \mathbb{I}_{\{a\}}(x_i)\log\frac{P(a)}{Q(a)} \\
&= n\sum_{a\in A}\frac{1}{n}\sum_{i=1}^n \mathbb{I}_{\{a\}}(x_i)\log\frac{P(a)}{Q(a)} \\
&= n\sum_{a\in A}\hat{P}_n(a)\log\Big(\frac{P(a)}{Q(a)}\frac{\hat{P}_n(a)}{\hat{P}_n(a)}\Big) \\
&= n\Big[D(\hat{P}_n\|Q) - D(\hat{P}_n\|P)\Big].
\end{aligned} \tag{18}$$

Substituting this identity in (14) yields (17). □

We close this section with some basic properties of the entropy and relative entropy. We will have much more to say about them in Sections 3 and 5.

Theorem 2.5 (Basic entropy bounds)**.** *(a) If $X \sim P$ is a discrete RV with values in the finite alphabet A:*

$$0 \le H(X) \le \log |A|.$$

Moreover, $H(X) = 0$ iff X is constant, and $H(X) = \log |A|$ iff X is uniformly distributed on A.

(b) If P, Q are PMFs on the same finite alphabet A, then,

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

with equality iff $P = Q$.

The proof of the theorem is based on the following useful bound, which is an easy consequence of Jensen's inequality:

Theorem 2.6 (Log-sum inequality)**.** *For arbitrary nonnegative constants, $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$, we have,*

$$\sum_i a_i \log \frac{a_i}{b_i} \ge \Big(\sum_i a_i \Big) \log \frac{(\sum_i a_i)}{(\sum_i b_i)},$$

with the convention that $0 \log 0 = 0 \log \frac{0}{0} = 0$. Moreover, we have equality if and only if the ratios a_i/b_i are the same for all $i = 1, 2, \ldots, n$.

Proof. Write $A = \sum_i a_i$ and $B = \sum_i b_i$. We assume without loss of generality that B and all the a_i are nonzero. (Check that this is indeed without loss of generality.) Define a RV:

$$X = \frac{a_i}{b_i}, \quad \text{with probability } \frac{b_i}{B}, \quad \text{for each } i = 1, 2, \ldots, n.$$

Let f be the function defined by $f(x) = x \log x$, for $x > 0$, and $f(0) = 0$, and note that it is strictly convex (e.g., its second derivative is strictly positive everywhere). Then,

$$\mathbb{E}(X) = \sum_i \frac{b_i}{B} \frac{a_i}{b_i} = \frac{A}{B},$$

similarly,

$$\mathbb{E}\big(f(X)\big) = \sum_i \frac{b_i}{B} \times \frac{a_i}{b_i} \log \frac{a_i}{b_i} = \frac{1}{B} \sum_i a_i \log \frac{a_i}{b_i},$$

and,

$$f\big(\mathbb{E}(X)\big) = \mathbb{E}(X) \log \mathbb{E}(X) = \frac{A}{B} \log \frac{A}{B}.$$

Then Jensen's inequality gives,

$$\frac{1}{B} \sum_i a_i \log \frac{a_i}{b_i} \ge \frac{A}{B} \log \frac{A}{B},$$

which is exactly the log-sum inequality. Since f is *strictly* convex the condition for equality is exactly the condition that X is constant, form Jensen's inequality. □

PROOF OF THEOREM 2.5. Since $\sum_a P(a) = \sum_a Q(a) = 1$, part (b) is an immediate consequence of the log-sum inequality. For the same reason, the condition for equality, i.e., $P(a)/Q(a) = \text{constant}$, can only be satisfied with constant $= 1$, i.e., iff $P = Q$.

The first inequality in (a) was already noted in Remark 3 in Section 1.1. For the second inequality, suppose $X \sim P$ and let Q be the uniform PMF $Q(a) = 1/|A|$, $a \in A$. Then,

$$D(P\|Q) = \sum_{x\in A} P(x) \log \frac{P(x)}{1/|A|} = \log|A| - H(P). \tag{19}$$

The second inequality in (a) now follows from (b). □

Remarks.

6. Suppose the size of the alphabet $|A| = m$. The fact that $H(X) \le \log m$ with equality iff X is uniformly distributed has a clear interpretation in terms of data compression. Any string in x_1^n from A can be described without any compression, simply by giving the binary representation of its index among all $|A^n| = m^n$ possible strings, with rate $\lceil n \log m \rceil / n \approx \log m$ bits/symbol. On the other hand, Theorem 1.6 states that it can be compressed with rate $H(X)$ bits/symbol, which is always better than $\log m$, unless X is uniformly distributed. In other words, a memoryless source can always be compressed, unless it is completely random.

7. Although stated for finite alphabets, it is easy to see that the results of Theorems 2.6 and 2.5 remain valid in the case of countably infinite alphabets.

3 Properties of entropy and relative entropy

In this section we describe the main properties of the entropy and relative entropy. We begin with an elementary but important observation.

Random variables and vectors. So far we have only defined the entropy $H(X)$ of a single discrete RV X. But any finite collection X_1^n of discrete RVs $(X_1, X_2, \ldots, X_n)$ with corresponding alphabets $A_1, A_2, \ldots, A_n$, say, can also be viewed as a single discrete RV: If all the A_i are finite or countably infinite, then the alphabet of X_1^n is the Cartesian product $A_1 \times A_2 \times \cdots \times A_n$, which is also a finite or countably infinite set. Therefore, we can simply define the *joint entropy* of a pair (X, Y) or of an arbitrary finite collection X_1^n of discrete RVs as the entropy of their joint PMF. Similar comments apply to the definitions of conditional entropy, mutual information, and all other information-theoretic functionals in this and in the following sections.

Example 3.1. Suppose X, Y are discrete RVs with alphabets A, B, respectively, and joint PMF $P_{XY}(x, y)$, $x \in A, y \in B$. Then their joint entropy is,

$$H(X,Y) = H(P_{XY}) = -\sum_{x\in A, y\in B} P_{XY}(x,y) \log P_{XY}(x,y) = \mathbb{E}[-\log P_{XY}(X,Y)]. \tag{20}$$

If X, Y are independent, then $P_{XY}(x, y) = P_X(x)P_Y(y)$ for all $(x, y) \in A \times B$, where P_X and P_Y are the (marginal) PMFs of X, Y, respectively, so that,

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

This makes sense; recalling the interpretation of the entropy as a measure of the (average) amount of information contained in a RV, then the information in two *independent* RVs (X, Y) is the sum of their individual information contents.

Now suppose X, Y are not independent, but have joint PMF given by:

X / Y	1	2	3	
0	1/10	1/5	1/4	11/20
1	1/5	1/20	1/5	9/20
	3/10	1/4	9/20	

We can compute that,

$$\begin{aligned} H(X) &= -\frac{3}{10}\log\frac{3}{10} - \frac{1}{4}\log\frac{1}{4} - \frac{9}{20}\log\frac{9}{20} \approx 1.539 \quad \text{bits}, \\ \text{and} \quad H(Y) &= -\frac{11}{20}\log\frac{11}{20} - \frac{9}{20}\log\frac{9}{20} \approx 0.993 \quad \text{bits}, \end{aligned}$$

whereas,

$$H(X,Y) = -\frac{1}{10}\log\frac{1}{10} - \frac{1}{5}\log\frac{1}{5} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{5}\log\frac{1}{5} - \frac{1}{20}\log\frac{1}{20} - \frac{1}{5}\log\frac{1}{5} \approx 2.441 \quad \text{bits}.$$

Observe that $H(X, Y) \approx 2.441$ is strictly smaller than $H(X) + H(Y) \approx 2.532$. That is, the amount of information in the pair (X, Y) is smaller than the sum of the amounts of information in X and Y taken individually. Again, this makes sense: Since X and Y are not independent, if we already know the value of X, say, then the amount of information we obtain from Y is smaller than $H(Y)$. Formally this can be expressed as follows. Recalling that the joint PMF

of (X, Y) can always be written, $P_{XY}(x, y) = P_X(x)P_{Y|X}(y|x)$, where $P_{Y|X}(\cdot|x)$ denotes the conditional PMF $P_{Y|X}(\cdot|x) = \mathbb{P}(Y = \cdot|X = x)$ of Y given $X = x$, we have,

$$\begin{aligned} H(X, Y) &= \mathbb{E}[-\log P_{XY}(X, Y)] \\ &= \mathbb{E}[-\log P_X(X) - \log P_{Y|X}(Y|X)] \\ &= H(X) + \mathbb{E}[-\log P_{Y|X}(Y|X)]. \end{aligned} \tag{21}$$

This motivates the following definition.

Definition 3.2 (Conditional entropy)**.** *If X, Y are discrete RVs with joint PMF P_{XY}, then the* conditional entropy *$H(Y|X)$ of X given Y is,*

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

where $P_{Y|X}$ is the conditional PMF of Y given X. Alternatively, as shown in (21):

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

For example, it is easy to compute, either via (21) or by direct calculation, that the amount of information in Y given X in the above example, is,

$$H(Y|X) \approx 0.902 \quad \text{bits}.$$

The fact that, here, $H(Y|X)$ is smaller than $H(Y) \approx 0.993$ bits, is intuitively satisfying: Knowing the information in a correlated RV X, reduces the amount of information revealed by Y. This is a general fact, as shown in Proposition 3.4 in the following section.

3.1 Entropy

The result of Proposition 3.3 stated next follows by inductively repeating the computations in equations (20) and (21) in Example 3.1 above, upon noting that the joint PMF $P_{X_1^n}$ of a collection of RVs X_1^n can be expanded, in the obvious notation, as:

$$P_{X_1^n}(x_1^n) = P_{X_1}(x_1)P_{X_2|X_1}(x_2|x_1) \cdots P_{X_n|X_1^{n-1}}(x_n|x_1^{n-1}) = \prod_{i=1}^{n} P_{X_i|X_1^{i-1}}(x_i|x_1^{i-1}).$$

Proposition 3.3 (Chain rule for entropy)**.** *For any finite collection X_1^n of discrete random variables $(X_1, X_2, \ldots, X_n)$:*

$$H(X_1^n) = \sum_{i=1}^{n} H(X_i|X_1^{i-1}) = H(X_1) + H(X_2|X_1) + \cdots + H(X_n|X_1^{n-1}).$$

If the X_1^n are independent:

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

Next we show that, as observed in Example 3.1, conditioning always reduces the entropy.

Proposition 3.4 (Conditioning reduces entropy)**.** *For any two discrete RVs X, Y,*

$$H(Y|X) \leq H(Y),$$

with equality iff X and Y are independent.

PROOF. In the notation of Definition 3.2, and letting P_X and P_Y denote the marginal PMFs of X and Y, respectively, we have,

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

Therefore, the difference $H(Y) - H(Y|X)$ is equal to the relative entropy between P_{XY} and the product PMF P_XP_Y, which, by Theorem 2.5 (b) is always nonnegative, with equality iff $P_{XY} = P_XP_Y$, i.e., iff X and Y are independent. □

Note that the conditional entropy $H(Y|X)$ can also be written as the average of the entropies of the conditional PMFs $P_{Y|X}(\cdot|x)$, over all x:

$$H(Y|X) = -\sum_{x,y} P_X(x)P_{Y|X}(y|x)\log P_{Y|X}(y|x) = \sum_x P_X(x)H(P_{Y|X=x}).$$

Therefore, writing B for the alphabet of Y, Theorem 2.5 (a) implies that we always have:

$$0 \le H(Y|X) \le \log|B|.$$

Next we establish some simple properties of the entropy of a function of a RV X. The main message here is that impossible to "add" more information to data simply by processing it.

Proposition 3.5 (Data processing inequality for entropy)**.** *Let X be a discrete RV on the alphabet A, and f be an arbitrary function on A. We have:*

(a) *$H(f(X)) \le H(X)$, with equality iff f is 1-1.*

(b) *$H(f(X)|X) = 0$.*

PROOF. Recall that in Remark 2 of Section 1.1 we noted that $H(f(X)) = H(X)$ if f is 1-1. Since for any function f the map $x \mapsto (x, f(x))$ is injective, we have that $H(f(X), X) = H(X)$, and combining with the chain rule,

$$H(X) = H(X, f(X)) = H(X) + H(f(X)|X).$$

This proves (b).

Now, applying the chain rule in the reverse direction we similarly obtain,

$$H(X) = H(X, f(X)) = H(f(X)) + H(X|f(X)) \ge H(f(X)), \tag{22}$$

where the last step follows from the nonnegativity of conditional entropy. Therefore, we always have $H(f(X)) \le H(X)$, and we know that equality holds if f is 1-1. For the "only if" part note that equality in (22) implies that $H(X|f(X)) = 0$, which can only happen if X is a deterministic function of $f(X)$, i.e., only if f is invertible. □

The following corollary is an immediate consequence of the chain rule in Proposition 3.3 combined with the Proposition 3.4.

Corollary 3.6 (Subadditivity of entropy)**.** *For any finite collection X_1^n of discrete random variables $(X_1, X_2, \ldots, X_n)$,*

$$H(X_1^n) \leq \sum_{i=1}^{n} H(X_i),$$

with equality iff the X_1^n are independent.

The following proposition states the natural generalisations of the results in Propositions 3.3, 3.4 and Corollary 3.6 to the case of conditional entropy. Its proof is left as an exercise.

Proposition 3.7 (Conditional entropy properties)**.** *For arbitrary discrete RVs X, Y, Z:*

(*i*) *(Chain rule) $H(X, Z|Y) = H(X|Y) + H(Z|X, Y)$*

(*ii*) *(Subadditivity) $H(X, Z|Y) \leq H(X|Y)+H(Z|Y)$ with equality iff X and Z are conditionally independent given Y.*

(*iii*) *(Conditioning reduces entropy) $H(X|Y, Z) \leq H(X|Y)$ with equality iff X and Z are conditionally independent given Y.*

Finally we present a simple but very useful inequality for the conditional entropy, which also admits an interesting interpretation. Suppose we are interested in the value of some RV X but we only get to observe a correlated RV Y, and we produce an estimate $\hat{X} = f(Y)$ of X based on Y. The probability of error of this estimate is $P_e := \mathbb{P}(\hat{X} \neq X)$. Fano's inequality says that, if P_e is small, that is, if we can estimate X accurately from Y, then the information $H(X|Y)$ in X given Y is small:

Theorem 3.8 (Fano's inequality)**.** *Let (X, Y) be an arbitrary pair of discrete RVs taking values in the finite alphabets A, B, respectively, let $\hat{X} = f(Y)$ for some function $f : B \to A$, and write $P_e := \mathbb{P}(\hat{X} \neq X)$. Then:*

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

where $h(p)$ denotes the binary entropy function; cf. (1)*.*

Proof. Let E denote the binary RV that takes the value 1 if there is an error, i.e., if $\hat{X} \neq X$, and is equal to zero otherwise. Then $E \sim \text{Bern}(P_e)$ and its entropy $H(E) = h(P_e)$. From the chain rule we have,

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

since E is a function of (X, Y). And using the chain rule in the other direction we similarly obtain,

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

since conditioning reduces entropy. Combining the above two equations yields,

$$H(X|Y) - h(P_e) \leq H(X|E, Y) = P_e H(X|E = 1, Y) + (1 - P_e)H(X|E = 0, Y).$$

But $E = 0$ means that, given Y, we can determine the value of X as $f(Y)$, so $H(X|E = 0, Y) = 0$. Similarly, $E = 1$ means that, given Y, we know that X takes one of the $|A| - 1$ possible values other than $f(Y)$, so $H(X|E = 1, Y) \leq \log(|A| - 1)$. Therefore,

$$H(X|Y) - h(P_e) \leq P_e \log(|A| - 1),$$

as claimed. □

3.2 Relative entropy

Returning for a moment to the hypothesis testing problem of Section 2, suppose we have data X and we wish to determine whether X came from distribution P or Q. And suppose also that someone suggests they can "help" us by first "processing" the data, giving us a processed version $f(X)$, say. The following result says that, no matter what the processor does, they cannot possibly make the two distributions more distinguishable: The relative entropy distance between the distributions of the processed data can only be smaller than that of the original data. This further reinforces our interpretation of the relative entropy as a measure of how different or how "distinguishable" two distributions are.

Theorem 3.9 (Data processing inequality for relative entropy)**.** *Let $X \sim P_X$ and $X' \sim Q_{X'}$ be two RVs with PMFs P_X and $Q_{X'}$, respectively, on the same alphabet A. Let $f : A \to B$ be an arbitrary function on A, and write $P_{f(X)}$ and $Q_{f(X')}$ for the corresponding PMFs of $f(X)$ and $f(X')$. Then:*

$$D(P_{f(X)} \| Q_{f(X')}) \leq D(P_X \| Q_{X'}).$$

Proof. Consider the set B of all possible values of $f(X)$ or $f(X')$. To each element $y \in B$ we can associate a subset $A_y \subset A$ that contains those values x such that $f(x) = y$, so that,

$$P_{f(X)}(y) = \sum_{x \in A_y} P_X(x) \quad \text{and} \quad Q_{f(X')}(y) = \sum_{x \in A_y} Q_{X'}(x).$$

Then we can write,

$$\begin{aligned}
D(P_{f(X)} \| Q_{f(X')}) &= \sum_{y \in B} P_{f(X)}(y) \log \frac{P_{f(X)}(y)}{Q_{f(X')}(y)} \\
&= \sum_{y \in B} \Big[\sum_{x \in A_y} P_X(x) \Big] \log \frac{\sum_{x \in A_y} P_X(x)}{\sum_{x \in A_y} Q_{X'}(x)} \\
&\overset{(a)}{\leq} \sum_{y \in B} \sum_{x \in A_y} P_X(x) \log \frac{P_X(x)}{Q_{X'}(x)} \\
&\overset{(b)}{=} \sum_{x \in A} P_X(x) \log \frac{P_X(x)}{Q_{X'}(x)} \;=\; D(P_X \| Q_{X'}),
\end{aligned}$$

where in (a) we applied the log-sum inequality to each term in the sum over y, and for (b) we observe that summing over each $y \in B$ and then over each $x \in A_y$ is the same as summing over all $x \in A$. □

As noted earlier, it is useful to think of the relative entropy as a measure of "distance" or "statistical dissimilarity" between distributions, but $D(P\|Q)$ is not a proper metric. [You should find easy examples where the symmetry and triangle inequality axioms for a metric are indeed violated by $D(P\|Q)$.]

On the other hand, relative entropy does have the useful property that $D(P\|Q) \geq 0$ with equality if and only if $P = Q$, and it can also provide quantitative bounds on the actual differences of the probabilities $|P(x) - Q(x)|$. This is made precise by the Pinsker's inequality, established next.

Definition 3.10 (Total variation)**.** *The* total variation distance $\|P - Q\|_{\mathrm{TV}}$ *between two PMFs P and Q on the same alphabet A is:*

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

A useful alternative representation of the total variation distance is the following. Let $B = \{x \in A \,:\, P(x) > Q(x)\}$. Then:

$$\begin{aligned} \|P - Q\|_{\mathrm{TV}} &= \sum_{x \in A} |P(x) - Q(x)| \\ &= \sum_{x \in B} (P(x) - Q(x)) + \sum_{x \in B^c} (Q(x) - P(x)) \\ &= P(B) - Q(B) + Q(B^c) - P(B^c) \\ &= 2\big(P(B) - Q(B)\big). \end{aligned} \tag{23}$$

Theorem 3.11 (Pinsker's inequality)**.** *For any two PMFs P, Q on the same alphabet A:*

$$\|P - Q\|_{\mathrm{TV}}^2 \le (2 \log_e 2) D(P\|Q).$$

Proof. First consider the case when P and Q are the Bern(p) and Bern(q) PMFs, respectively. Despite the fact that $D(P\|Q)$ is non symmetric, we can assume without loss of generality that $q \le p$ for, otherwise, we can simply interchange $p \leftrightarrow (1-p)$ and $q \leftrightarrow (1-q)$. Let $\Delta(p, q)$ denote the difference,

$$\Delta(p, q) = (2 \log_e 2) D(P\|Q) - \|P - Q\|_{\mathrm{TV}}^2 = 2p \log_e \frac{p}{q} + 2(1-p) \log_e \frac{1-p}{1-q} - \big(2(p-q)\big)^2,$$

so that our goal is to show that, $\Delta(p, q) \ge 0$, for all $0 \le q \le p \le 1$. And since $\Delta(p, p) = 0$ for all p, it suffices to show that $\partial \Delta(p, q)/\partial q \le 0$. Indeed, differentiating and simplifying,

$$\frac{\partial \Delta(p, q)}{\partial q} = -\frac{2p}{q} + \frac{2(1-p)}{(1-q)} + 8(p-q) = 2(q-p) \left(\frac{1}{q(1-q)} - 4 \right),$$

which is indeed ≤ 0 since the first factor $(q-p)$ in the right-hand side above is ≤ 0, and the second factor is ≥ 0 because $q(1-q) \le 1/4$ for all $q \in [0, 1]$.

The case of two general PMFs P, Q can be reduced to the Bernoulli case as follows. Recall the alternative representation of the total variation distance in (23) in terms of the set B. Define two RVs, $X \sim P$ and $X' \sim Q$, and let $f(x) = \mathbb{I}_B(x)$. We will apply the data processing inequality of Theorem 3.9. Write P_f and Q_f for the PMFs of the RVs $f(X)$ and $f(X')$, respectively, and note that $f(X) \sim$ Bern(p) and $f(X') \sim$ Bern(q), with $p = P(B)$ and $q = Q(B)$.

We then have,

$$\begin{aligned} (2 \log_e 2) D(P\|Q) &\overset{(a)}{\ge} (2 \log_e 2) D(P_f\|Q_f) \\ &\overset{(b)}{\ge} \|P_f - Q_f\|_{\mathrm{TV}}^2 \\ &\overset{(c)}{=} \big(2(p-q)\big)^2 \\ &= \big(2(P(B) - Q(B))\big)^2 \\ &\overset{(d)}{=} \|P - Q\|_{\mathrm{TV}}^2, \end{aligned}$$

where (a) follows from Theorem 3.9; (b) is the binary special case of this theorem already established; in (c) we simply compute the TV distance between two Bernoulli distributions; and (d) follows from the alternative representation (23). □

We close this section with two simple convexity results for the entropy and relative entropy. The proof of Theorem 3.12 is left as an exercise; it is similar to, and somewhat simpler than, that of Theorem 3.9.

Theorem 3.12 (Convexity of relative entropy)**.** *Relative entropy $D(P\|Q)$ is jointly convex in its two arguments: For any P_1, P_2, Q_2, Q_2 and any $0 < \lambda < 1$:*

$$D\big(\lambda P_1 + (1-\lambda)P_2\big\|\lambda Q_1 + (1-\lambda)Q_2\big) \le \lambda D(P_1\|Q_1) + (1-\lambda)D(P_2\|Q_2).$$

Recall that we saw earlier in (19) that, writing Q for the uniform PMF on an alphabet A, we have, $D(P\|Q) = \log m - H(P)$. This together with Theorem 3.12 immediately imply the following:

Corollary 3.13 (Concavity of entropy)**.** *The entropy $H(P)$ is a concave function on the set of all PMFs P on a finite alphabet.*

In fact, as shown in Exercise 12.8, $H(P)$ is a *strictly* concave function of P. Also, it is easy to see that all the results of this section remain valid in the case of infinite alphabets, under the additional assumption that all entropies involved are finite and all relative entropies involved satisfy (9).

4 Poisson approximation

Before continuing our study of information measures and their properties, we take a small detour and present an application of some of the results we have seen so far to the classical problem of Poisson approximation.

Probability theory is, to a large extent, based on the premise that "a bit of randomness is random, a lot of randomness brings order." If you toss a fair coin a few times it is nearly impossible to make any predictions about the results. But if you toss it a thousand times, you can be quite certain that the proportion of Heads will be close to 50%.

Perhaps the most important instantiation of this phenomenon is in the central limit theorem: If you look at the average of a lot of independent random quantities with similar statistical characteristics, its distribution will be close to a Gaussian.

The other basic limit theorem one encounters in elementary probability is the binomial-to-Poisson convergence. It says that, if each one of a large number, n say, of independent possible events occurs with probability λ/n. Then the distribution of total number of events that occur is close to $\mathrm{Po}(\lambda)$, the Poisson distribution with mean λ.

4.1 Poisson approximation via entropy

Suppose $X_1, X_2, \ldots, X_n$ are IID $\mathrm{Bern}(\lambda/n)$ RVs and let $S_n = X_1 + X_2 + \cdots + X_n$ denote their sum. Then the distribution of S_n converges to the $\mathrm{Po}(\lambda)$:

$$\mathbb{P}(S_n = k) \to e^{-\lambda}\frac{\lambda^k}{k!}, \quad \text{as } n \to \infty, \text{ for each } k \geq 0. \tag{24}$$

This phenomenon occurs much more generally: The distribution P_{S_n} of the sum $S_n = X_1 + X_2 + \cdots + X_n$ of possibly dependent RVs $X_i \sim \mathrm{Bern}(p_i)$ is close to the $\mathrm{Po}(\lambda)$, with $\lambda = \mathbb{E}(S_n) = p_1 + p_2 + \cdots + p_n$, as long as:

(i) All the p_i are small;

(ii) The X_i are weakly dependent.

A version of this result is proved in Theorem 4.3 below, essentially only using the data processing property of relative entropy. This result is more general than (24), it establishes convergence in a stronger sense, and its proof is simpler than the usual argument used to establish (24).

We find it easier in this section to work with the relative entropy $D_e(P\|Q)$, defined with the natural logarithm $\log_e$ in place of $\log = \log_2$:

$$D_e(P\|Q) = \sum_x P(x)\log_e\frac{P(x)}{Q(x)} = (\log_e 2)D(P\|Q).$$

Since $D_e(P\|Q)$ is simply $D(P\|Q)$ times a constant, it enjoys the same properties as $D(P\|Q)$. In particular, all the results of the previous section remain valid for $D_e(P\|Q)$, and Pinsker's inequality takes the cleaner form,

$$D_e(P\|Q) \geq \frac{1}{2}\|P - Q\|_{\mathrm{TV}}^2.$$

Theorem 4.1 (Poisson approximation #1)**.** *Suppose $S_n = X_1 + X_2 + \cdots + X_n$ is the sum of n independent Bernoulli RVs, with each $X_i \sim \mathrm{Bern}(p_i)$. Then,*

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

with $\lambda = \mathbb{E}(S_n) = p_1 + p_2 + \cdots + p_n$,

Remarks.

1. The result of Theorem 4.1 easily implies the binomial-to-Poisson convergence (24): If all the p_i are equal to λ/n, then, writing P_λ for the $\mathrm{Po}(\lambda)$ PMF, using Pinsker's inequality we have,
$$\|P_{S_n} - P_\lambda\|_{\mathrm{TV}}^2 \leq 2D_e(P_{S_n}\|\mathrm{Po}(\lambda)) \leq 2\sum_{i=1}^n \frac{\lambda^2}{n^2} = \frac{2\lambda^2}{n}.$$
Therefore, for each k,
$$|\mathbb{P}(S_n = k) - P_\lambda(k)| \leq \|P_{S_n} - P_\lambda\|_{\mathrm{TV}} \leq \frac{\sqrt{2}\lambda}{\sqrt{n}}, \tag{25}$$
which of course goes to zero as $n \to \infty$.

2. Theorem 4.1 is stronger than (24) in that it is not asymptotic. In particular, instead of asymptotic binomial-to-Poisson convergence, we get the inequality (25) which holds for *all* n. It is also stronger than (24) in that it provides an easily computable, quantitative bound on the difference between P_{S_n} and the limiting Poisson distribution. This makes it useful in applications. Moreover, Theorem 4.1 is more general than (24), since it does not assume all the p_i are equal.

3. On a more theoretical note it is interesting to mention that Theorem 4.1 captures the "essence" of Poisson convergence in the following sense: It is known[1] that, under the hypotheses of Theorem 4.1, the distribution of S_n converges to a Poisson if and only if $\sum_i p_i^2 \to 0$ as $n \to \infty$. Therefore, the condition for Poisson convergence implied by Theorem 4.1 is best possible.

PROOF. Let $A = \{0, 1, \ldots\}$, and define n independent RVs $Z_1, Z_2, \ldots, Z_n$, with each $Z_i \sim \mathrm{Po}(p_i)$, so that the distribution P_{T_n} of their sum $T_n = Z_1 + Z_2 + \cdots + Z_n$ is $\mathrm{Po}(\lambda)$. Now, let us think of the collection X_1^n as a single RV taking values in A^n. Letting $f : A^n \to A$ be the function,
$$f(x_1^n) = \sum_{i=1}^n x_i, \quad x_1^n \in A^n,$$
we have that $S_n = f(X_1^n)$ and $T_n = f(Z_1^n)$. By the data processing property of relative entropy in Theorem 3.9 (and its extension to countably infinite alphabets in Exercise 12.9),
$$D_e\big(P_{S_n}\big\|\mathrm{Po}(\lambda)\big) = D_e\big(P_{f(X_1^n)}\big\|P_{f(Z_1^n)}\big) \leq D_e\big(P_{X_1^n}\big\|P_{Z_1^n}\big),$$
where we used the obvious notation $P_{X_1^n}$ for the joint distribution of X_1^n and similarly for $P_{Z_1^n}$.

Writing P_{X_i} for the $\mathrm{Bern}(p_i)$ distribution of each X_i and P_{Z_i} for the $\mathrm{Po}(p_i)$ distribution of each Z_i, we can write out the relative entropy as an expectation, to obtain,
$$D_e\big(P_{S_n}\big\|\mathrm{Po}(\lambda)\big) \leq \mathbb{E}\left[\log_e \frac{P_{X_1^n}(X_1^n)}{P_{Z_1^n}(X_1^n)}\right] = \mathbb{E}\left[\log_e\left(\prod_{i=1}^n \frac{P_{X_i}(X_i)}{P_{Z_i}(X_i)}\right)\right], \tag{26}$$
where we used the independence of the X_1^n and of the Z_1^n to write their distributions as products. Therefore,
$$D_e\big(P_{S_n}\big\|\mathrm{Po}(\lambda)\big) \leq \sum_{i=1}^n \mathbb{E}\left[\log_e \frac{P_{X_i}(X_i)}{P_{Z_i}(X_i)}\right] = \sum_{i=1}^n D_e\big(P_{X_i}\big\|P_{Z_i}\big) = \sum_{i=1}^n D_e\big(\mathrm{Bern}(p_i)\big\|\mathrm{Po}(p_i)\big). \tag{27}$$
The proof is completed by substituting in (27) the elementary bound stated in the following lemma. Its proof is left as an exercise. □

[1] This result is way beyond the scope of these notes, but if you are interested you can find out much more about it and about Poisson approximation in general in the monograph [2].

Lemma 4.2. *For every $p \in [0,1]$: $D_e\big(\mathrm{Bern}(p)\big\|\mathrm{Po}(p)\big) \le p^2$.*

Next we describe a more general Poisson approximation bound that works for sums of dependent Bernoulli RVs. As with the relative entropy, we find it more convenient to work with the entropy $H_e(X)$ defined in terms of natural logarithms below: If $X \sim P$ on A,

$$H_e(X) = -\sum_{x\in A} P(x)\log_e P(x) = (\log_e 2)H(X).$$

Theorem 4.3 (Poisson approximation #2)**.** *Suppose $S_n = X_1 + X_2 + \cdots + X_n$ is the sum of n possibly dependent Bernoulli RVs, with each $X_i \sim \mathrm{Bern}(p_i)$. Then, if $\mathbb{E}(S_n) = \lambda$,*

$$D_e(P_{S_n}\|\mathrm{Po}(\lambda)) \le \sum_{i=1}^n p_i^2 + \Big[\sum_{i=1}^n H_e(X_i) - H_e(X_1^n)\Big].$$

Before we give the proof we note that this result is quite satisfying, as it gives a precise, quantitative version of the assertion that was made in the beginning of this section: P_{S_n} is close to a Poisson distribution, as long as all the p_i are small and the X_i are weakly dependent. The first term above clarifies the first part: The p_i must be small enough that the sum of their squares be small. And the second term gives a quantitative version of what it means to be weakly dependent: The term $\sum_i H_e(X_i) - H_e(X_1^n)$ is always nonnegative, and it is only zero if the X_i are independent. So the closer it is to zero, the smaller the dependence between the X_i, and the closer the approximation of P_{S_n} by a Poisson distribution.

Proof. Following the same steps as in the proof of Theorem 4.1 leading to equation (26), we have:

$$D_e\big(P_{S_n}\big\|\mathrm{Po}(\lambda)\big) \le \mathbb{E}\left[\log_e \frac{P_{X_1^n}(X_1^n)}{P_{Z_1^n}(X_1^n)}\right] = \mathbb{E}\left[\log_e\left(\frac{P_{X_1^n}(X_1^n)}{\prod_{i=1}^n P_{Z_i}(X_i)} \times \frac{\prod_{i=1}^n P_{X_i}(X_i)}{\prod_{i=1}^n P_{X_i}(X_i)}\right)\right].$$

Therefore, $D_e\big(P_{S_n}\big\|\mathrm{Po}(\lambda)\big)$ can be bounded above by the sum of three terms:

$$\mathbb{E}\left[\log_e\left(\prod_{i=1}^n \frac{P_{X_i}(X_i)}{P_{Z_i}(X_i)}\right)\right] + \mathbb{E}\left[\log_e\left(\frac{1}{\prod_{i=1}^n P_{X_i}(X_i)}\right)\right] - \mathbb{E}\left[\log_e\left(\frac{1}{P_{X_1^n}(X_1^n)}\right)\right].$$

The first term is already treated in the proof of Theorem 4.1 where we found it is $\le \sum_i p_i^2$. The second term can be expanded into,

$$\mathbb{E}\left[\log_e\left(\frac{1}{\prod_{i=1}^n P_{X_i}(X_i)}\right)\right] = \sum_{i=1}^n \mathbb{E}\left[\log_e\left(\frac{1}{P_{X_i}(X_i)}\right)\right] = \sum_{i=1}^n H_e(X_i).$$

And the third term is simply $H_e(X_1^n)$. Combining these three yields the desired result. □

4.2 What is the Poisson distribution?

The importance of the Poisson distribution stems from its "inevitability" when we try to describe a random number of events happening "as randomly as possible." The way this is usually formulated in introductory classes is as follows. Suppose we want to model the number of events that occur in a certain time interval, say $[0,1]$, assuming only that the mean number of events, say λ, is known, and that they occur in the most random possible manner. A natural construction that describes these assumptions is to divide up $[0,1]$ into a large number n of subintervals, $[\frac{i-1}{n}, \frac{1}{n})$, $1 \le i \le n$, and to say that in each subinterval we either have an event

with probability λ/n, or no events with probability $(1-\lambda/n)$, independently between different subintervals.

Then the distribution of the total number of events S_n is $\mathrm{Bin}(n, \lambda/n)$, the binomial with parameters n and λ/n, which converges to the $\mathrm{Po}(\lambda)$ law as $n \to \infty$.

Here we present a different "maximally random" characterisation of the Poisson distribution. Theorem 4.4 states that the maximal entropy that can be achieved by sums $S_n = \sum_{i=1}^n X_i$ of independent $\mathrm{Bern}(p_i)$ RVs with mean $\mathbb{E}(S_n) = \lambda$, is the entropy of the $\mathrm{Po}(\lambda)$ distribution.

Theorem 4.4 (Poisson maximum entropy)**.**

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

The main technical ingredient of the proof is the following maximum entropy property of the binomial. Exercise 12.23 contains an outline of the proof of Proposition 4.5. Both of these results were established by Harremoës in [13].

Proposition 4.5 (Binomial maximum entropy)**.** *Let $B_n(\lambda)$ consist of all distributions on $\{0, 1, \ldots\}$ that arise from sums $\sum_{i=1}^n X_i$ of independent $X_i \sim \mathrm{Bern}(p_i)$ RVs, with $\sum_{i=1}^n p_i = \lambda$. Then for all $n \geq \lambda$:*

$$\sup\{H_e(P) \ : \ P \in B_n(\lambda)\} = H_e(\mathrm{Bin}(n, \lambda/n)).$$

Proof of Theorem 4.4. First we note that the supremum in the theorem can be written as,

$$H^* = \sup_{n \geq 1} \sup\{H_e(P) \ : \ P \in B_n(\lambda)\},$$

and that the collections $B_n(\lambda)$ satisfy $B_n(\lambda) \subset B_{n+1}(\lambda)$, for all n, so that, in view of Proposition 4.5,

$$H^* = \lim_{n \to \infty} \sup\{H_e(P) \ : \ P \in B_n(\lambda)\} = \lim_{n \to \infty} H_e(\mathrm{Bin}(n, \lambda/n)).$$

Therefore, it only remains to show that $H_e(\mathrm{Bin}(n, \lambda/n)) \to H_e(\mathrm{Po}(\lambda))$, as $n \to \infty$.

Let P_n and Q denote the $\mathrm{Bin}(n, \lambda/n)$ and $\mathrm{Po}(\lambda)$ PMFs, respectively. The simple bound $k! \leq k^k \leq e^{k^2}$ implies that $H_e(Q) = H_e(\mathrm{Po}(\lambda))$ is finite:

$$H_e(Q) = \sum_k Q(k) \log_e \frac{k!}{e^{-\lambda}\lambda^k} \leq \sum_k Q(k)[\lambda - k \log_e \lambda + k^2] = \lambda^2 + 2\lambda - \lambda \log_e \lambda < \infty,$$

where we used the fact that the first and second moments of the $\mathrm{Po}(\lambda)$ distribution are λ and $\lambda + \lambda^2$, respectively.

Below we give a proof of the fact that $H_e(P_n) \to H_e(Q)$, as $n \to \infty$, using the dominated convergence theorem. An alternative, direct, elementary proof is given in Exercise 12.24.

By Theorem 4.1 and Pinsker's inequality we know that, as $n \to \infty$, we have $P_n(k) \to Q(k)$ so that $-P_n(k) \log_e Q(k) \to -Q(k) \log_e Q(k)$. Also, for any $0 \leq k \leq n$,

$$P_n(k) = \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \leq \frac{n!}{n^k(n-k)!} \frac{\lambda^k}{k!} \leq Q(k) e^{\lambda}.$$

Therefore, $-P_n \log_e Q$ is dominated by $-e^{\lambda} Q \log_e Q$, which is summable because $H_e(Q)$ is finite, and by the dominated convergence theorem (with respect to the σ-finite counting measure) we have, as $n \to \infty$,

$$-\sum_k P_n(k) \log_e Q(k) \to H_e(Q). \tag{28}$$

Finally, from Theorem 4.1 we know that $D_e(P_n\|Q) \to 0$ as $n \to \infty$, and combining this with (28) we obtain that, as $n \to \infty$,

$$H_e(P_n) = -\sum_k P_n(k) \log_e P_n(k) = -\sum_k P_n(k) \log_e Q(k) - D_e(P_n\|Q) \to H_e(Q),$$

as required, completing the proof. □

5 What is mutual information?

Suppose X, Y are two correlated discrete RVs. How much common information is contained in both X and Y?

Recall that in Example 3.1 we had a pair of RVs X, Y with $H(X) \approx 1.539$ bits and,

$$H(X|Y) = H(X,Y) - H(Y) \approx 1.449 \quad \text{bits.}$$

We could interpret this as saying that, of the 1.539 bits of uncertainty in X, 1.449 bits remain uncertain after having observed Y. Therefore, we can think of the difference,

$$H(X) - H(X|Y) \approx 0.091 \quad \text{bits,}$$

and the amount of *mutual information* in X and Y.

Definition 5.1 (Mutual information)**.** *Let X, Y, Z be three discrete RVs. The* mutual information $I(X;Y)$ *between X and Y is:*

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

The conditional mutual information $I(X;Y|Z)$ *between X and Y given Z is:*

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

Before interpreting $I(X;Y)$ further, we note that it admits several different, useful representations.

Proposition 5.2 (Mutual information representations)**.** *For any two discrete RVs X, Y with joint PMF P_{XY} and marginal PMFs P_X and P_Y, respectively:*

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(X) + H(Y) - H(X,Y) \qquad (29) \\ &= H(Y) - H(Y|X) \\ &= D(P_{XY} \| P_X P_Y). \qquad (30) \end{aligned}$$

Proof. The first expression is the definition of $I(X;Y)$, the next two follow immediately by the chain rule, and the last one follows from (29) as,

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

□

Although the full glory of the operational meaning of $I(X;Y)$ comes from its connection with *channel capacity*, we note that the expressions in Proposition 5.2 already lead to several interesting interpretations. First we observe that, although not obvious from its definition, equation (29) shows that $I(X;Y)$ is in fact symmetric in X, Y. Moreover, the sum of the information contained in X and Y separately, minus the information contained in the pair, indeed describes the amount of *mutual* information shared by both. Finally, recalling Stein's lemma (Theorem 2.2), we can view $I(X;Y) = D(P_{XY} \| P_X P_Y)$ as measure of how well data generated from P_{XY} can be distinguished from independent pairs (X', Y') generated by the product distribution $P_X P_Y$. In this sense, $I(X;Y)$ can be though of as a measure of how far (X, Y) from being independent.

5.1 Basic properties

Proposition 5.3 (Chain rule for mutual information)**.** *For any collection of discrete RVs* $X_1, X_2, \ldots, X_n, Y$*:*

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

PROOF. This follows easily from the definitions and the chain rule for the entropy:

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

□

Proposition 5.4 (Information inequalities)**.** *For any two discrete RVs* X, Y*:*

(*a*) $I(X;Y) \geq 0$*, with equality iff* X, Y *are independent.*

(*b*) $I(X;Y) \leq H(X)$.

PROOF. From (30) in Proposition 5.2, $I(X;Y) = D(P_{XY} \| P_X P_Y)$, so (*a*) follows from the basic bound (*b*) in Theorem 2.5. Also, by definition, $I(X;Y) = H(X) - H(X|Y)$, so (*b*) follows from the nonnegativity of conditional entropy. □

The results of Propositions 5.2, 5.3 and 5.4 have natural generalisations for the case of conditional mutual information. These are stated in Exercise 12.25, but you may want to try to guess what they are before looking there. One of them that will be used in the proof of Proposition 5.5, is that $I(X;Z|Y) \geq 0$, with equality if and only if X and Z are conditionally independent given Y.

Notation: Conditional independence. As usual, we say that two discrete RVs X and Z are *conditionally independent given a third RV* Y, if, for all possible values (x, y, z),

$$\begin{aligned} P_{XZ|Y}(x, z|y) &= \mathbb{P}(X = x, Z = z | Y = y) \\ &= \mathbb{P}(X = x | Y = y) \mathbb{P}(Z = z | Y = y) \\ &= P_{X|Y}(x|y) P_{Z|Y}(z|y). \end{aligned}$$

Equivalently, we say the X, Y *and* Z *form a Markov chain*, which we sometimes denote as $X - Y - Z$. Note that, $X - Y - Z$ is equivalent to $Z - Y - X$, but not to $X - Z - Y$. Also note that the above condition can also equivalently be stated as,

$$P_{X|YZ}(x|y, z) = P_{X|Y}(x|y), \qquad \text{for all } x, y, z.$$

An important special case of conditional independence which we will encounter in several places is that, for any two RVs X, Y and any function F, we always have: $X - Y - F(Y)$.

Proposition 5.5 (Data processing inequalities for mutual information)**.** *If* X, Z *are conditionally independent given* Y*, i.e., if* $X - Y - Z$ *form a Markov chain, then:*

$$\begin{aligned} I(X;Z) &\leq I(X;Y) \\ \text{and} \quad I(X;Y|Z) &\leq I(X;Y). \end{aligned}$$

PROOF. Using the chain rule, we can expand $I(X;Y,Z)$ in two different ways to obtain:

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

Since $I(X;Z|Y) = 0$ by conditional independence, we have,

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

Both results of the proposition now follow from the nonnegativity of mutual information and conditional mutual information. □

Remark. Note that all the properties of mutual information established in this section remain valid in the case of RVs taking values in countably infinite alphabets, at least as long as all the entropies involved are finite.

5.2 Synergy and redundancy

Consider the following question that comes up in neuroscience. Suppose X is a visual stimulus, and Y_1, Y_2 are the responses to this stimulus produced by two neurons in your visual cortex. How does the pair (Y_1, Y_2) capture (some of) the information in X? Specifically, how does the mutual information $I(X;Y_1,Y_2)$ compare with the sum $I(X;Y_1) + I(X;Y_2)$ of the individual mutual informations? As it turns out (see Exercise 12.26), all three possibilities are possible: $I(X;Y_1,Y_2)$ can be strictly greater, strictly smaller, or equal to $I(X;Y_1) + I(X;Y_2)$.

Notice that we can use the chain rule for mutual information to express the difference,

$$S(X;Y_1,Y_2) := I(X;Y_1,Y_2) - I(X;Y_1) - I(X;Y_2) = I(X;Y_2|Y_1) - I(X;Y_2). \tag{31}$$

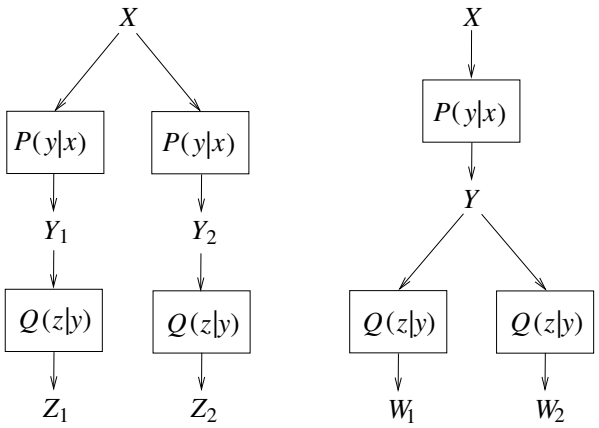
Asking whether the expression on the right-hand-side is positive or negative also admits a nice interpretation: Does knowing Y_1 make Y_2 more or less useful in determining X? In Exercise 12.26 you will show that $S(X;Y_1,Y_2)$ can in general be positive, negative, or zero. In other words, conditioning can decrease, increase, or not affect the mutual information. If $S(X;Y_1,Y_2)$ is negative, we say that Y_1, Y_2 contain *redundant* information about X; if it is zero, we say they are *orthogonal*; and if it is strictly positive, we say that the descriptions of X in Y_1 and Y_2 are *synergistic*: Knowing Y_1 already, makes the information in Y_2 even more valuable.

Example 5.6 (Are independent samples better?)**.** Suppose we are allowed to observe two equally noisy versions of some signal X, where the level of the noise is fixed. We are given a choice: We can either observe two independent noisy versions of X, or two correlated ones. Which is better?

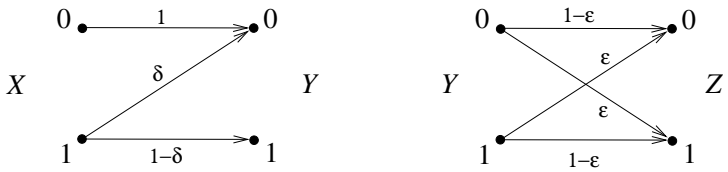
Consider the following two scenarios. Suppose $P_{Y|X}$ and $Q_{Z|Y}$ are two collections of conditional PMFs, with all variables taking values in the same alphabet A, and let X, the variable of interest, have PMF P_X on A. In the first scenario, corresponding to "independent observations," let Y_1, Y_2 be conditionally independent given X, each with distribution $P_{Y|X}(\cdot|X)$, and let Z_1 and Z_2 be distributed according to $Q_{Z|Y}(\cdot|Y_1)$ and $Q_{Z|Y}(\cdot|Y_2)$, respectively. In the second scenario, that of "correlated observations," let Y be distributed according to $P_{Y|X}(\cdot|X)$, and W_1, W_2 be conditionally independent given Y and distributed according to $Q_{Z|Y}(\cdot|Y)$. See the diagram below.

The joint distribution of (X, Z_1) is of course the same as the joint distribution of (X, W_1), so there is no difference in observing Z_1 or W_1. But the joint distribution of (X, Z_1, Z_2) is different from that of (X, W_1, W_2), and the amount of information we obtain from the second observation is different in these two scenarios. It is perhaps natural to expect that the two (conditionally) independent observations (Z_1, Z_2) would always be "better," in the sense that

the mutual information $I(X; Z_1, Z_2)$ will always be at least as large as $I(X; W_1, W_2)$. Somewhat surprisingly, it turns out that this is not always the case.



As a specific example, suppose $P_{Y|X}$ is the so-called Z-channel: If $X = 0$, then $Y = 0$ with probability 1, and if $X = 1$, then $Y \sim \text{Bern}(1-\delta)$, for some $\delta \in (0, 1)$. And take $Q_{Z|Y}$ to be a binary symmetric channel: Given Y, $Z = Y$ with probability $(1-\epsilon)$ and $Z = 1 - Y$ with probability ϵ, for some $\epsilon \in (0, 1)$:



If $X \sim \text{Bern}(1/2)$, $\delta = 0.85$ and $\epsilon = 0.1$, then a direct (tedious but instructive, so you may want to try it) calculation shows that, in this case,

$$I(X; W_1, W_2) \approx 0.047 > I(X; Z_1, Z_2) \approx 0.039,$$

so the correlated observations (W_1, W_2) are better than independent ones (Z_1, Z_2). In fact this is not a particularly rare phenomenon, as observed in [15].

Next we establish a sufficient condition for this counter-intuitive phenomenon to occur. A weaker condition, which is also necessary, is given in Exercise 12.29.

Proposition 5.7. *In the setting of Example 5.6, let $P_{Y|X}$ and $Q_{Z|Y}$ be arbitrary conditional PMFs, with all variables in the same discrete alphabet, and let X have an arbitrary distribution on A. If the synergy $S(X; W_1, W_2)$ is positive, then the correlated observations are better than the independent ones, in that $I(X; W_1, W_2) > I(X; Z_1, Z_2)$.*

Proof. In view of (31), having $S(X; W_1, W_2) > 0$ means that $I(X; W_2|W_1) > I(X; W_2)$, and since the joint distribution of $(X; W_2)$ is the same as that of $(X; Z_2)$, we have that,

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

Also, since Z_1 and Z_2 are conditionally independent given X, by the data processing property of mutual information we have, $I(X; Z_2|Z_1) \leq I(X; Z_2)$, therefore,

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

Adding $I(X; Z_1) = I(X; W_1)$ to both sides and using the chain rule completes proof. □

6 Entropy rate

In this brief section we define the *entropy rate* of a random process, or source, $\boldsymbol{X} = \{X_n\}$ as the average number of bits per symbol, in $\boldsymbol{X}$, and we derive some simple properties.

Definition 6.1 (Entropy rate)**.** *For an arbitrary sequence of discrete RVs* $\boldsymbol{X} = \{X_n\}$ *the* entropy rate *of* $\boldsymbol{X}$ *is defined as,*

$$H(\boldsymbol{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1^n) \quad \textit{bits/symbol,}$$

whenever the limit exists.

Example 6.2 (Memoryless source)**.** Suppose $\boldsymbol{X} = \{X_n\}$ is a sequence of IID RVs with common entropy $H = H(X_i)$. Then the entropy rate exists,

$$H(\boldsymbol{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1^n) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i) = H,$$

and $H(\boldsymbol{X}) = H$.

Example 6.3 (Ergodic Markov chain)**.** Suppose $\boldsymbol{X} = \{X_n \ ; \ n \geq 1\}$ is an irreducible and aperiodic Markov chain on a finite state space A, with transition matrix $Q = (Q_{xx'})$, where $Q_{xx'} = \mathbb{P}(X_{n+1} = x' | X_n = x)$, $x, x' \in A$. Suppose the chain has an arbitrary initial distribution $X_1 \sim P_{X_1}$, and unique stationary distribution π. Then, by the chain rule and conditional independence,

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

By the convergence theorem for Markov chains we know that $P_{X_n} \to \pi$ as $n \to \infty$, so the joint distribution of (X_{n-1}, X_n) converges as,

$$P_{X_{n-1}X_n}(x, x') = P_{X_n}(x) Q_{xx'} \to \pi(x) Q_{xx'}, \qquad \text{for all } x, x' \in A.$$

It is easy to see from the definition that conditional entropy $H(X|Y)$ is a continuous function of the joint distribution P_{XY} of the RVs X, Y, so, here, $H(X_n|X_{n-1}) \to H(\bar{X}_2|\bar{X}_1)$ as $n \to \infty$, where $\bar{X}_1 \sim \pi$ and $\mathbb{P}(\bar{X}_2 = x' | \bar{X}_1 = x) = Q_{xx'}$. Therefore,

$$\frac{1}{n} H(X_1^n) = \frac{H(X_1) - H(X_{n+1}|X_n)}{n} + \frac{1}{n} \sum_{i=1}^{n} H(X_{i+1}|X_i).$$

As $n \to \infty$, the first term above converges to zero (because the values of the numerator form a bounded sequence) and, since the summands in the second term converge to $H(\bar{X}_2|\bar{X}_1)$, their averages also converge to the same limit by the Cesàro lemma. So the entropy rate of $\boldsymbol{X}$ exists and $H(\boldsymbol{X}) = H(\bar{X}_2|\bar{X}_1)$, regardless of the chain's initial distribution.

One of the most general classes of sources for which the entropy rate exists is the class of stationary sources.

Definition 6.4 (Stationary source)**.** *A source* $\boldsymbol{X} = \{X_n\}$ *is* stationary *if, for each* $n \geq 1$*, the distribution of* X_{k+1}^{k+n} *is the same for all* k*.*

Theorem 6.5 (Entropy rate of stationary source)**.** *If* $\boldsymbol{X} = \{X_n\ ;\ n \geq 1\}$ *is a stationary source on a finite alphabet* A*, then the entropy rate,*

$$H(\boldsymbol{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1^n),$$

exists and it is equal to,

$$H(\boldsymbol{X}) = \lim_{n\to\infty} H(X_n|X_1^{n-1}).$$

Proof. Consider the sequence $\{H(X_n|X_1^{n-1})\}$. It is bounded below by zero, and it is also nonincreasing,

$$H(X_{n+1}|X_1^n) \leq H(X_{n+1}|X_2^n) = H(X_n|X_1^{n-1}),$$

where we used the fact that conditioning reduces entropy and stationarity. Hence, the limit $\lim_n H(X_n|X_1^{n-1})$ exists. By the chain rule,

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

and since the summands converge, so do their averages by the Cesàro lemma. Therefore $H(X_1^n)/n$ and $H(X_n|X_1^{n-1})$ both converge to the same limit, $H(\boldsymbol{X})$, as $n \to \infty$. □

We close this section with yet another representation of the entropy rate for a stationary source $\boldsymbol{X} = \{X_n\ ;\ n \in \mathbb{Z}\}$. Write $P(\cdot|x_{-n}^{-1})$ for the conditional distribution of X_0 given $X_{-n}^{-1} = x_{-n}^{-1}$, and similarly let $P(\cdot|x_{-\infty}^{-1})$ denote its conditional distribution given the infinite past $X_{-\infty}^{-1} = x_{-\infty}^{-1}$. Lévy's 'upward' martingale convergence theorem [21] tells us that, for any $x \in A$,

$$P(x|X_{-n}^{-1}) \to P(x|X_{-\infty}^{-1}) \quad \text{a.s., as } n \to \infty,$$

where 'a.s.' stands for 'almost surely.' And since the function $p \log p$ is continuous and bounded for $p \in [0, 1]$, by the bounded convergence theorem we have that, as $n \to \infty$,

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

Noting that, by stationarity, $H(X_0|X_{-n}^{-1}) = H(X_{n+1}|X_1^n)$, and using the natural notation $H(X_0|X_{-\infty}^{-1})$ for $\mathbb{E}[-\log P(X_0|X_{-\infty}^{-1})]$, we have established the following:

Theorem 6.6 (* Entropy rate and the infinite past)**.** *If* $\boldsymbol{X} = \{X_n\ ;\ n \in \mathbb{Z}\}$ *is a stationary source on a finite alphabet* A*, then its entropy rate* $H(\boldsymbol{X})$ *can also be expressed as:*

$$H(\boldsymbol{X}) = \lim_{n\to\infty} H(X_0|X_{-n}^{-1}) = H(X_0|X_{-\infty}^{-1}).$$

7 * The Shannon-McMillan-Breiman theorem

In this section we present the general version of Theorem 1.1 and the Asymptotic Equipartition Property of Section 1.1, for discrete stationary sources.

For the statement of the Shannon-McMillan-Theorem we need to assume that the source is *ergodic*, a property which we will define precisely after the theorem. Suffice it for now to say that, intuitively, ergodicity means that the source satisfies the general form of the strong law of large numbers (SLLN).

Theorem 7.1 (Shannon-McMillan-Breiman theorem)**.** *Suppose $\boldsymbol{X} = \{X_n\ ;\ n \geq 1\}$ is a stationary and ergodic source on a finite alphabet, with entropy rate $H = H(\boldsymbol{X})$, and for each $n \geq 1$ write P_n for the PMF of X_1^n. Then, as $n \to \infty$:*

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

We find it convenient to work with a two-sided process $\{X_n\ ;\ n \in \mathbb{Z}\}$. In fact we can assume without loss of generality that our source $\boldsymbol{X} = \{X_n\}$ is defined for all $n \in \mathbb{Z}$, since Kolmogorov's extension theorem [5] guarantees that a one-sided stationary process $\boldsymbol{X} = \{X_n\ ;\ n \geq 1\}$ always admits a unique two-sided extension.

Definition 7.2 (Ergodicity)**.** *Consider a stationary source $\boldsymbol{X} = \{X_n\ ;\ n \in \mathbb{Z}\}$ on a finite alphabet A, and let the shift operator $T : A^{\mathbb{Z}} \to A^{\mathbb{Z}}$ on the set of doubly-infinite sequences $\boldsymbol{x} = x_{-\infty}^{\infty} \in A^{\mathbb{Z}}$ be defined by $(T\boldsymbol{x})_n = x_{n+1}$, for all $n \in \mathbb{Z}$.*

The source $\boldsymbol{X}$ is ergodic *if all T-invariant (measurable) sets are trivial, i.e., if $T^{-1}B = B$ always implies that $\mathbb{P}(X_{-\infty}^{\infty} \in B) = 0$ or 1.*

The importance of ergodicity for us is that it guarantees that the following strong version of the SLLN holds:

Theorem 7.3 (Birkhoff ergodic theorem)**.** *Suppose $\boldsymbol{X} = \{X_n\ ;\ n \in \mathbb{Z}\}$ is a stationary and ergodic source. Then for any function $f : A^{\mathbb{Z}} \to \mathbb{R}$ such that $\mathbb{E}|f(X_{-\infty}^{\infty})| < \infty$ we have,*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} f(T^i X_{-\infty}^{\infty}) \to \mathbb{E}[f(X_{-\infty}^{\infty})], \quad a.s.$$

For example, recalling that every IID source is ergodic and taking $f(x_{-\infty}^{\infty}) = x_1$ in Theorem 7.3, we recover the classical SLLN.

The obvious first approach to proving Theorem 7.1 is to use the expansion,

$$P_n(X_1^n) = \prod_{i=1}^{n} P(X_i|X_1^{i-1}),$$

and try to prove the convergence of,

$$-\frac{1}{n}\log P_n(X_1^n) = \frac{1}{n}\sum_{i=1}^{n}[-\log P(X_i|X_1^{i-1})], \tag{32}$$

where, for the sake of avoiding unnecessarily cumbersome notation, we omit the subscripts in the conditional PMFs like $P(x_i|x_1^{i-1}) = P_{X_i|X_1^{i-1}}(x_i|x_1^{i-1})$ for the rest of this section. But the ergodic theorem cannot be directly applied here, since the summands $-\log P(x_i|x_1^{i-1})$ cannot be expressed in the form $f(T^i x_{-\infty}^{\infty})$.

So, instead, we consider two related expressions. The first is the same as (32) but conditioned on the infinite past. By the ergodic theorem,

$$-\frac{1}{n}\log P(X_1^n|X_{-\infty}^0) = \frac{1}{n}\sum_{i=1}^{n}[-\log P(X_i|X_{-\infty}^{i-1})] \to \mathbb{E}[-\log P(X_0|X_{-\infty}^{-1})] = H, \quad \text{a.s.}, \tag{33}$$

as $n \to \infty$, where we recall the representation of the entropy rate H in Theorem 6.6. For the second one, rather than conditioning on a past sequence X_1^{i-1} of increasing length, we condition on the most recent k symbols. Again by the ergodic theorem, for each fixed $k \geq 1$, we have,

$$\frac{1}{n}\sum_{i=1}^{n}[-\log P(X_i|X_{i-k}^{i-1})] \to \mathbb{E}[-\log P(X_0|X_{-k}^{-1})] = H(X_0|X_{-k}^{-1}), \quad \text{a.s., as } n \to \infty. \tag{34}$$

It is instructive to compare the three different expressions in the left-hand sides of (32), (33) and (34).

The main idea of the proof is to show that the left-hand side (32) can be asymptotically "sandwiched" between (33) and (34), and then recall from Theorem 6.6 that $H(X_0|X_{-k}^{-1}) \to H(X_0|X_{-\infty}^{-1}) = H$ as $k \to \infty$. This proof strategy is due to Algoet and Cover [1]; see [19] for more background material on ergodicity and other related ergodic properties.

Proof of Theorem 7.1. First we claim that,

$$\limsup_{n\to\infty}\left[-\frac{1}{n}\log P(X_1^n|X_{-\infty}^0)\right] - \left[-\frac{1}{n}\log P_n(X_1^n)\right] \leq 0, \quad \text{a.s.} \tag{35}$$

Let $\epsilon > 0$ arbitrary. We have,

$$\begin{aligned}\mathbb{P}\left\{-\frac{1}{n}\log P(X_1^n|X_{-\infty}^0) - \left[-\frac{1}{n}\log P_n(X_1^n)\right] > \epsilon\right\} &= \mathbb{P}\left\{\frac{P(X_1^n)}{P(X_1^n|X_{-\infty}^0)} > 2^{n\epsilon}\right\} \\ &\leq 2^{-n\epsilon}\mathbb{E}\left[\frac{P(X_1^n)}{P(X_1^n|X_{-\infty}^0)}\right],\end{aligned} \tag{36}$$

where we used Markov's inequality. Now, for any $x_{-\infty}^0$, write $S(x_{-\infty}^0) \subset A^n$ for the set of all x_1^n such that $P(x_1^n|x_{-\infty}^0) \neq 0$. Then the conditional expectation,

$$\mathbb{E}\left\{\frac{P(X_1^n)}{P(X_1^n|X_{-\infty}^0)}\middle| X_{-\infty}^0 = x_{-\infty}^0\right\} = \sum_{x_1^n \in S(x_{-\infty}^0)} P(x_1^n|x_{-\infty}^0)\frac{P(x_1^n)}{P(x_1^n|x_{-\infty}^0)} \leq 1,$$

and hence also,

$$\mathbb{E}\left[\frac{P(X_1^n)}{P(X_1^n|X_{-\infty}^0)}\right] = \mathbb{E}\left[\mathbb{E}\left\{\frac{P(X_1^n|X_{-\infty}^0)}{P(X_1^n)}\middle| X_{-\infty}^0\right\}\right] \leq 1.$$

Therefore, the probability in (36) is bounded above by $2^{-n\epsilon}$, which is summable over $n \geq 1$, and since $\epsilon > 0$ was arbitrary the Borel-Cantelli lemma gives (35).

For each fixed $k \geq 1$, we define a family of distributions that can be thought of as a 'kth order Markov approximation' to the distribution of $\boldsymbol{X}$. Let $Q_{n,k} = P_n$ for $1 \leq n \leq k$, and for $n > k$ let,

$$Q_{n,k}(x_1^n) = P_k(x_1^k)\prod_{i=k+1}^{n} P(x_i|x_{i-k}^{i-1}), \quad x_1^n \in A^n. \tag{37}$$

Our second claim is that,

$$\limsup_{n\to\infty}\left[-\frac{1}{n}\log P_n(X_1^n)\right] - \left[-\frac{1}{n}\log Q_{n,k}(X_1^n)\right] \leq 0, \quad \text{a.s.} \tag{38}$$

Again with $\epsilon > 0$ arbitrary, using Markov's inequality we have,

$$\begin{aligned}
\mathbb{P}\left\{-\frac{1}{n}\log P_n(X_1^n) + \frac{1}{n}\log Q_{n,k}(X_1^n) > \epsilon\right\} &= \mathbb{P}\left\{\frac{Q_{n,k}(X_1^n)}{P_n(X_1^n)} > 2^{n\epsilon}\right\} \\
&\leq 2^{-n\epsilon}\mathbb{E}\left[\frac{Q_{n,k}(X_1^n)}{P_n(X_1^n)}\right] \\
&= 2^{-n\epsilon}\sum_{x_1^n : P_n(x_1^n)\neq 0} P_n(x_1^n)\frac{Q_{n,k}(x_1^n)}{P_n(x_1^n)} \\
&\leq 2^{-n\epsilon},
\end{aligned}$$

and since $\epsilon > 0$ was arbitrary the Borel-Cantelli lemma gives (38). In fact, noting the similarity between the terms in the left-hand side of (34) and the definition of $Q_{n,k}$ in (37), we see that the difference,

$$\frac{1}{n}\sum_{i=1}^{n}[-\log P(X_i|X_{i-k}^{i-1})] - \left[-\frac{1}{n}\log Q_{n,k}(X_1^n)\right] = \frac{1}{n}\sum_{i=1}^{k}\log\frac{P(X_i|X_1^{i-1})}{P(X_i|X_{i-k}^{i-1})} \to 0, \quad \text{a.s.}$$

as $n \to \infty$, since the sum does not depend on n and is a.s. finite. Combining this with (38) yields,

$$\limsup_{n\to\infty}\left[-\frac{1}{n}\log P_n(X_1^n)\right] - \frac{1}{n}\sum_{i=1}^{n}[-\log P(X_i|X_{i-k}^{i-1})] \leq 0, \quad \text{a.s.} \tag{39}$$

We are now in a position to complete the proof of the theorem. We have shown that, for any $k \geq 1$,

$$H \leq \liminf_{n\to\infty} -\frac{1}{n}\log P_n(X_1^n) \leq \limsup_{n\to\infty} -\frac{1}{n}\log P_n(X_1^n) \leq H(X_0|X_{-k}^{-1}), \quad \text{a.s.},$$

where the first inequality follows from combining the first claim (35) with the result of the ergodic theorem in (33), and the third inequality follows from the result of the second claim (39) combined with (34). Taking $k \to \infty$ and recalling Theorem 6.6 yields the desired result. □

It is worth noting that the statements and proofs of the AEP of Corollary 1.3 and the fixed-rate coding theorem in Theorem 1.6 remain valid exactly as given in the IID case, for an arbitrary stationary and ergodic source $\boldsymbol{X}$ with entropy rate H instead of a memoryless source.

8 Types and large deviations

In this section we will develop a series of combinatorial identities and bounds, that express probabilities related to IID strings in terms of the entropy and relative entropy. These tools then will be used to derive two important results for probabilities of large deviations, and to compute the best achievable exponent of the probability of error in fixed-rate data compression.

8.1 The method of types

Given a finite string $x_1^n \in A^n$ taking values in a finite alphabet A, we refer to the empirical distribution $\hat{P}_n$ it induces on A as its *type*. Recall (16): Each $\hat{P}_n(a)$ is the frequency with which a appears in x_1^n. Sometimes we will also write $\hat{P}_n$ for the random type induced by the RVs X_1^n, but the difference should always be clear from the relevant context.

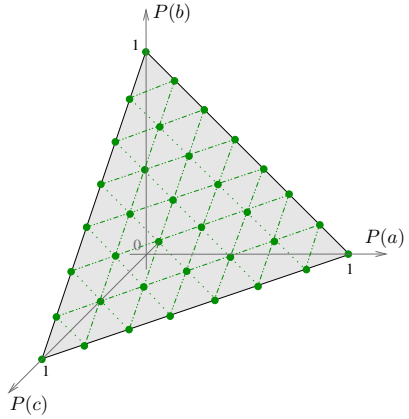
For fixed alphabet A of size $|A| = m$, the type $\hat{P}_n$ of any string $x_1^n \in A^n$ is a PMF on A, so it belongs to the probability simplex,

$$\mathcal{P} = \Big\{ P \in [0,1]^m \ : \ \sum_{a \in A} P(a) = 1 \Big\}.$$

In fact, $\hat{P}_n$ necessarily belongs to a finite subset $\mathcal{P}_n$ of $\mathcal{P}$ consisting of all *n-types*:

$$\mathcal{P}_n = \big\{ P \in \mathcal{P} \ : \ nP(a) \text{ is an integer, for all } a \in A \big\}.$$

The figure on the right shows the probability simplex $\mathcal{P}$ as a shaded triangle, in the case of a 3-letter alphabet $A = \{a, b, c\}$. The n-types in $\mathcal{P}_n$ for $n = 7$ are shown as bold green dots. Note that, in general, for an alphabet A of size $|A| = m$, there are only polynomially many n types in $\mathcal{P}_n$: Since for $P \in \mathcal{P}_n$ and any $a \in A$ we have $P(a) = k/n$ for some $k \in \{0, 1, \ldots, n\}$, in fact,



$$|\mathcal{P}_n| \leq (n+1)^m. \tag{40}$$

Of course this is a crude overestimate, but it will be sufficient for most of the discussion in this section.

Our first result shows that the probability of an IID string x_1^n can always be expressed in terms of its type.

Proposition 8.1. (*a*) *For any PMF Q on A, the probability under Q^n of a string x_1^n with type $\hat{P}_n$ is:*

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

(*b*) *If $\hat{P}_n = Q$:*

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

Proof. For part (*a*) we essentially repeat the computation for the Neyman-Pearson region in (18): First we re-write,

$$\log Q^n(x_1^n) = \log \prod_{i=1}^n Q(x_i) = \sum_{i=1}^n \log Q(x_i) = \sum_{i=1}^n \sum_{a \in A} \mathbb{I}_{\{a\}}(x_i) \log Q(a),$$

and using the definition of $\hat{P}_n$,

$$\log Q^n(x_1^n) = n \sum_{a \in A} \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{\{a\}}(x_i) \log \Big(Q(a) \frac{\hat{P}_n(a)}{\hat{P}_n(a)} \Big) = -n \Big[H(\hat{P}_n) + D(\hat{P}_n \| Q) \Big].$$

Part (*b*) is immediate from (*a*) with $Q = \hat{P}_n$. □

A rather obvious but important observation here is that all strings of the same type have the same probability under an IID distribution. This motivates as to define:

Definition 8.2 (Type class)**.** *For a type $P \in \mathcal{P}_n$, the* type class $T(P)$ *consists of all strings x_1^n with type P:*

$$T(P) = \left\{ x_1^n \in A^n \ : \ \hat{P}_n = P \right\}.$$

Next we derive simple upper and lower bounds on the size and probability of $T(P)$. In fact it is easy to write out an exact expression for $|T(P)|$. Let $A = \{a_1, a_2, \ldots, a_m\}$. Since $T(P)$ consists of all possible arrangements of $nP(a_1)$ a_1's, of $nP(a_2)$ a_2's, and so on, $|T(P)|$ is given by the multinomial coefficient:

$$|T(P)| = \frac{n!}{\prod_{j=1}^m (nP(a_j))!}. \tag{41}$$

Proposition 8.3 (Size of type class)**.** *For any type $P \in \mathcal{P}_n$ over an alphabet of size m:*

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

For the proof of the proposition we will need the following natural bound:

Lemma 8.4. *The most likely type class under P^n is $T(P)$:*

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

Proof. Let $P, P' \in \mathcal{P}_n$ and write $A = \{a_1, a_2, \ldots, a_m\}$. In view of (41),

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

Using the simple inequality $k!/\ell! \geq \ell^{k-\ell}$ (to see this, consider $k \geq \ell$ and $k < \ell$ separately),

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

therefore,

$$\frac{P^n(T(P))}{P^n(T(P'))} \geq n^{n \sum_{j=1}^m [P'(a_j) - P(a_j)]} = 1,$$

as desired. □

Proof of Proposition 8.3. The proof of the upper bound is similar to the proof of the upper bound for $|B_n^*|$ in the direct part of the AEP. Using Proposition 8.1,

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

For the lower bound we have,

$$1 = \sum_{x_1^n \in A^n} P^n(x_1^n) = \sum_{P' \in \mathcal{P}_n} P^n(T(P')) \leq |\mathcal{P}_n| \max_{P' \in \mathcal{P}_n} P^n(T(P')) = |\mathcal{P}_n| P^n(T(P)),$$

where in the last step we used Lemma 8.4. Therefore, using the observation (40) and Proposition 8.1 again,

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

which completes the proof. □

The last two results above, Proposition 8.1 and Proposition 8.3, can be viewed as giving a more precise, finite-n version of the AEP: They tell us that the set $T(P)$ of all strings with type P all have the same probability, $2^{-nH(P)}$, and their number $|T(P)|$ is approximately $2^{nH(P)}$.

Finally we establish corresponding bounds for the probability of a type class.

Proposition 8.5 (Probability of type class)**.** *For any type $P \in \mathcal{P}_n$ over an alphabet of size m, and any PMF Q on A:*

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

Proof. Let x_1^n be any string in $T(P)$. Then, by Proposition 8.1,

$$Q^n(T(P)) = |T(P)| Q^n(x_1^n) = |T(P)| 2^{-n[H(P)+D(P\|Q)]}.$$

The result follows from this combined with the bounds of Proposition 8.3. □

8.2 Sanov's theorem

Here we will use the method of types to establish one of the central results in large deviations, Sanov's theorem. *Large deviations* is a sub-area of probability theory, concerned with estimating the probabilities of rare events. The following is a standard example.

Example 8.6 (Empirical averages and types)**.** Suppose X_1^n are IID with common PMF Q on a finite alphabet A, and let $f : A \to \mathbb{R}$ be a function with mean $\mu = \mathbb{E}[f(X_1)]$. The law of large numbers tells us that the probability that the empirical averages $S_n = (1/n)\sum_{i=1}^n f(X_i)$ of the $f(X_i)$ significantly exceed μ goes to zero, as $n \to \infty$. We wish to estimate exactly how small this probability is as a function of n.

Let B denote the event,

$$B = \{S_n \geq \mu + \epsilon\}, \tag{42}$$

where $\epsilon > 0$ and $\mu + \epsilon < f^* := \max_{a \in A} f(a)$, so that the probability of B is nonzero. A standard way to bound the probability of B above is as follows: Switching for a moment to natural logarithms and exponentials, for any $\lambda > 0$, by Markov's inequality,

$$\mathbb{P}(S_n \geq \mu + \epsilon) = \mathbb{P}(e^{\lambda n S_n} \geq e^{n\lambda(\mu+\epsilon)}) \leq e^{-n\lambda(\mu+\epsilon)} \mathbb{E}(e^{n\lambda S_n}).$$

Let,

$$\Lambda(\lambda) := \log_e \mathbb{E}(e^{\lambda f(X_1)}), \qquad \lambda \in \mathbb{R},$$

denote the log-moment generating function of the $f(X_i)$. Then, by independence,

$$\mathbb{E}(e^{n\lambda S_n}) = \mathbb{E}\Big[e^{\lambda \sum_{i=1}^n f(X_i)}\Big] = e^{n\Lambda(\lambda)},$$

and hence,

$$\mathbb{P}(S_n \geq \mu + \epsilon) \leq e^{-n\lambda(\mu+\epsilon)} e^{n\Lambda(\lambda)}.$$

Since this holds for all $\lambda > 0$, optimizing over λ we obtain the **Chernoff bound**:

$$\mathbb{P}(S_n \geq \mu + \epsilon) \leq e^{-n\Lambda^*(\mu+\epsilon)},$$

where,

$$\Lambda^*(x) := \sup_{\lambda > 0} [\lambda x - \Lambda(\lambda)], \qquad \mu < x < f^*.$$

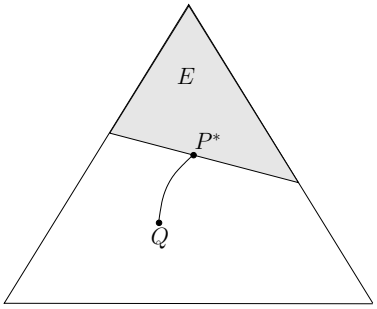
Alternatively, the empirical averages S_n can be expressed in terms of the type $\hat{P}_n$ of X_1^n as,

$$S_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i) = \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in A}\mathbb{I}_{\{a\}}(X_i)f(a) = \sum_{a\in A}\hat{P}_n(a)f(a) = \mathbb{E}_{X\sim\hat{P}_n}[f(X)].$$

So B can be expressed as, $B = \{\hat{P}_n \in E\}$, where $E \subset \mathcal{P}$ is,

$$E = \{P \in \mathcal{P} \,:\, \mathbb{E}_{X\sim P}[f(X)] \geq \mu + \epsilon\}. \tag{43}$$

In view of the above example, in order to understand the probabilities of events of "large exceedances," it suffices to look at events of types, of the form $\{\hat{P}_n \in E\}$, for appropriate sets $E \subset \mathcal{P}$. For example, the figure on the right shows the event E defined as the linear half-space in (43).



Sanov's theorem states that, for nice sets E, the probability $Q^n(\hat{P}_n \in E)$ is approximately equal to $2^{-nD(P^*\|Q)}$ for large n, where P^* minimises $D(P\|Q)$ over all $P \in E$. An attractive interpretation of this is as saying that, the probability of the rare event $\{\hat{P}_n \in E\}$ is exponentially small, and the exponent is determined by the least unlikely way in which this event can occur.

In order to avoid uninteresting technicalities, throughout this section we will assume that Q has full support, $Q(a) > 0$ for all $a \in A$.

Theorem 8.7 (Sanov's theorem)**.** *Let $\hat{P}_n$ denote the empirical distribution of the IID RVs X_1^n with PMF Q which has full support on an alphabet A of size m. For any $E \subset \mathcal{P}$:*

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

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

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

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

Proof. Since Q has full support, for any P we have $D(P\|Q) \leq -\sum_{a\in A}\log Q(a)$, so $\inf_{P\in E} D(P\|Q)$ is always finite. And since $D(P\|Q)$ is convex in P for a fixed Q, if E is also closed then the the infimum is achieved by some $P^* \in E$. In general, since $\mathcal{P}$ is compact, there is again always a minimising P^*, although it may not belong to E. We choose and fix such a P^* throughout the proof.

For the upper bound, we have,

$$Q^n(\hat{P}_n \in E) = Q^n(\hat{P}_n \in E \cap \mathcal{P}_n) = \sum_{P\in E\cap\mathcal{P}_n} Q^n(\hat{P}_n = P) = \sum_{P\in E\cap\mathcal{P}_n} Q^n(T(P)), \tag{44}$$

and using Proposition 8.5,

$$Q^n(\hat{P}_n \in E) \leq |E\cap\mathcal{P}_n| \max_{P\in E\cap\mathcal{P}_n} Q^n(T(P)) \leq (n+1)^m \sup_{P\in E} 2^{-nD(P\|Q)},$$

where we also used the bound on $|\mathcal{P}_n|$ from (40). Therefore, $Q^n(\hat{P}_n \in E) \leq (n+1)^m 2^{-nD(P^*\|Q)}$, as claimed, and, in particular,

$$\liminf_{n\to\infty} -\frac{1}{n}\log Q^n(\hat{P}_n \in E) \geq D(P^*\|Q). \tag{45}$$

For the lower bound, we observe that the set $\cup_{n\geq 1}\mathcal{P}_n$ of n-types over all $n \geq 1$ is dense in $\mathcal{P}$ and hence also in E; to see this, just look at the figure in the beginning of this section and imagine n growing. In fact, the $\mathcal{P}_n$ eventually intersect with every open subset of $\mathcal{P}$. This means that we can find a sequence $\{P_n\}$ with each $P_n \in \mathcal{P}_n \cap E$, such that $P_n \to P^*$ and $D(P_n\|Q) \to D(P^*\|Q)$ as $n \to \infty$. From (44), using Proposition 8.5,

$$Q^n(\hat{P}_n \in E) \geq Q^n(T(P_n)) \geq (n+1)^{-m}2^{-nD(P_n\|Q)},$$

so,

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

Combining this with (45) completes the proof. □

Example 8.6 continued (Empirical averages and types) Consider again the event $B = \{S_n \geq \mu + \epsilon\}$, for which the Chernoff bound tells us that,

$$Q^n(B) \leq e^{-n\Lambda^*(\mu+\epsilon)}.$$

With E as in (43), we see that E is equal to the closure of its interior (because $\mu < \mu+\epsilon < f^*$), so from Sanov's theorem we have,

$$Q^n(B) \leq (n+1)^m e^{-nD_e(P^*\|Q)},$$

where $D_e(P^*\|Q) = \inf_{P\in E} D_e(P\|Q)$. Moreover, Sanov's theorem tells us that $D_e(P^*\|Q)$ is asymptotically the right exponent, so we must have $\Lambda^*(\mu+\epsilon) \leq D_e(P^*\|Q)$. As it turns out, these two exponents are in fact equal.

Proposition 8.8. *Let X_1^n be IID RVs with common PMF Q on A. For events B of the form (42),*

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

the exponents in Chernoff's bound and in Sanov's theorem are equal:

$$\Lambda^*(\mu+\epsilon) = D_e(P^*\|Q).$$

Proof. We already showed that,

$$\Lambda^*(\mu+\epsilon) \leq D_e(P^*\|Q). \tag{46}$$

(This, incidentally, can also be shown directly using the result of Exercise 12.40.) To prove that $\Lambda^*(\mu+\epsilon) \geq D_e(P^*\|Q)$ as well, we first compute the supremum in the definition of $\Lambda^*(x)$. For each $\lambda > 0$, define the PMF,

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

Then,

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

and a similar computation (do it!) shows that,

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

Therefore, $\Lambda'(\lambda)$ is nondecreasing from $\Lambda'(0+) = \mu$ to $\lim_{\lambda\to\infty}\Lambda'(\lambda) = f^*$, so there must exist a $\lambda^* > 0$ such that $\Lambda'(\lambda^*) = \mu + \epsilon$. This λ^* must necessarily achieve the supremum in the definition of $\Lambda^*(\mu+\epsilon)$,

$$\Lambda^*(\mu+\epsilon) = \lambda^*(\mu+\epsilon) - \Lambda(\lambda^*).$$

Then $\mathbb{E}_{Y\sim P_{\lambda^*}}[f(Y)] = \Lambda'(\lambda^*) = \mu+\epsilon$, so $P_{\lambda^*} \in E$, and hence $D_e(P_{\lambda^*}\|Q) \geq D_e(P^*\|Q)$, i.e.,

$$\begin{aligned} D_e(P^*\|Q) &\leq D_e(P_{\lambda^*}\|Q) && (47)\\ &= \mathbb{E}_{Y\sim P_{\lambda^*}}\Big[\log\frac{P_{\lambda^*}(Y)}{Q(Y)}\Big] \\ &= \mathbb{E}_{Y\sim P_{\lambda^*}}\Big[\log\frac{e^{\lambda^* f(Y)}}{\mathbb{E}[e^{\lambda^* f(X_1)}]}\Big] \\ &= \lambda^*\mathbb{E}_{Y\sim P_{\lambda^*}}[f(Y)] - \Lambda(\lambda^*) \\ &= \Lambda^*(\mu+\epsilon), && (48)\end{aligned}$$

as required. □

Looking closely at the above proof, we see that we have actually identified the minimising P^* explicitly:

Corollary 8.9. *Let X_1^n be IID RVs with common PMF Q on A. For events B as in Proposition 8.8, the minimising P^* in Sanov's theorem is unique and given by,*

$$P^*(a) = \frac{e^{\lambda^* f(a)}}{\mathbb{E}[e^{\lambda^* f(X_1)}]}Q(a), \qquad a \in A,$$

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

Proof. Since $D(P\|Q)$ is strictly convex in P for fixed Q (Exercise 12.8), and E is nonempty, convex and closed, the minimising P^* is unique. From equations (46), (47) and (48) we see that $D_e(P^*\|Q) = D_e(P_{\lambda^*}\|Q)$, where P_{λ^*} is as stated. □

The following L^2-style bound is an interesting result about the minimiser P^* in Sanov's theorem, viewed as the "orthogonal projection" of Q onto E with respect to relative entropy. Of course this is only a heuristic interpretation since $D(P\|Q)$ is not the metric of a Hilbert space, but the result itself is both intuitively satisfying and useful.

Theorem 8.10 (Pythagorean identity for relative entropy)**.** *Suppose the set $E \subset \mathcal{P}$ is closed and convex, take $Q \notin E$, assume Q has full support, and let P^* achieve the minimum in Sanov's theorem,*

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

Then:

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

Before giving the proof we note that, by the strict convexity of $D(P\|Q)$ in P (Exercise 12.8) and the fact that E is convex and closed, the minimiser P^* exists and is unique. One of the uses of the Pythagorean identity is to provide a quantitative version of this statement: Not only $D(P\|Q) = D(P^*\|Q)$ implies $P = P^*$, but also, if $D(P\|Q)$ is close to $D(P^*\|Q)$, say $D(P\|Q) \leq D(P^*\|Q) + \delta$, then P is necessarily close to P^*, in that $D(P\|P^*) \leq \delta$.

Proof. For simplicity, we give the proof for D_e instead of D. The stated result follows from the one we prove simply by multiplying both sides by $(\log e)$.

Let $P \in E$ arbitrary, and consider the mixture $P_\lambda = \lambda P + (1-\lambda)P^*$, $0 \le \lambda \le 1$. By convexity, $P_\lambda \in E$ for all λ, and by the definition of P^*, $D(P_\lambda \| Q) \ge D(P^* \| Q) = D(P_0 \| Q)$. Therefore, the derivative of $D(P_\lambda \| Q)$, if it exists, must be nonnegative at $\lambda = 0+$. We have,

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

which proves the theorem. □

8.3 The Gibbs conditioning principle

Suppose you roll a fair die 1000 times and, instead of getting an average of 3.5, you get something much larger, say 5 or more. Clearly the empirical distribution of the results could not have been close to uniform. So, given that you got an average ≥ 5, what proportion of the time did you get a 6?

The Gibbs conditioning principle states that, conditional on the rare event $\{\hat{P}_n \in E\}$, the type $\hat{P}_n$ is close to the minimiser P^* in Sanov's theorem, with high probability:

Theorem 8.11 (Gibbs' conditioning principle)**.** *Let $\hat{P}_n$ denote the empirical distribution of the IID RVs X_1^n with PMF Q which has full support on A. Suppose $E \subset \mathcal{P}$ is closed, convex, has nonempty interior, and does not contain Q. Then, as $n \to \infty$,*

$$\mathbb{E}(\hat{P}_n(a) | \hat{P}_n \in E) = \mathbb{P}(X_1 = a | \hat{P}_n \in E) \to P^*(a),$$

for each $a \in A$, where $P^* \in E$ uniquely achieves,

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

Proof. Recall that, as noted earlier, the minimiser P^* here exists and is unique. Also, by symmetry, the conditional distribution of each X_i given $\{\hat{P}_n \in E\}$ is the same, so,

$$\mathbb{E}(\hat{P}_n(a) | \hat{P}_n \in E) = \frac{1}{n} \sum_{i=1}^n \mathbb{P}(X_i = a | \hat{P}_n \in E) = \mathbb{P}(X_1 = a | \hat{P}_n \in E).$$

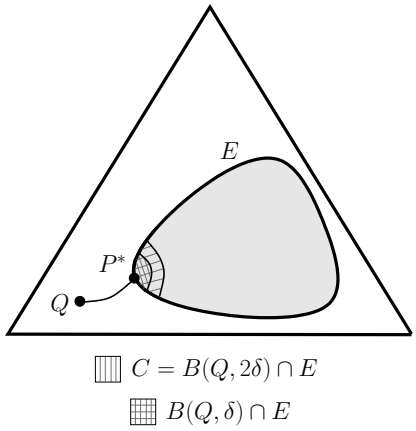
Consider the relative entropy neighborhoods,

$$B(Q, \delta) := \{P \in \mathcal{P} \;:\; D(P\|Q) \le D(P^*\|Q) + \delta\},$$

and define the subsets of $\mathcal{P}$,

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

Our strategy is to show that $\hat{P}_n$ is close to P^* with high probability, in that, for large n, we have $Q^n(\hat{P}_n \in C | \hat{P}_n \in E) \approx 1$, or, equivalently, that $Q^n(\hat{P}_n \in D | \hat{P}_n \in E) \approx 0$.



We begin by noting that,

$$\begin{aligned} Q^n(\hat{P}_n \in D | \hat{P}_n \in E) &= \frac{Q^n(\hat{P}_n \in D \cap E)}{Q^n(\hat{P}_n \in E)} \\ &\leq \frac{Q^n(\hat{P}_n \in D)}{Q^n(\hat{P}_n \in C)}. \end{aligned} \tag{49}$$

For the numerator we observe that, if $P \in D$ then $D(P\|Q) > D(P^*\|Q) + 2\delta$, so by the upper bound of Sanov's theorem applied to the event D,

$$\begin{aligned} Q^n(\hat{P}_n \in D) &\leq (n+1)^m 2^{-n \inf_{P \in D} D(P\|Q)} \\ &\leq (n+1)^m 2^{-n[D(P^*\|Q)+2\delta]}. \end{aligned} \tag{50}$$

For the denominator in (49), first note that, since $\cup_{n \geq 1} \mathcal{P}_n$ is dense in $\mathcal{P}$, the $\mathcal{P}_n$ eventually intersect every open set in $\mathcal{P}$, and E has nonempty interior, then $B(Q, \delta) \cap E \cap \mathcal{P}_n$ is nonempty eventually. Using Proposition 8.5 with $P_n \in B(Q, \delta) \cap E \cap \mathcal{P}_n$,

$$Q^n(\hat{P}_n \in C) \geq Q^n(\hat{P}_n \in B(Q, \delta) \cap E) \geq Q^n(T(P_n)) \geq (n+1)^{-m} 2^{-nD(P_n\|Q)},$$

and since $P_n \in B(Q, \delta)$,

$$Q^n(\hat{P}_n \in C) \geq (n+1)^{-m} 2^{-n[D(P^*\|Q)+\delta]}. \tag{51}$$

Substituting (50) and (51) in (49),

$$Q^n(\hat{P}_n \in D | \hat{P}_n \in E) \leq (n+1)^{2m} 2^{-n\delta}, \tag{52}$$

which of course $\to 0$ as $n \to \infty$. Therefore, conditional on E, $\hat{P}_n$ will almost certainly be close to P^* in the sense that $D(\hat{P}_n\|Q) \leq D(P^*\|Q) + 2\delta$. The last step of the proof uses the Pythagorean identity of Theorem 8.10 to make this precise.

Let $P \in C$. Then,

$$D(P^*\|Q) + 2\delta \geq D(P\|Q) \geq D(P\|P^*) + D(P^*\|Q),$$

where the first inequality follows from the definition of C and the second from the Pythagorean identity. Hence, $D(P\|P^*) \leq 2\delta$, and from (52) we conclude that,

$$Q^n(D(\hat{P}_n\|P^*) \leq 2\delta | \hat{P}_n \in E) \to 1, \qquad \text{as } n \to \infty.$$

Since $\delta > 0$ was arbitrary, Pinsker's inequality in turn implies, that, for any $a \in A$,

$$Q^n\big(|\hat{P}_n(a) - P^*(a)| > \epsilon \big| \hat{P}_n \in E\big) \to 0, \qquad \text{as } n \to \infty.$$

This means that, conditional on $\{\hat{P}_n \in E\}$, the RVs $\hat{P}_n(a)$ converge to $P^*(a)$ in probability. Since they are bounded by 1, the bounded convergence theorem implies they also converge in L^1, so that, in particular, $\mathbb{E}(\hat{P}_n(a) | \hat{P}_n \in E) \to P^*(a)$, as $n \to \infty$, as desired. □

Remark on bounded convergence. If you do not like using (or do not know) the bounded convergence theorem, here is how to establish the last step in the proof above directly.

Lemma 8.12. *Suppose a sequence of RVs $\{Z_n\}$ converges to constant c in probability, and assume that all the $|Z_n| \leq M$ for some finite constant M. Then $\mathbb{E}(Z_n) \to c$ as $n \to \infty$.*

PROOF. For $\epsilon > 0$ arbitrary,

$$|\mathbb{E}(Z_n) - c| \leq \mathbb{E}(|Z_n - c|) = \mathbb{E}(|Z_n - c|\mathbb{I}_{\{|Z_n - c| > \epsilon\}}) + \mathbb{E}(|Z_n - c|\mathbb{I}_{\{|Z_n - c| \leq \epsilon\}}).$$

The first term is bounded above by,

$$(M + |c|)\mathbb{E}(\mathbb{I}_{\{|Z_n - c| > \epsilon\}}) = (M + |c|)\mathbb{P}(|Z_n - c| > \epsilon),$$

which goes to zero as $n \to \infty$, by the assumption $Z_n \to c$ in probability, and the second term is bounded above by ϵ. Therefore, $\limsup_{n\to\infty} |\mathbb{E}(Z_n) - c| \leq \epsilon$, and since $\epsilon > 0$ was arbitrary this completes the proof. □

Example 8.13. Now we can return (and answer) the question that we asked in the beginning of this section: If you roll a fair die 1000 times and get an average that is 5 or larger, then, according to the Gibbs conditioning principle, the proportion of 6s in your sample should be $\approx P^*(6)$, where P^* achieves the infimum in Sanov's theorem.

To compute P^* explicitly, observe that the current setting is a special case of the more general setting in Example 8.6. Let X_1^n be IID RVs with uniform distribution Q on $A = \{1, 2, 3, 4, 5, 6\}$, take $f(x) = x$, $x \in A$, $\mu = 3.5$, $\epsilon = 1.5$, and let $E = \{P \in \mathcal{P} \ : \ \mathbb{E}_{X\sim P}[X] \geq 5\}$. Then Corollary 8.9 tells us that the distribution P^* which achieves $\inf_{P\in E} D(P\|Q)$ is unique, and it is given by,

$$P^*(a) = \frac{e^{\lambda^* a}}{\sum_{b=1}^{6} e^{\lambda^* b}}, \qquad a \in A,$$

where $\lambda^* > 0$ is chosen so that $\mathbb{E}_{Y\sim P_{\lambda^*}}[Y] = 5$. Direct computation gives $\lambda^* \approx 0.6296$, and consequently,

$$P^* \approx (0.0205, 0.0385, 0.0723, 0.1357, 0.2548, 0.4781).$$

Therefore, if in 1000 die rolls we observe an average of at least 5, we should expect that we rolled a 6 around 47.8% of the time.

8.4 Error probability in fixed-rate data compression

In Section 1.2 we showed that, for a memoryless source, the best achievable compression rate was the source's entropy H: For any rate $R > H$ there is a code with rate R and vanishing probability of error, and any code with vanishing probability of error has rate at least H. Here we identify the fastest speed at which the probability of error can decay to zero: Theorem 8.14 states that, given a target rate $R > H$, the probability of error $P_e^{(n)}$ will, at best, decay like $P_e^{(n)} \approx 2^{-nD^*}$, where the optimal exponent D^* is given in (53).

Theorem 8.14 (Error exponents for fixed-rate compression)**.** *Suppose $\boldsymbol{X} = \{X_n\}$ is a memoryless source with entropy $H = H(X_i)$ and distribution Q which has full support on a finite alphabet A. Let R be any rate $H < R < \log|A|$.*

($\Rightarrow$) *There is a fixed-rate code $\{B_n^*\}$ with asymptotic rate no more than R bits/symbol,*

$$\limsup_{n\to\infty} \frac{1}{n}\big(1 + \lceil \log|B_n^*| \rceil\big) = \limsup_{n\to\infty} \frac{1}{n} \log|B_n^*| \leq R,$$

and probability of error $P_e^{(n)}$ that decays to zero exponentially fast,

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

with exponent,

$$D^* = \inf_{P:H(P)\geq R} D(P\|Q). \tag{53}$$

($\Leftarrow$) *If $\{B_n\}$ is any fixed-rate code with asymptotic rate no more than R bits/symbol,*

$$\limsup_{n\to\infty} \frac{1}{n}\big(1 + \lceil \log |B_n| \rceil\big) = \limsup_{n\to\infty} \frac{1}{n} \log |B_n| \leq R,$$

then its probability of error $P_e^{(n)}$ cannot decay faster than exponentially with exponent D^*:

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

PROOF. Write m for the alphabet size $|A|$. For each $n \geq 1$ we define the codebook,

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

By Proposition 8.3, each type class $T(P)$ in the above union has size $|T(P)| \leq 2^{nH(P)} \leq 2^{nR}$. Since the union is over no more than $|\mathcal{P}_n| \leq (n+1)^m$ types, we have,

$$|B_n^*| \leq (n+1)^m 2^{nR},$$

and hence,

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

For the probability of error, observe that,

$$P_e^{(n)} = \mathbb{P}(X_1^n \notin B_n^*) = \mathbb{P}\left(X_1^n \in \bigcup_{P\in\mathcal{P}_n : H(P)\geq R} T(P)\right) = \sum_{P\in\mathcal{P}_n : H(P)\geq R} Q^n(T(P)).$$

By Proposition 8.5, each type class in the above sum has probability,

$$Q^n(T(P)) \leq 2^{-nD(P\|Q)} \leq 2^{-nD^*},$$

therefore, using again the polynomial bound on $|\mathcal{P}_n|$ we obtain that,

$$\limsup_{n\to\infty} \frac{1}{n} \log P_e^{(n)} \leq \limsup_{n\to\infty} \frac{1}{n} \log \left[(n+1)^m 2^{-nD^*}\right] = -D^*.$$

This completes the proof of the direct part.

For the converse, let $\epsilon > 0$ arbitrary. By continuity, there is a $\delta > 0$ such that,

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

Since the n-types $\{\mathcal{P}_n \,;\, n \geq 1\}$ are dense in $\mathcal{P}$ and the $\mathcal{P}_n$ eventually intersect every open set in $\mathcal{P}$, for all n large enough we can find a $P_n \in \mathcal{P}_n$ such that $H(P_n) \geq R + \delta/2$ and,

$$D(P_n\|Q) \leq D^* + 2\epsilon. \tag{54}$$

First we claim that B_n is necessarily much smaller than the type class $T(P_n)$. Let $\{r_n\}$ be a sequence tending to zero as $n \to \infty$, such that $\frac{1}{n} \log |B_n| \leq R + r_n$, for each $n \geq 1$. By Proposition 8.3,

$$\frac{|B_n|}{|T(P_n)|} \leq \frac{2^{n(R+r_n)}}{(n+1)^{-m} 2^{nH(P_n)}} = (n+1)^m 2^{n[R-H(P_n)+r_n]} \leq (n+1)^m 2^{n\left[-\frac{\delta}{2}+r_n\right]}.$$

Since the right-hand side above $\to 0$ as $n \to \infty$, we have that, in particular,

$$\frac{|B_n|}{|T(P_n)|} \leq \frac{1}{2}, \qquad \text{eventually.} \tag{55}$$

We can now obtain the required lower bound on $P_e^{(n)}$ as follows. Letting $x_1^n \in A^n$ denote any string in $T(P_n)$, we have, eventually,

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

where (a) follows from Proposition 8.5 and (b) follows from the bounds in (54) and (55). Therefore,

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

and since $\epsilon > 0$ was arbitrary this completes the proof. □

Remarks.

1. An important observation here is that the code with the specific codebooks $\{B_n^*\}$ as described in the proof above is *universal*, in that it achieves the optimal error probability at the target rate R *simultaneously* for all memoryless sources with entropy $H < R$.

2. A reasonable question is to ask what happens if we try to compress a memoryless source at a rate R below the entropy. The fixed-rate coding theorem of Section 1.2 tells us that in this case the probability of error $P_e^{(n)}$ cannot tend to zero as $n \to \infty$, but if in fact it is possible to establish something much stronger: If the rate R of a code is smaller than the entropy, then its probability of error necessarily converges to 1 as $n \to \infty$, and it does so exponentially fast. Exercise 12.41 contains an outline of the proof of this result, which is sometimes referred to as a "strong converse" coding theorem.

9 Variable-rate lossless data compression

Let A be an arbitrary finite alphabet, and write $\{0,1\}^*$ for the set of all finite-length binary strings.

Definition 9.1 (Variable-rate codes)**.** *A* variable-rate lossless compression code *with blocklength n on the alphabet A is a pair (C_n, L_n) where the* encoder $C_n : A^n \to \{0,1\}^*$ *is a 1-1 function that maps source strings x_1^n to variable-length binary* codewords $C_n(x_1^n)$*, and the* length function $L_n : A^n \to \mathbb{N}$*,*

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

for each $x_1^n \in A^n$.

For both practical and theoretical reasons we find it convenient to concentrate on the class of *prefix-free codes*. From now on, we will always assume that all codes considered are prefix-free.

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

Example 9.3 (Codes)**.** Consider the four codes on $A = \{a, b, c, d\}$ given by:

X	$C(x)$
a	0
b	1
c	00
d	11

X	$C(x)$
a	0
b	00
c	110
d	111

X	$C(x)$
a	00
b	01
c	10
d	11

X	$C(x)$
a	0
b	10
c	110
d	111

The first two are not prefix-free, the last two are.

One obvious advantage of prefix-free codes is that they can be used repeatedly. For example, if a decoder receives '00' from the first code above, they cannot tell if the intended message was c or aa. With a prefix-free code, on the other hand, once a codeword is encountered, it can be immediately and unambiguously decoded, since no other codeword can have that as its prefix.

The following result expresses the fundamental combinatorial limitation on possible codeword lengths. The Kraft inequality is a precise version of the statement that "not all codelengths $L_n(x_1^n)$ can be short."

Theorem 9.4 (Kraft inequality)**.**

($\Rightarrow$) *For any length function $L_n : A^n \to \mathbb{N}$ satisfying* Kraft's inequality*,*

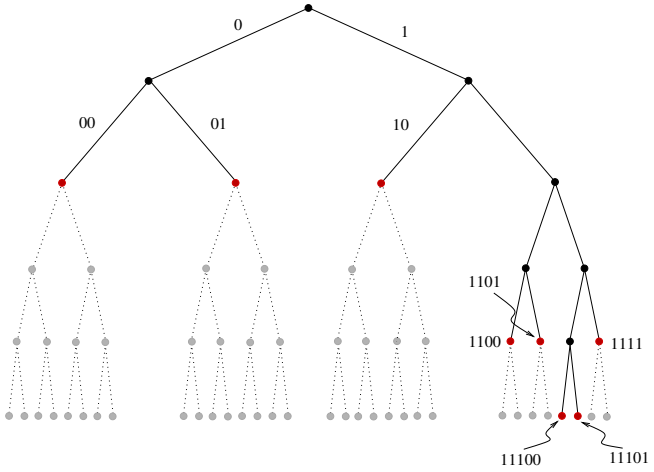
$$\sum_{x_1^n \in A^n} 2^{-L_n(x_1^n)} \leq 1, \tag{56}$$

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

($\Leftarrow$) *The length function L_n of any prefix free code (C_n, L_n) satisfies Kraft's inequality* (56)*.*

Proof. We first prove the converse.

Given a prefix-free code (C_n, L_n), let $L^* = \max_{x_1^n \in A^n} L_n(x_1^n)$ denote the maximum codeword length, and consider the complete binary tree of depth L^*. If we mark all the codewords on the tree then, since the code is prefix-free, no codeword will be a descendant of any other codeword. The figure below shows this construction for a code with codewords $\{00, 01, 10, 1100, 1101, 11100, 11101, 1111\}$ so that $L^* = 5$:



Every codeword with length $L_n(x_1^n)$ has $2^{L^*-L_n(x_1^n)}$ leaves at depth L^* as its descendants, and the prefix-free property guarantees that the collections of leaves corresponding to different codewords are disjoint. Therefore, the total number of leaves that descend from codewords must be no more that the total number of leaves at depth L^*, i.e.,

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

which is exactly (56).

For the direct part, given a collection of codeword lengths $\{L_n(x_1^n) \ ; \ x_1^n \in A^n\}$, we can define a prefix-free code following the same procedure. Let, again, $L^* = \max_{x_1^n \in A^n} L_n(x_1^n)$, fix an ordering of the strings x_1^n, and for each x_1^n assign $C_n(x_1^n)$ to be the lexicographically first available node at depth $L_n(x_1^n)$, where *available* means that it is neither a prefix nor a descendant of any of the previously assigned codewords. It is easy to see that Kraft's inequality guarantees there will always be such an available node. □

9.1 The codes-distributions correspondence

The following elegant result is a re-statement of the Kraft inequality.

Theorem 9.5 (Codes-distributions correspondence)**.**

(⇒) *For any distribution P_n on A^n, there is a prefix-free code (C_n^*, L_n^*) with:*

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

(⇐) *For any prefix-free code (C_n, L_n), there is a distribution Q_n on A^n such that:*

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

Proof. The direct part comes from the observation that, since,

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

the Kraft inequality guarantees that there is a code (C_n^*, L_n^*) with codelengths,

$$L_n^*(x_1^n) = \lceil -\log P_n(x_1^n) \rceil < -\log P_n(x_1^n) + 1 \quad \text{bits.}$$

For the converse part, given a prefix-free code (C_n, L_n), the Kraft inequality must be satisfied,

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

Therefore, if we define a PMF Q_n on A^n by

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

then,

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

as claimed. □

From Theorem 9.5 we easily deduce the well-known bounds establishing the entropy as the best achievable rate in lossless data compression:

Theorem 9.6 (Entropy bounds)**.** *Suppose X_1^n has an arbitrary distribution P_n on A^n.*

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

$$\mathbb{E}[L_n^*(X_1^n)] < H(X_1^n) + 1 \quad \textit{bits.}$$

($\Leftarrow$) *For any prefix-free code (C_n, L_n) on A^n:*

$$\mathbb{E}[L_n(X_1^n)] \geq H(X_1^n) \quad \textit{bits.}$$

Proof. For the direct part, taking P_n to be the true distribution of X_1^n in Theorem 9.5, we are guaranteed the existence of a code (C_n^*, L_n^*) that satisfies:

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

For the converse, given (C_n, L_n), let Q_n be the distribution obtained in Theorem 9.5. Then:

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

□

Another way of stating the results of the theorem is:

$$H(X_1^n) \leq \inf_{\text{prefix-free } (C_n, L_n)} \mathbb{E}[L_n(X_1^n)] < H(X_1^n) + 1.$$

If $\{X_n\}$ is a stationary source, so that its entropy rate exists, then dividing by n and letting $n \to \infty$ we obtain:

Corollary 9.7 (Entropy rate bound)**.** *Let $\boldsymbol{X} = \{X_n\}$ be a stationary source. The entropy rate $H = H(\boldsymbol{X})$ is the best asymptotically achievable compression rate, among all variable-rate, prefix-free codes:*

$$\lim_{n\to\infty} \inf_{\text{prefix-free } (C_n, L_n)} \frac{1}{n} \mathbb{E}[L_n(X_1^n)] = H \quad \textit{bits/symbol.}$$

9.2 Shannon codes and their properties

We begin with some observations.

Remarks.

1. Theorem 9.5 has a very important conceptual consequence: It says that the performance of *any* prefix-free code on A^n can be dominated by a code with codelengths of the form,

$$L_n(x_1^n) \approx -\log Q_n(x_1^n), \tag{57}$$

for some probability distribution Q_n on A^n, and that for any distribution Q_n such a code exists. This implies that the problem of finding a good code can be reduced to the problem of finding a good distribution Q_n. And this holds in complete generality, since the result assumes nothing about the distribution of the source X_1^n or the blocklength n. Indeed, this is precisely the foundation of the connection between information theory and statistics.

2. Given the importance implied by Theorem 9.5 for codes with codelengths of the form (57), it is appropriate to give them a name:

Definition 9.8 (Shannon code)**.** *Given a distribution Q_n on A^n, the code with codelengths,*

$$L_n(x_1^n) = \lceil -\log Q_n(x_1^n) \rceil, \quad \textit{for all } x_1^n \in A^n, \tag{58}$$

is called the Shannon code for Q_n.

Note that the terminology *Shannon-Fano code for* Q_n is also common. Remember, we are guaranteed the existence of such a code because the codelengths (58) satisfy the Kraft inequality as we saw in the proof of Theorem 9.5.

3. It is easy to see by the construction implied by the Kraft inequality that, despite the fact that we call it "the" Shannon code, it is in fact not unique; it is a simple but useful exercise to come up with examples of different codes C_n, C_n' with the same length function (58).

4. It is also easy to show via simple examples that, even if we take Q_n to be the true distribution of the data X_1^n, the Shannon code does not achieve the optimal (i.e., the minimal) expected description length $\mathbb{E}[L_n(X_1^n)]$. That is achieved by the Huffman code, which is not considered further in these notes. Nevertheless, the difference between them will always be less than 1 bit.

5. Although obvious, it is worth pointing out the fact that the Shannon code gives shorter descriptions to higher probability messages, and longer description to more unlikely ones.

6. Since the difference between $-\log Q_n(x_1^n)$ and $\lceil -\log Q_n(x_1^n) \rceil$ is always less than 1 bit, we will ignore the integer constraint and, from on, consider any length function $L_n : A^n \to [1, \infty)$ that satisfies Kraft's inequality to correspond to a prefix-free code. In the case of the Shannon code it is also convenient to give these non-necessarily-integer codelengths a name:

Definition 9.9 (Ideal Shannon codelengths)**.** *We call,*

$$L_n(x_1^n) = -\log Q_n(x_1^n), \qquad x_1^n \in A^n,$$

the ideal Shannon codelengths *for distribution Q_n.*

Apart from its near optimality in terms of the expected description length, the (ideal) Shannon code also enjoys a remarkably strong competitive optimality property:

Theorem 9.10 (Competitive optimality of the Shannon code)**.** *Suppose that $X_1^n \sim P_n$ for an arbitrary distribution P_n on A^n. Then for any other distribution Q_n:*

$$\mathbb{P}\big(-\log Q_n(X_1^n) \leq -\log P_n(X_1^n) - K \quad \text{bits}\big) \leq 2^{-K}. \tag{59}$$

Note that the result of Theorem 9.10 is extremely general, as it makes no assumptions on the source distribution P_n, and it applies to *any* coding distribution Q_n, or, via the codes-distributions correspondence, to any prefix-free code.

The bound in (59) is particularly strong, as it is independent of the blocklength n. For example, when compressing a 10 MByte file of $n = 8 \times 10^7$ bits, the probability that *any* compression algorithm will beat the Shannon code by more than a mere $K = 20$ bits is at most one-in-a-million.

Finally, it is quite satisfying that such a general result admits a very short and elementary proof:

Proof. Using Markov's inequality:

$$\begin{aligned}
\mathbb{P}\big(-\log Q_n(X_1^n) \leq -\log P_n(X_1^n) - K\big) &= \mathbb{P}\left(\frac{Q_n(X_1^n)}{P_n(X_1^n)} \geq 2^K\right) \\
&\leq 2^{-K}\mathbb{E}\left(\frac{Q_n(X_1^n)}{P_n(X_1^n)}\right) \\
&\leq 2^{-K}\sum_{x_1^n \in A^n} P_n(x_1^n) \times \frac{Q_n(x_1^n)}{P_n(x_1^n)} \\
&= 2^{-K}. \qquad \square
\end{aligned}$$

10 Entropy and additive combinatorics: Sumset bounds

Suppose A and B are two finite subsets of an abelian group $(G, +)$. What types of properties of A and B can help us estimate the cardinality of the *sumset* $A+B = \{a+b : a \in A, b \in B\}$? One of the main toolboxes in the important area of *additive combinatorics* is the so-called Plünnecke-Ruzsa *sumset theory*, which, among other things, provides results and techniques for estimating the sizes of sumsets in various circumstances. For example, there are the obvious bounds,

$$\max\{|A|, |B|\} \leq |A + B| \leq |A|\,|B|, \tag{60}$$

as well as more subtle results, like the Ruzsa triangle inequality,

$$|A - C| \leq \frac{|A - B|\,|B - C|}{|B|}, \tag{61}$$

which hold for arbitrary finite subsets A, B, C of G, and where the *difference set* $A-B$ is defined as $A - B = \{a - b : a \in A, b \in B\}$.

In this context, it is interesting to recall one of the interpretations of the entropy provided by the AEP. If X_1^n is a large collection of IID RVs with common PMF P on an alphabet A, then, writing $H = H(X_i)$ for their entropy, we know that the smallest subset $B_n \subset A^n$ with P^n-probability close to 1 has size $|B_n| \approx 2^{nH}$. In other words, the "effective support" of P^n has size $(2^H)^n$, as opposed to the size $|A|^n$ of the set of all possible strings of length n. In this sense, we can think of 2^H rather than $|A|$ as the "effective support size" of a RV with entropy H.

This interpretation suggests a correspondence between bounds for cardinalities of sets, and corresponding bounds for the entropy of discrete RVs. The idea is to start with a sumset bound such as (60) or (61), replace subsets of G by independent RVs taking finitely many values on G, and then replace the log-cardinality, $\log |A|$, of each set A by the entropy of the corresponding RV. Thus, e.g., the simple bounds (60) become,

$$\max\{H(X), H(Y)\} \leq H(X + Y) \leq H(X) + H(Y), \tag{62}$$

and Ruzsa's bound (61) becomes,

$$H(X - Z) \leq H(X - Y) + H(Y - Z) - H(Y). \tag{63}$$

In this section we will see that this intriguing correspondence does indeed lead to new inequalities for the entropy, and we will illustrate how the tools we have developed so far – especially the data processing inequality for mutual information – can be used to establish (62), (63), and some other interesting entropy bounds.

10.1 Simple sumset entropy bounds

First we prove the elementary bounds in (62); also see Exercise 12.16.

Proposition 10.1. *If X, Y are independent RVs taking values in a finite subset of G, then:*

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

Proof. For the lower bound we have,

$$\begin{aligned} H(X) + H(Y) &\overset{(a)}{=} H(X, Y) \\ &\overset{(b)}{=} H(Y, X + Y) \\ &\overset{(c)}{=} H(X + Y) + H(Y|X + Y) \\ &\overset{(d)}{\geq} H(X + Y) + H(Y), \end{aligned}$$

where (a) follows by independence, (b) follows by data processing because the map $(x,y) \mapsto (x, x+y)$ is 1-1, (c) follows from the chain rule, and (d) from the fact that conditioning reduces entropy. Therefore $H(X+Y) \geq H(X)$, and the same bound with $H(Y)$ in the right-hand side follows by interchanging the roles of X and Y.

A similar argument using the chain rule and the nonnegativity of conditional entropy gives,

$$H(X) + H(Y) = H(Y, X+Y) = H(X+Y) + H(Y|X+Y) \geq H(X+Y),$$

which yields the upper bound, $H(X+Y) \leq H(X) + H(Y)$. □

Definition 10.2 (Entropic Ruzsa distance)**.** *Let X, Y be random variables taking finitely many values on G. The (entropic)* Ruzsa *distance between X and Y is*

$$d_{\mathrm{R}}(X,Y) = H(X' - Y') - \frac{1}{2}H(X) - \frac{1}{2}H(Y),$$

where X', Y' are independent and have the same marginals as X and Y.

Next we prove (63).

Theorem 10.3 (Ruzsa triangle inequality for the entropy)**.** *If X, Y, and Z are independent G-valued RVs, then:*

$$H(X - Z) \leq H(X - Y) + H(Y - Z) - H(Y). \tag{64}$$

In other words,

$$d_{\mathrm{R}}(X,Z) \leq d_{\mathrm{R}}(X,Y) + d_{\mathrm{R}}(Y,Z).$$

First we need a lemma.

Lemma 10.4. *If X, Y, and Z are independent G-valued RVs, then,*

$$H(X,Y,Z) + H(X - Z) \leq H(X - Y, Y - Z) + H(X,Z), \tag{65}$$

which follows from,

$$I(X; (X - Y, Y - Z)) \geq I(X; X - Z). \tag{66}$$

Proof. The bound (66) follows easily from the data processing inequality for mutual information, since $X - Z = (X - Y) + (Y - Z)$, so X and $X - Z$ are conditionally independent given $(X - Y, Y - Z)$. To see that it implies (65), first note that the right-hand side of (66) is,

$$H(X - Z) - H(X - Z|X) = H(X - Z) - H(Z|X) = H(X - Z) - H(Z),$$

where we used the fact that, given $X = x$, the map $x - z \mapsto z$ is 1-1, and the independence of X, Z. And then observe that the left-hand side of (66) is,

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

where the first equality above follows from data processing for the entropy because the map $(x, x-y, y-z) \mapsto (x,y,z)$ is 1-1, and the second equality follows from the fact that $H(X,Y,Z) = H(X) + H(Y) + H(Z)$ by independence. □

Proof of Theorem 10.3. Rearranging (65) and using independence we obtain,

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

as claimed, where in the last step we used the subadditivity of the entropy. □

10.2 The doubling-difference inequality for entropy

If X_1, X_2 are IID G-valued RVs, then we know from Proposition 10.1 that both $H(X_1 + X_2)$ and $H(X_1 - X_2)$ are at least as large as $H(X_1)$. Here we ask the natural question, *does the sum or the difference of two IID RVs increase the entropy more*?

If we consider the "entropy-increase" due to addition of subtraction,

$$\begin{aligned} \Delta_+ &= H(X_1 + X_2) - H(X_1), \\ \Delta_- &= H(X_1 - X_2) - H(X_1), \end{aligned}$$

then Theorem 10.5 asserts that the ratio Δ_+/Δ_- must always lie between $1/2$ and 2.

Theorem 10.5 (Doubling-difference inequality)**.** *If X_1 and X_2 are IID G-valued RVs, then:*

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

Equivalently:

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

The proof will be based on two lemmas.

Lemma 10.6. *If X, Y, Z are independent G-valued RVs, then:*

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

Proof. This follows immediately from the triangle inequality in Theorem 10.3, upon replacing Y by $-Y$, since $H(-W) = H(W)$ for any RV W. □

Lemma 10.7. *If X, Y, Z are independent G-valued RVs, then,*

$$H(X + Y + Z) + H(Y) \leq H(X + Y) + H(Y + Z), \tag{67}$$

or, equivalently,

$$I(X + Y + Z; X) \leq I(X + Y; X). \tag{68}$$

Proof. For the equivalence of the two stated inequalities, first observe that,

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

where the first equality is the definition of mutual information, the second follows from the fact that, given $X = x$, the map $x + y + z \mapsto y + z$ is 1-1, and the third follows from the independence of X and $Y + Z$. And for the same reasons we similarly have,

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

Therefore, (67) and (68) are equivalent.

Since $X - (X + Y, Z) - (X + Y + Z)$ form a Markov chain, by the data processing inequality and the chain rule for mutual information we have,

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

and since Z is (conditionally) independent of X given $X + Y$, $I(Z; X|X + Y) = 0$, so

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

which proves (68) and hence the lemma. □

PROOF OF THEOREM 10.5. For the upper bound, from Proposition 10.1 and Lemma 10.7 we have,

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

so that, taking $X, -Y$ and Z IID,

$$H(X+Z)+H(X) \le 2H(X-Z),$$

or,

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

which holds for arbitrary IID X and Z, as claimed.

For the lower bound, Lemma 10.6 with X, Y, Z IID yields,

$$H(X-Y)+H(X) \le 2H(X+Y),$$

i.e.,

$$H(X-Y)-H(X) \le 2[H(X+Y)-H(X)],$$

which again holds for arbitrary IID X and Y, as stated.

Finally, the fact that the entropy bounds are equivalent to the corresponding mutual information bounds can be established easily as in the first part of the proof of Lemma 10.7: Since, conditional on $X_2 = x_2$, the map $(x_1 + x_2) \mapsto x_1$ is 1-1, by data processing and independence, $H(X_1 + X_2|X_2) = H(X_1|X_2) = H(X_2)$, therefore,

$$I(X_1 + X_2; X_2) = H(X_1 + X_2) - H(X_1),$$

and similarly we find that $I(X_1 - X_2; X_2) = H(X_1 - X_2) - H(X_1)$. □

Remarks.

1. It is tempting to think that, by taking our RVs to be uniformly distributed on certain sets, so that their entropies are equal to the log-cardinalities of those sets, we could recover the original sumset bounds from their entropy versions. As it turns out, this is generally not possible. The entropy versions are neither strictly weaker nor strictly stronger than the corresponding inequalities for set cardinalities.

2. The Ruzsa triangle inequality for the entropy in Proposition 10.3 was first established by Tao [20] using the fact that the entropy satisfies a *functional submodularity* property described in Exercise 12.46. The proof given here based on data processing was obtained a few years later in [16]. It is interesting to observe that Ruzsa's original proof in [18] of the triangle inequality (63) has a "data processing" flavor, based on the representation $a - c = (a - b) + (b - c)$ of elements of $A - C$ in terms of elements of $A - B$ and $B - C$. It is given below.

PROOF OF RUZSA'S TRIANGLE INEQUALITY. We will deduce that $|B||A-C| \le |A-B||B-C|$ from the fact that we can define an injective map $f : B \times (A - C) \to (A - B) \times (B - C)$.

Any element $y \in A - C$ has at least one representation $y = a - c$ for a pair $(a, c) \in A \times C$. For each $y \in A - C$ choose and fix a unique such representation, denoted $y = a(y) - c(y)$. We define f on $B \times (A - C)$ by $(x, y) \mapsto (a(y) - x, x - c(y))$. To see that f is invertible observe that, if $f(x, y) = (v, w) \in (A - B) \times (B - C)$, we can uniquely recover (x, y) as $y = v + w$ and $x = w + c(y)$. □

10.3 Small Ruzsa distance

The results developed in this section so far go in one direction: use known bounds for sets and obtain analogous inequalities for random variables and entropy. The other direction is, however, perhaps more important. Recently, Marton's conjecture was proved [10, 11] using a number of estimates for entropy. Some of the theory was developed in [12]. Marton was an information theorist, famous for the development of the entropy method in concentration inequalities. However, she formulated the aforementioned conjecture in additive combinatorics, which is also known in the literature as the polynomial Freiman-Ruzsa conjecture. The exact formulation is beyond the scope of this course. Nevertheless, at the heart of the proof lies a structural characterization of random variables with small entropic Ruzsa distance, which makes crucial use of the entropy–set correspondence and is reminiscent of the derivation of the typical set.

Proposition 10.8. *Let X, Y be independent random variables taking finitely many values in an abelian group G. Let $C \geq 4$ be a constant. There is a finite non-empty subset S of G such that, if U_S is a uniform random variable on S, then*

$$d_{\mathrm{R}}(U_S, Y) \leq (C+2)d_{\mathrm{R}}(X, Y) + h(1-2/C) \tag{69}$$

and

$$\log \frac{|S-S|}{|S|} \leq (2C+4)d_{\mathrm{R}}(X, Y) + 2h(1-2/C). \tag{70}$$

Proof. Suppose X, Y are independent and let $d = d_{\mathrm{R}}(X, Y)$ as shorthand. We have

$$H(X-Y) = \frac{1}{2}H(X) + \frac{1}{2}H(Y) + d \leq \frac{1}{2}H(X-Y) + \frac{1}{2}H(Y) + d, \tag{71}$$

and hence $H(X-Y) - H(Y) \leq 2d$. Writing P_X, P_{x-Y}, P_{X-Y} for the PMFs of $X, x-Y$ and $X-Y$ respectively, and rearranging the finite sums we have

$$\begin{aligned} H(X-Y) - H(Y) &= \sum_{t\in G}\sum_{x\in G} P_X(x)P_{x-Y}(t)\log\frac{1}{P_{X-Y}(t)} - \sum_{x\in G} P_X(x)H(x-Y) && (72)\\ &= \sum_{x\in G} P_X(x)D(x-Y\|X-Y) && (73)\\ &\leq 2d. && (74) \end{aligned}$$

For any $C \geq 4$,

$$S := \{x \in G : P_X(x) > 0, D(x-Y\|X-Y) \leq Cd\}. \tag{75}$$

Let $p := \mathbb{P}(X \in S) = \mathbb{P}(\mathbb{I}_S(X) = 1)$. By Markov's inequality and (74), it follows that

$$p = \mathbb{P}(X \in S) \geq 1 - 2/C \geq 1/2. \tag{76}$$

Note first that

$$H(X) = H(X, \mathbb{I}_S) = pH(X|X \in S) + (1-p)H(X|X \notin S) + h(p). \tag{77}$$

Also, since Y is independent of X and $\mathbb{I}_S(X)$, it follows that $H(X-Y|\mathbb{I}_S = i) \geq H(Y), H(X|\mathbb{I}_S = i)$ for $i = 0, 1$. Therefore

$$\begin{aligned} H(X-Y) &\geq H(X-Y|\mathbb{I}_S) = pH(X-Y|X\in S) + (1-p)H(X-Y|X\notin S)\\ &\geq pH(Y) + \frac{(1-p)}{2}(H(Y) + H(X|X\notin S)). && (78) \end{aligned}$$

Combining with (77) and using the definition of d, we obtain

$$2d \geq p(H(Y) - H(X|X \in S)) - h(p) \tag{79}$$

and since $\log|S| \geq H(X|X \in S)$, we have

$$\log|S| \geq H(Y) - \frac{2}{p}d - \frac{1}{p}h(p) \tag{80}$$

$$\geq H(Y) - 4d - 2h(1 - \frac{2}{C}), \tag{81}$$

where we used the monotonicity of $h(\cdot)$ for $C \geq 4$.

Now we claim that for any random variable Z taking values in S, independent of Y,

$$H(Y - Z) - H(Y) \leq Cd. \tag{82}$$

Combined with (80), (82) implies (69) by letting $Z = U_S$.

To prove the claim, assume Y, Z are independent and note that as in (73),

$$H(Y - Z) - H(Y) \leq \sum_{z \in S} P_Z(z) D(z - Y \| X - Y). \quad \text{(Why?)} \tag{83}$$

By definition of S, the right-hand side of (83) is at most Cd.

Finally, we establish (70). We may define (Z, Z') on $S \times S$, so that $Z - Z'$ is uniform on $S - S$ (how?). Note that Z and Z' may not be independent. Using (81) and (64), which holds even if X, Z there are not necessarily independent (Exercise 12.48),

$$\log\frac{|S - S|}{|S|} = H(Z - Z') - \log|S|$$

$$\leq H(Z - Z') - H(Y) + 4d + 2h(1 - \frac{2}{C})$$

$$\leq H(Z - Y) + H(Y - Z') - 2H(Y) + 4d + 2h(1 - \frac{2}{C}) \tag{84}$$

$$\leq 2Cd + 4d + 2h(1 - \frac{2}{C}), \tag{85}$$

where the last inequality follows by (82). Therefore,

$$\log\frac{|S - S|}{|S|} \leq (2C + 4)d + 2h(1 - \frac{2}{C}), \tag{86}$$

which is (70). □

11 The Entropy Power Inequality

In the previous section we saw several bounds for the entropy of sums of discrete random variables. In this section we will discuss a fundamental inequality, known as the *Entropy Power Inequality (EPI)*, which gives a lower estimate for the entropy of the sum of independent continuous RVs on $\mathbb{R}$. For continuous RVs the analogous notion is that of *differential entropy*:

Definition 11.1 (Differential entropy and relative entropy)**.** *For a continuous RV X on $\mathbb{R}$ with density f, the* differential entropy of X *is:*

$$h(X) = -\int_{\mathbb{R}} f(x) \log_e f(x)\, dx, \tag{87}$$

Let Y be another RV with density g. The relative entropy *between X and Y is*

$$D_e(X\|Y) = \int_{\mathbb{R}} f(x) \log_e \frac{f(x)}{g(x)}\, dx.$$

It is a common convention and somewhat more convenient to choose the natural base for the logarithm in the continuous setting. It is immediate to verify that the general definition of relative entropy in Remark 3 following the definition of relative entropy for discrete RVs reduces to the above definition in the case when the RVs have densities with respect to the Lebesgue measure.

The entropy power of X is

$$N(X) = \frac{1}{2\pi e} e^{2h(X)}.$$

A Gaussian random variable of variance σ^2 has entropy power σ^2 (verify this). The entropy power is therefore the variance of a Gaussian that has the same entropy as X, i.e., the effective variance of X.

Note that it is not always the case that the integral in (87) is well-defined. A usual convention is that we put $h(X) = -\infty$ and $N(X) = 0$ if X does not have a density or if the integral is not well-defined.

The differential entropy satisfies the scaling property

$$h(aX) = h(X) + \log_e |a|, a \neq 0 \tag{88}$$

and can thus be negative.

The Gaussian distribution has maximum entropy under fixed variance:

Proposition 11.2. *Let X be a random variable on $\mathbb{R}$ and Z a Gaussian with the same mean and variance. Then*

$$h(X) \leq h(Z).$$

Proof.

$$0 \leq D_e(X\|Z) = h(Z) - h(X). \text{ (Why?)}$$

This is the claimed inequality. □

It is useful to compare the above argument with the argument in Theorem 2.5.

Now suppose X has a continuously differentiable density f. In that case, we define the *score function* of X by

$$\rho_X(x) = \frac{f'(x)}{f(x)} = \frac{d}{dx} \log_e f(x), \quad x \in \mathbb{R}.$$

The *Fisher information* of X (with respect to the location parameter) is

$$I(X) = \mathbb{E}[\rho_X(X)^2] = \int_{\mathbb{R}} \frac{f'(x)^2}{f(x)}\, dx.$$

For $a \neq 0$, we have the scaling property

$$I(aX) = \frac{1}{a^2} I(X).$$

Lemma 11.3 (de Bruijn's identity)**.** *Let X have finite variance, and let $Z \sim \mathcal{N}(0,1)$ be independent of X. For $t > 0$,*

$$\frac{d}{dt} h(X + \sqrt{t}Z) = \frac{1}{2} I(X + \sqrt{t}Z).$$

PROOF. Let f_t be the density of $X + \sqrt{t}Z$. Thus, $f_t = f * \varphi_t$, where

$$\varphi_t(x) = \frac{1}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}}.$$

For $t > 0$, the density f_t is smooth and satisfies the heat equation

$$\frac{\partial}{\partial t} f_t \overset{*}{=} \frac{1}{2} \frac{\partial^2}{\partial x^2} f_t.$$

Therefore

$$\begin{aligned}
\frac{d}{dt} h(X + \sqrt{t}Z) &= -\frac{d}{dt} \int_{\mathbb{R}} f_t(x) \log_e f_t(x)\ dx \\
&\overset{*}{=} -\int_{\mathbb{R}} (\frac{\partial}{\partial t} f_t(x)) \log_e f_t(x)\ dx - \int_{\mathbb{R}} \frac{\partial}{\partial t} f_t(x)\ dx \\
&\overset{*}{=} -\frac{1}{2} \int_{\mathbb{R}} f_t''(x) \log_e f_t(x)\ dx.
\end{aligned}$$

We have not justified the exchange of differentiation and integration. The interested reader may find a justification in [3]. We also used $\int \frac{\partial}{\partial t} f_t = \frac{\partial}{\partial t} \int f_t = 0$. Integrating by parts* gives

$$-\int_{\mathbb{R}} f_t''(x) \log_e f_t(x)\ dx = \int_{\mathbb{R}} \frac{(f_t'(x))^2}{f_t(x)}\ dx.$$

This completes the proof. The justification for the steps in the proof marked with * are not examinable.

□

Lemma 11.4 (Stam's Fisher information inequality)**.** *Let X, Y be independent random variables with everywhere positive, continuously differentiable densities and finite Fisher information. Then, for every $a \in [0,1]$,*

$$I(X+Y) \leq a^2 I(X) + (1-a)^2 I(Y).$$

Consequently,

$$\frac{1}{I(X+Y)} \geq \frac{1}{I(X)} + \frac{1}{I(Y)}.$$

PROOF. Let $S = X + Y$. Write f, g for the densities of X, Y, and write ρ_X, ρ_Y, ρ_S for the corresponding score functions. The density of S is $f * g$, and

$$(f * g)'(s) \overset{*}{=} \int_{\mathbb{R}} f'(x) g(s-x)\, dx.$$

Therefore

$$\begin{aligned}\rho_S(s) &= \frac{(f*g)'(s)}{(f*g)(s)} \\ &= \int_{\mathbb{R}} \frac{f'(x)}{f(x)} \frac{f(x)g(s-x)}{(f*g)(s)}\, dx \\ &= \mathbb{E}[\rho_X(X) \mid X+Y=s].\end{aligned}$$

Similarly,

$$\rho_S(s) = \mathbb{E}[\rho_Y(Y) \mid X+Y=s].$$

Hence, for every $a \in [0,1]$,

$$\rho_S(S) = \mathbb{E}[a\rho_X(X) + (1-a)\rho_Y(Y) \mid S].$$

By Cauchy–Schwarz,

$$\begin{aligned}I(S) &\le \mathbb{E}\left[\big(a\rho_X(X) + (1-a)\rho_Y(Y)\big)^2\right] \\ &= a^2 I(X) + (1-a)^2 I(Y),\end{aligned}$$

because X and Y are independent and

$$\mathbb{E}[\rho_X(X)] = \int_{\mathbb{R}} f'(x)\, dx = 0.$$

This proves the first inequality. For the second see Exercise 12.50.

□

We shall also need continuity of entropy under small Gaussian perturbations.

Lemma 11.5 (Continuity of entropy)**.** *Let X be a RV with finite variance. Let $Z \sim \mathcal{N}(0,1)$ be independent of X. Then*

$$h(X + \sqrt{t}Z) \longrightarrow h(X) \qquad \text{as } t \downarrow 0.$$

Proof (non-examinable).

We will only prove the result in the special case when the density of X is continuous and bounded.

Note first that we always have (see Exercise 12.49 *(ii)*)

$$h(X + \sqrt{t}Z) \ge h(X).$$

For the reverse inequality, assume without loss of generality that $\mathrm{Var}(X) = 1$. By the scaling property of entropy, it suffices to show that

$$\limsup_{t \downarrow 0} h\Big(\sqrt{1-t}X + \sqrt{t}Z\Big) \le h(X).$$

Let f be the density of X, and let f_t be the density of $\sqrt{1-t}X + \sqrt{t}Z$. Since $\sqrt{1-t}X + \sqrt{t}Z \Rightarrow X$ as $t \to 0$ in distribution,

$$f_t(x) \to f(x)$$

pointwise. Let φ be the standard Gaussian density. Abusing the notation for relative entropy, it suffices (why?) to show that

$$\liminf_{t \downarrow 0} D_e(f_t \| \varphi) \ge D_e(f \| \varphi). \tag{89}$$

Since

$$u \log_e u \ge -e^{-1} \quad \text{uniformly},$$

we have by Fatou's lemma

$$\liminf_{t\downarrow 0} \int_{\mathbb{R}} \frac{f_t(x)}{\varphi(x)} \log_e\Big(\frac{f_t(x)}{\varphi(x)}\Big)\, \varphi(x)\, dx \geq \int_{\mathbb{R}} \frac{f(x)}{\varphi(x)} \log_e\Big(\frac{f(x)}{\varphi(x)}\Big)\, \varphi(x)\, dx,$$

which is (89).

□

Remark. For the interested reader, we note that in the general case, one has to justify a.e. convergence of densities in a different way as weak convergence is no longer enough.

Perhaps the cleanest way to proceed, however, is to justify (89) using lower semicontinuity of relative entropy with respect to weak convergence. This is a standard result and can be proved via the Donsker-Varadhan representation principle for relative entropy.

Theorem 11.6 (Entropy Power Inequality)**.** *Let X, Y be independent real-valued random variables with finite variances. Then*

$$N(X+Y) \geq N(X) + N(Y). \tag{90}$$

Equivalently, for every $\lambda \in [0,1]$,

$$h(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) \geq \lambda h(X) + (1-\lambda)h(Y). \tag{91}$$

Proof. We will prove the second inequality. Let Z_1, Z_2 be independent standard Gaussians, independent of X, Y. For $0 < t < 1$, define

$$X_t = \sqrt{t}X + \sqrt{1-t}Z_1 \qquad \text{and} \qquad Y_t = \sqrt{t}Y + \sqrt{1-t}Z_2.$$

Let

$$\Phi(t) = h\left(\sqrt{\lambda}X_t + \sqrt{1-\lambda}Y_t\right) - \lambda h(X_t) - (1-\lambda)h(Y_t).$$

At $t = 0$, all variables are Gaussian and therefore

$$\Phi(0) = 0. \quad \text{(Verify this.)}$$

Observe that by scaling

$$h(X_t) = \frac{1}{2}\log_e t + h\left(X + \sqrt{\frac{1-t}{t}}Z_1\right).$$

By de Bruijn's identity,

$$\frac{d}{ds}h(X + \sqrt{s}Z_1) = \frac{1}{2}I(X + \sqrt{s}Z_1).$$

Changing the variable as $s = \frac{1-t}{t}$, we have $\frac{ds}{dt} = -\frac{1}{t^2}$. Thus,

$$\frac{d}{dt}h(X_t) = \frac{1}{2t} - \frac{1}{2t^2} I\left(X + \sqrt{\frac{1-t}{t}}Z_1\right).$$

By the scaling property of Fisher information,

$$I\left(X + \sqrt{\frac{1-t}{t}}Z_1\right) = tI(X_t).$$

Therefore,

$$\frac{d}{dt}h(X_t) = \frac{1}{2t}\big(1 - I(X_t)\big) \qquad \text{and} \qquad \frac{d}{dt}h(Y_t) = \frac{1}{2t}\big(1 - I(Y_t)\big).$$

Now

$$\sqrt{\lambda}X_t + \sqrt{1-\lambda}Y_t = \sqrt{t}\left(\sqrt{\lambda}X + \sqrt{1-\lambda}Y\right) + \sqrt{1-t}\left(\sqrt{\lambda}Z_1 + \sqrt{1-\lambda}Z_2\right).$$

Since $\sqrt{\lambda}Z_1 + \sqrt{1-\lambda}Z_2 \sim \mathcal{N}(0,1)$, the same calculation gives

$$\frac{d}{dt}h\left(\sqrt{\lambda}X_t + \sqrt{1-\lambda}Y_t\right) = \frac{1}{2t}\left(1 - I\left(\sqrt{\lambda}X_t + \sqrt{1-\lambda}Y_t\right)\right).$$

Consequently,

$$\Phi'(t) = \frac{1}{2t}\left[\lambda I(X_t) + (1-\lambda)I(Y_t) - I\left(\sqrt{\lambda}X_t + \sqrt{1-\lambda}Y_t\right)\right].$$

By Stam's inequality, Lemma 11.4,

$$\begin{aligned} I\left(\sqrt{\lambda}X_t + \sqrt{1-\lambda}Y_t\right) &\leq \lambda^2 I(\sqrt{\lambda}X_t) + (1-\lambda)^2 I(\sqrt{1-\lambda}Y_t) \\ &= \lambda I(X_t) + (1-\lambda)I(Y_t). \end{aligned}$$

Hence

$$\Phi'(t) \geq 0,$$

which means that $\Phi(t)$ is increasing on $(0,1)$.

By entropy continuity, Lemma 11.5, all involved quantities are continuous in $t \in [0,1]$, and so we conclude that

$$\Phi(1) \geq \Phi(0) = 0.$$

But this reads

$$h\left(\sqrt{\lambda}X + \sqrt{1-\lambda}Y\right) \geq \lambda h(X) + (1-\lambda)h(Y).$$

For the equivalence of the two statements, see Exercise 12.51.

□

12 Examples papers

EXAMPLES PAPER #1: [Sections 1–3]

Exercise 12.1 (Gambling)**.** Consider the following game played sequentially by a large number of people. An initial $S_0 = 1$ dollar is placed on the table. The first player flips a fair coin and if it comes up Heads they take half of S_0, leaving $S_1 = 1/2$ on the table, while if it comes up Tails they add half of S_0 to the pot leaving $S_1 = 1.5$ dollars on the table. Then the second player repeats the same process, either taking half the pot or adding to it another half of its value, leaving a pot of $S_2 = (1/2)S_1$ or $S_2 = (3/2)S_1$ dollars on the table, each with probability $1/2$. Let S_n denote the pot size after n turns.

(*a*) What is the expected size of the pot $\mathbb{E}(S_n)$ after n players play?

(*b*) How does S_n behave for large n?

(*c*) How does the answer in part (*b*) compare with part (*a*)? Explain.

(*d*) Now consider the setting of the AEP in Theorem 1.1. Compute $\mathbb{E}[\frac{1}{P^n(X_1^n)}]$. How does this compare with the result of Theorem 1.1 for $\log \frac{1}{P^n(X_1^n)}$? Explain.

Exercise 12.2 (High-probability sets)**.** Suppose $\{X_n\}$ are IID RVs with distribution P and entropy H on the alphabet A, and for any $t > 0$ let,

$$C_n(t) = \Big\{ x_1^n \in A^n \ : \ P^n(x_1^n) \geq 2^{-nt} \Big\},$$

denote the subset of n-sequences with probabilities $\geq 2^{-nt}$.

(*a*) Show that $|C_n(t)| \leq 2^{nt}$, for all $n \geq 1$.

(*b*) For what values of t does $\mathbb{P}(X_1^n \in C_n(t)) \to 1$ as $n \to \infty$?

(*c*) Observe that the direct part of the fixed-rate coding theorem (Theorem 1.6) can be established using the sets $C_n(t)$ as codebooks instead of the typical sets B_n^*.

Exercise 12.3 (An AEP-like limit)**.** Let $\{X_n\}$ be IID $\sim P$ on A. Find:

$$\lim_{n\to\infty} \big[P^n(X_1^n)\big]^{1/n}.$$

Exercise 12.4 (Infinite entropy)**.** (*a*) Let $X \sim P$ on $A = \{2, 3, \ldots\}$, where $P(k) = \frac{C}{k(\log k)^2}$, $k \geq 2$, for an appropriate normalising constant C. Show that $H(X) = \infty$.

(*b*) Suppose that the PMF of a RV X with values in $A = \{1, 2, \ldots\}$ has nonincreasing probabilities, $P(k+1) \leq P(k)$, for all $k \geq 1$. Show that, if $H(X) < \infty$, the $\mathbb{E}[\log X] < \infty$.

Exercise 12.5 (AEP and source coding)**.** An IID source produces random bits with probabilities $P(1) = 0.005$ and $P(0) = 0.995$. The digits are taken 100 at a time and a binary codeword is provided for every sequence of 100 digits containing three or fewer 1s.

(*a*) Assuming that all codewords are the same length, find the minimum length required to provide codewords for all sequences with three or fewer 1s. How good is the rate of this code compared to the source's entropy?

(*b*) Calculate the probability of error for this code, i.e., the probability of observing a source sequence for which no codeword has been assigned.

(*c*) Use Chebyshev's inequality to bound the probability of observing a source sequence for which no codeword has been assigned. Compare this bound with the actual probability computed in part (*b*).

Exercise 12.6 (Random codebooks)**.** Let $\{X_n\}$ be a memoryless source with distribution P on the alphabet A, and with entropy $H = H(X_i)$. Consider a *random* fixed-rate code $\{B_n\}$ with each codebook B_n consisting of $K_n := \lceil 2^{nR} \rceil$ IID strings $Y_1^n(1), \ldots, Y_1^n(K_n)$, where each $Y_1^n(i)$ has distribution P^n. We assume that the codebooks $\{B_n\}$ are independent of the source $\{X_n\}$.

Show that, if $R > H$, then the probability of error of this code $\mathbb{P}(X_1^n \notin B_n)$ tends to zero as $n \to \infty$, where the probability is now over both X_1^n and the random codebook B_n. *Hint.* Recall the simple inequality $(1-p)^{1/p} \leq e^{-1}$, $p \in (0,1)$.

Exercise 12.7 (Monotonic convergence of the empirical distribution)**.**
Let $\hat{P}_n$ denote the empirical distribution corresponding to IID $X_1^n \sim P$ on A, defined in (16). Since the X_1^n are random, the PMF $\hat{P}_n$ is also random.

(*a*) Show that:

$$\mathbb{E}\big[D(\hat{P}_{2n} \| P)\big] \leq \mathbb{E}\big[D(\hat{P}_n \| P)\big].$$

Thus, the expected relative entropy "distance" from the empirical distribution to the true distribution decreases when we double the sample size. (*Hint*: Write $\hat{P}_{2n} = \frac{1}{2}\hat{P}_n + \frac{1}{2}\hat{P}'_n$.)

(*b*) Show that, in fact:

$$\mathbb{E}\big[D(\hat{P}_n \| P)\big] \leq \mathbb{E}\big[D(\hat{P}_{n-1} \| P)\big].$$

Thus, the expected relative entropy between the empirical and the true distribution actually decreases with n. (*Hint*: Write $\hat{P}_n$ as the average of n empirical PMFs, with each of the n samples deleted in turn.)

Exercise 12.8 (Convexity and concavity)**.** (*a*) Write out the proof of Theorem 3.12.

(*b*) Show that $H(P)$ is in fact *strictly* concave, using the log-sum inequality directly, without appealing to Theorem 3.12.

(*c*) Similarly, show that, for a fixed Q of full support (i.e., $Q(a) > 0$ for all $a \in A$) $D(P\|Q)$ is *strictly* convex in P.

Exercise 12.9 (Infinite alphabets)**.** Let P, Q be PMFs on the same countably infinite alphabet A, such that:

$$\sum_{x \in A} P(x) \left| \log \frac{P(x)}{Q(x)} \right| < \infty. \tag{92}$$

(*a*) Argue that, assuming $D(P_X\|Q_X)$ satisfies (92), the result and proof of the the data processing property for relative entropy in Theorem 3.9 remain valid in the case of countably infinite alphabets. Similarly, argue that, assuming $D(P\|Q)$ satisfies (92), Pinsker's inequality in Theorem 3.11 remains valid for countably infinite alphabets.

(*b*) Using Pinsker's inequality, show that:

$$\sum_{x \in A} P(x) \left| \log \frac{P(x)}{Q(x)} \right| \leq D(P\|Q) + \big[(2 \log e) D(P\|Q)\big]^{1/2}.$$

This bound was first derived by Barron [4].

Exercise 12.10 (Simple entropy estimation)**.** Let $X_1, X_2, \ldots$ be IID RVs with distribution $(1/2, 1/4, 1/8, 1/8)$ on $A = \{1, 2, 3, 4\}$.

(*a*) Use your favourite computer program to simulate about 100 such RVs X_i, and calculate the empirical distribution $\hat{P}_n$ from your samples $X_1, X_2, \ldots, X_{100}$.

(*b*) Pretend you do not know the true distribution, and use the empirical distribution to come up with an estimate $\hat{H}$ for the entropy, by taking:

$$\hat{H} = H(\hat{P}_n).$$

Compare your estimate with the true value of the entropy.

(*c*) Repeat steps (*a*) and (*b*) 10 or 20 times. Comment on the estimates you get. Do you see any systematic trends?

Exercise 12.11 (More about entropy estimation)**.** Show that the entropy estimate in Exercise 12.10 is biased "downwards:" Let X_1^n be IID RVs with distribution P and entropy $H(P)$ over the finite alphabet A. Prove that, unless $H(P) = 0$, we have:

$$\mathbb{E}_P[H(\hat{P}_n)] < H(P).$$

Exercise 12.12 (No unbiased estimator for the entropy)**.** Exercise 12.11 showed that the "plug-in" entropy estimate, where we simply compute the entropy of the empirical distribution, was biased "downwards." Here, you will show that in fact there is *no* unbiased estimator for the entropy of a binary IID process: Let X_1^n be binary IID Bern(θ) RVs, and suppose that, for a fixed n, $T : \{0,1\}^n \to \mathbb{R}$ is a given entropy estimator. Show that it is impossible to have,

$$\mathbb{E}_P[T(X_1^n)] = H(P),$$

for every $P \sim \text{Bern}(\theta)$, i.e., for every $\theta \in (0,1)$.

Exercise 12.13 (Hypothesis testing exponent)**.** Suppose $\{X_n\}$ are IID RVs taking values in $A = \{1, 2, \ldots\}$. Consider two hypotheses: $H_0 : X_i \sim P$, where $P(k) = 2^{-k}$, and $H_1 : X_i \sim Q$, where $Q(k) = (1-p)p^{k-1}$, $k \geq 1$, for some $p \neq 1/2$.

Compute the optimal error exponent in Stein's lemma for this test.

Exercise 12.14 (Testing for the wrong distribution)**.** We wish to test IID data $X_1, X_2, \ldots$ for having distribution either P_1 or P_2 on a finite alphabet A, based on a sample x_1^n. Suppose we use the Neyman-Pearson test with an appropriately chosen threshold, corresponding to the setting of Stein's lemma. What would happen if the data actually came from a third distribution Q? Describe the possible results of the test, not the associated error probabilities.

Exercise 12.15 (Conditional entropy properties)**.** Prove the three inequalities for conditional entropy stated in (*i*)–(*iii*) of Proposition 3.7.

Exercise 12.16 (Entropy of sums)**.** Suppose X, Y are discrete RVs with alphabets A, B, respectively, where A and B are finite subsets of $\mathbb{R}$.

(*a*) Show that $H(X+Y|X) = H(Y|X)$. Argue that, if X, Y are independent, then $H(X+Y) \geq \max\{H(X), H(Y)\}$. Thus, the addition of *independent* RVs adds uncertainty.

(*b*) Give an example of RVs X, Y such that $H(X+Y)$ is strictly smaller than both $H(X)$ and $H(Y)$.

(*c*) Give a simple set of sufficient conditions for $H(X+Y) = H(X) + H(Y)$.

Exercise 12.17 (Fano without conditioning)**.** Suppose X is a RV with PFM P on $A = \{1, 2, \ldots, m\}$, such that $p := P(1) > P(k)$ for all $k > 1$. The minimum-probability-of-error predictor of X is $\hat{X} = 1$, with probability of error $P_e = 1 - p$. Find the maximal entropy $H(X)$ subject to the above conditions, for a fixed $p \in (0, 1)$, and use it to obtain a bound on P_e in terms of $H(X)$. This is Fano's inequality without conditioning.

Exercise 12.18 (Conditioning that does not decrease entropy)**.** Proposition 3.4 showed that we always have $H(Y|X) \leq H(Y)$. Give an example where $H(Y|X = x) > H(Y)$ for a fixed value x.

Exercise 12.19 (Maximum entropy with given marginals)**.** Let (X, Y) be a pair of random variables with joint PMF P_{XY}, such that the marginal distribution of X is uniform on $\{1, 2, 3, 4, 5\}$ and the marginal of Y is $(1/2, 1/4, 1/8, 1/8)$ on $\{1, 2, 3, 4\}$. What is the maximum entropy PMF P_{XY} given these constraints?

Exercise 12.20 (Relative entropy and χ^2 distance)**.** Since $D(P\|Q)$ is nonnegative and it is only zero when $P = Q$, it is natural to expect that it is locally quadratic around $P \approx Q$. Pinsker's inequality gives a quadratic lower bound. Here you will prove a quadratic upper bound with respect to the χ^2 distance between two PMFs P, Q on the same alphabet A:

$$\chi^2(P\|Q) := \sum_{x \in A} \frac{(P(x) - Q(x))^2}{Q(x)}.$$

Show that:

$$D_e(P\|Q) \leq \chi^2(P\|Q).$$

EXAMPLES PAPER #2: [Sections 4–7]

Exercise 12.21 (A simple calculus lemma)**.** Prove Lemma 4.2

Exercise 12.22 (Poisson approximation)**.** Suppose X_1^n are IID Bern$(1/n^{3/4})$ RVs, and let S_n denote their sum $X_1 + X_2 + \cdots + X_n$.

(*a*) Give reasons why we cannot expect the distribution P_{S_n} of the sum to converge to any reasonable distribution as $n \to \infty$.

(*b*) Explain why it still makes sense to approximate P_{S_n} by a Poisson distribution, and derive an appropriate bound.

Exercise 12.23 (Binomial maximum entropy)**.** In this question you will prove the maximum entropy property of the binomial distribution stated in Proposition 4.5.

(*a*) Suppose a function $f(x, y)$ is defined for x, y such that $x + y =$ a constant c, let $z = x - \frac{c}{2}$, and define a new function $g(z) = f(\frac{c}{2} + z, \frac{c}{2} - z)$. Show that if $g''(z) \leq 0$ then f is jointly concave in (x, y).

(*b*) Suppose that f is also symmetric in x, y, i.e., $f(x, y) = f(y, x)$. Show that f achieves its maximum at the point $(x, y) = (\frac{c}{2}, \frac{c}{2})$.

(*c*) Take $n \geq \lambda > 0$ fixed, let $S_n = \sum_{i=1}^n X_i$ where the X_i are independent Bern(p_i) RVs such that $\sum_i p_i = \lambda$, and let $h(p_1, p_2, \ldots, p_n) = H_e(S_n)$. Using part (a), you will prove the following claim: *For any two $i \neq j$, h is a concave function of (p_i, p_j) when all the rest of the p's are kept fixed.*

i. Without loss of generality (why?) take $i = 1$ and $j = 2$, so that $p_1 + p_2$ is equal to the constant $c = \lambda - p_3 - p_4 - \cdots - p_n$. Let $f(p_1, p_2) = h(p_1, p_2, \ldots, p_n)$, $z = p_1 - \frac{c}{2}$, and $g(z) = f(\frac{c}{2} + z, \frac{c}{2} - z)$. Define $u = z^2$, and explain why $\frac{d^2}{du^2} H_e(S_n) \leq 0$.

ii. Express $\mathbb{P}(S_n = \ell)$ as the product of the symmetric functions s_ℓ^n (evaluated at suitable arguments) and a simple product involving the p_i. Recall that the s_k^n are defined by,

$$s_k^n(x_1, x_2, \ldots, x_n) = \sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq n} x_{i_1} x_{i_2} \cdots x_{i_k},$$

for all positive x_i, and that they satisfy,

$$s_k^n \, s_{k+2}^n \leq (s_{k+1}^n)^2. \tag{93}$$

iii. Expressing $\frac{d}{du} H_e(S_n)$ as a telescoping sum show that,

$$\frac{d}{du} H_e(S_n) = \sum_{\ell=0}^{n} \log_e \left(\frac{\mathbb{P}(S_n = \ell)\mathbb{P}(S_n = \ell + 2)}{\mathbb{P}(S_n = \ell + 1)^2} \right) \mathbb{P}(X_3 + X_4 + \cdots + X_n = \ell),$$

and use part ii. to conclude that $\frac{d}{du} H_e(S_n) \leq 0$.

iv. Observe that,

$$\frac{d^2}{dz^2} H_e(S_n) = 2 \frac{d}{du} H_e(S_n) + \left(\frac{du}{dz} \right)^2 \frac{d^2}{du^2} H_e(S_n), \tag{94}$$

and use ii. and iii. to show that $g''(z) \leq 0$. Conclude that the claim in (c) is true.

(*d*) Prove the result of Proposition 4.5.

Exercise 12.24 (Binomial to Poisson entropy)**.** Let P, Q be two PMFs on the same (finite or countably infinite) alphabet A, write $\delta(x) = |P(x) - Q(x)|$, $x \in A$, for the differences, and let Δ denote their total variation distance, $\Delta = \|P - Q\|_{\mathrm{TV}} = \sum_{x \in A} \delta(x)$.

(*a*) By considering the concave function $f(t) := -t \log t$, $0 \le t \le 1$, show that, if $\Delta \le 1/2$, then:

$$|H(P) - H(Q)| \le -\sum_{x \in A} \delta(x) \log \delta(x).$$

(*b*) Let P_n, Q be the $\mathrm{Bin}(n, \lambda/n)$ and the $\mathrm{Po}(\lambda)$ PMFs, respectively. Recall from the proof of Theorem 4.5 that $H(Q) < \infty$, and from (25) that the total variation distance $\Delta_n = \|P_n - Q\|_{\mathrm{TV}} \le \sqrt{2}\lambda/\sqrt{n}$.

By considering the series in part (*a*) separately for $0 \le x \le n$ and for $x > n$, prove that $H(P_n) \to H(Q)$, as $n \to \infty$.

Exercise 12.25 (Conditional mutual information)**.** Prove the following properties of conditional mutual information, generalising the results in Propositions 5.2, 5.3 and 5.4.

(*a*) The conditional mutual information can alternatively be expressed as:

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

(*b*) The following version of the chain rule holds:

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

(*c*) $I(X;Z|Y) \ge 0$ with equality if and only if X and Z are conditionally independent given Y, and $I(X;Y|Z)$ is always bounded above by $H(X|Z)$.

Exercise 12.26 (Conditioning and mutual information)**.** Give three different examples of RVs (X, Y_1, Y_2) such that the "synergy" $S(X; Y_1, Y_2)$ defined in (31) is positive, negative, and zero.

Exercise 12.27 (Even binary strings)**.** Suppose X_1^n is a string selected uniformly among all strings $\{0,1\}^n$ with an even number of 1s. Find the mutual informations:

$$I(X_1; X_2),\ I(X_2; X_3|X_1),\ \ldots,\ I(X_{n-1}; X_n|X_1^{n-2}).$$

Exercise 12.28 (Mutual information inequalities)**.** Prove the two inequalities:

$$\begin{aligned} &(a) \qquad I(X,Y;Z) \ge I(X;Z), \\ &(b) \qquad I(X;Z|Y) \ge I(Z;Y|X) - I(Z;Y) + I(X;Z), \end{aligned}$$

and give conditions for equality.

Exercise 12.29 (Synergy)**.** Show that the conclusion of Proposition 5.7 is valid if and only if the synergy $S(X; W_1, W_2)$ satisfies $S(X; W_1, W_2) > -I(Z_1; Z_2)$.

Exercise 12.30 (Arrow of time)**.** As we know, for example, from Markov chains, there are non-reversible processes $\{X_n\ ;\ n \in \mathbb{Z}\}$, i.e., processes for which the distribution of $(X_1, \ldots, X_n)$ is different from that of $(X_n, \ldots, X_1)$. Nevertheless, show that, if $\{X_n\}$ is stationary, then we always have:

$$H(X_0|X_{-1}, \ldots, X_{-n}) = H(X_0|X_1, \ldots, X_n).$$

Exercise 12.31 (Past and future)**.** Suppose $\{X_n\ ;\ n \geq 1\}$ is stationary. Although the existence of the entropy rate tells us that the information $H(X_1^n)$ in X_1^n grows linearly with n, show that the mutual information between X_1^n and X_{n+1}^{2n} does not:

$$\lim_{n\to\infty} \frac{1}{n} I(X_1^n; X_{n+1}^{2n}) = 0.$$

Exercise 12.32 (Entropy rate given the recent past)**.** Suppose $\{X_n\ ;\ n \in \mathbb{Z}\}$ is a stationary process with entropy rate H. Show that, for any fixed $k \geq 0$:

$$\lim_{n\to\infty} \frac{1}{n} H(X_1^n | X_{-k}^0) = H.$$

Exercise 12.33 (Entropy rate of a mixture)**.** Suppose $\boldsymbol{X}^{(0)} = \{X_n^{(0)}\}$ are IID Bern(p) and $\boldsymbol{X}^{(1)} = \{X_n^{(1)}\}$ are IID Bern(q), independent of the $\boldsymbol{X}^{(0)}$, for some $q \neq p$. Let $I \sim \text{Bern}(1/2)$ be independent of $\boldsymbol{X}^{(0)}$ and $\boldsymbol{X}^{(1)}$, and define a new process $\{Y_n\}$ as $Y_n = X_n^{(I)}$, $n \geq 1$.

(*a*) Is $\{Y_n\}$ stationary?

(*b*) Is $\{Y_n\}$ an IID sequence?

(*c*) Is $\{Y_n\}$ ergodic?

(*d*) What is the entropy rate H of $\{Y_n\}$?

(*e*) Let P_n denote the PMF of Y_1^n, for each $n \geq 1$. As $n \to \infty$, does,

$$-\frac{1}{n} \log P_n(Y_1^n) \to H?$$

Now let $\{I_n\}$ be an IID Bern(1/2) sequence, and define a new process $\{Z_n\}$ by $Z_n = X_n^{(I_n)}$, $n \geq 1$. Answer (*a*)–(*e*) again, with $\{Z_n\}$ in place of $\{Y_n\}$.

EXAMPLES PAPER #3: [Sections 8–11]

Exercise 12.34 (Size of type class)**.** In Proposition 8.3 we saw that:

$$\frac{1}{(n+1)^m}2^{nH(P)} \le |T(P)| \le 2^{nH(P)}.$$

Here we will see that in fact,

$$|T(P)| \approx n^{-\frac{m-1}{2}}2^{nH(P)}.$$

More precisely, assuming P is an n-type with all positive probabilities on a alphabet A of size $m \ge 2$, show that,

$$\log|T(P)| = nH(P) - \frac{m-1}{2}\log(2\pi n) - \frac{1}{2}\sum_{a\in A}\log P(a) - \frac{\vartheta(n,P)}{12}(\log e)m,$$

where $0 \le \vartheta(n,P) \le 1$. *Hint.* Use Robbins' sharpening of Stirling's formula [9, p. 54],

$$\sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n+\frac{1}{12(n+1)}} \le n! \le \sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n+\frac{1}{12n}},$$

noticing that $P(a) \ge \frac{1}{n}$ whenever $P(a) > 0$.

Exercise 12.35 (A tighter upper bound in Sanov's theorem)**.** Suppose X_1^n are IID with distribution Q on A, and let $E \subset \mathcal{P}$ be a convex set of probability distributions on A. In this case you will show that the upper bound of Sanov's theorem holds even without the polynomial term $(n+1)^m$:

$$Q^n(\hat{P}_{X_1^n} \in E) \le 2^{-nD(P^*\|Q)}.$$

(*i*) Let $X \sim Q$ be an arbitrary discrete RV, take any event B with $Q(B) \ne 0$, and let $P_{X|B}$ denote the conditional distribution of X given the event B. Show that

$$D(P_{X|B}\|Q) = -\log Q(B).$$

(*ii*) Derive the desired upper bound by justifying each of the steps below and explaining all the extra notation introduced:

$$\begin{aligned}
-\log Q^n(\hat{P}_{X_1^n} \in E) &\overset{(a)}{=} -\log Q^n(B)\\
&\overset{(b)}{=} D(P_{X_1^n|B}\|Q^n)\\
&\overset{(c)}{=} \sum_{i=1}^n \sum_{x_1^{i-1}} P_{X_1^{i-1}|B}(x_1^{i-1}) D\Big(P_{X_i|B,X_1^{i-1}}(\cdot|x_1^{i-1})\Big\|Q(\cdot)\Big)\\
&\overset{(d)}{\ge} n\sum_{i=1}^n \frac{1}{n}D(P_{X_i|B}\|Q)\\
&\overset{(e)}{\ge} nD\Big(\Big(\frac{1}{n}\sum_{i=1}^n P_{X_i|B}\Big)\Big\|Q\Big)\\
&\overset{(f)}{\ge} nD(P^*\|Q).
\end{aligned}$$

Exercise 12.36 (Counting)**.** Let $A = \{1,2,\ldots,m\}$, let $g : A \to \mathbb{R}$ be a given function, and $\alpha \in \mathbb{R}$. Show that the number of strings $x_1^n \in A^n$ satisfying,

$$\frac{1}{n}\sum_{i=1}^n g(x_i) \ge \alpha, \tag{95}$$

is approximately equal to 2^{nH^*}, to first order in the exponent, where,

$$H^* = \sup_{P\,:\,\sum_{j=1}^m P(j)g(j)\geq\alpha} H(P).$$

Exercise 12.37 (Bent coins)**.** Let $\{X_i\}$ be IID $\sim Q$ where Q is the Bin(m,q) PMF on $A = \{0,1,\ldots,m\}$, for some $q \in (0,1)$. Let $\alpha \in (mq, m)$. Show that, as $n \to \infty$,

$$\mathbb{P}\Big(X_1 = k \,\Big|\, \frac{1}{n}\sum_{i=1}^n X_i \geq \alpha\Big) \to P^*(k), \qquad \text{for each } k \in A,$$

where P^* is the Bin(m,λ) PMF for some $\lambda \in [0,1]$. Identify λ.

Exercise 12.38 (Large deviations below the mean)**.** Let $\{X_n\}$ be IID with distribution Q on a finite alphabet A, and $f : A \to \mathbb{R}$ arbitrary. State, without proof:

(a) The corresponding result to Chernoff's bound for the probabilities,

$$\mathbb{P}(S_n \leq \mu - \epsilon),$$

defining all the relevant terms and quantities.

(b) The definition of the relevant log-moment generating function $\Lambda(\lambda)$ and relevant exponent $\Lambda^*(c)$ here.

Exercise 12.39 (Rare events)**.** Suppose you are a public health official, and you want to get some statistical information about what would happen in the event of a major biological disaster, such as a major epidemic, or in case of a biological weapons attack. How do you simulate events in a very unlikely scenario?

Here is a cartoon version of this problem: Suppose you roll a fair die 10,000 times, and instead of getting a 6 approximately 1/6 of the time, you get a proportion of 6s at least 25%. Clearly this is a very rare event. Now suppose someone asks you, "If I rolled a die 10,000 times and got 6s more than 25% of the time, what should I expect the overall average of my 10,000 die rolls to be?". If you tried to directly simulate this situation, you would be in trouble – it would almost never happen that you get $\geq$25% 6s in 10,000 simulated rolls. Instead, use the Gibbs conditioning principle to answer this question analytically.

Exercise 12.40 (Functional lower bound to relative entropy)**.** Let $X \sim P$ and $Y \sim Q$ be two RVs with values on the same finite alphabet A. Show that for any $g : A \to \mathbb{R}$:

$$D_e(P\|Q) \geq \mathbb{E}[g(X)] - \log_e \mathbb{E}[e^{g(Y)}].$$

Exercise 12.41 (Strong converse in data compression)**.** Let $\boldsymbol{X} = \{X_n\}$ be a memoryless source with entropy $H = H(X_i)$ and distribution Q with full support on a finite alphabet A of size $|A| = m$. Here you will prove the following "strong converse" coding theorem for fixed-rate data compression: If $\{B_n\}$ is any fixed-rate code with asymptotic rate,

$$\limsup_{n\to\infty} \frac{1}{n}\big(1 + \lceil \log|B_n| \rceil\big) = \limsup_{n\to\infty} \frac{1}{n}\log|B_n| \leq R,$$

where R is strictly smaller than H, then its probability of error $P_e^{(n)} = Q^n(B_n^c)$ converges to 1 exponentially fast, namely,

$$\limsup_{n\to\infty} \frac{1}{n}\log Q^n(B_n) \leq -D_*(R),$$

where,

$$D_*(R) = \inf_{P:H(P)\leq R} D(P\|Q) > 0.$$

(*a*) Show that for any n-type P, the proportion of sequences in $T(P)$ that belong to B_n is at most $(n+1)^m 2^{-n|H(P)-R_n|^+}$, where, $R_n = \frac{1}{n}\log|B_n|$ and, as usual, $|x|^+ = x$ if $x \geq 0$ and $= 0$ otherwise.

(*b*) Use the result of part (*a*) to show that,

$$\limsup_{n\to\infty} \frac{1}{n}\log Q^n(B_n) \leq -\inf_{\text{all } P}[D(P\|Q) + |H(P) - R|^+].$$

(*c*) Show that,

$$D_*(R) = \inf_{\text{all } P}[D(P\|Q) + |H(P) - R|^+].$$

Hints. 1. Before starting, read the proof of the converse part of Theorem 8.14 carefully. 2. For part (c), observe that it suffices to show that the infimum over all P above is actually achieved when $H(P) \leq R$. 3. To prove this, show that for any P with $H(P) > R$ there is always a P^* with $H(P^*) = R$ that does at least as well. 4. One way to construct such a P^* is as a mixture of P and the point mass at an appropriately chosen $y^* \in A$.

Exercise 12.42 (Shannon codes)**.** (*a*) Give an explicit example of two different prefix-free codes C_n, C_n' and a distribution Q_n, such that both achieve the Shannon codelengths (58) for Q_n.

(*b*) Construct a Shannon code for the distribution P on the alphabet $A = \{a, b, r, c, d\}$, where:

$$P = (3/10, 1/4, 1/4, 1/10, 1/10).$$

(*c*) Encode the message "abracadabra" (in binary) using the code in part (*b*). How many bits did you need? Are you happy with the compression performance of this code? Why, or why not?

Exercise 12.43 (Expected redundancy)**.** Suppose that we have data X_1^n with distribution P_n on A^n, but we compress X_1^n with respect to the optimal code for a different distribution Q_n on A^n. How many extra bits does the (ideal) Shannon code for Q_n use compared to the entropy?

Exercise 12.44 (Slackness in Kraft inequality)**.** Suppose (C_n, L_n) is a prefix-free code such that we have strict inequality is Kraft's inequality:

$$\sum_{x_1^n} 2^{-L_n(x_1^n)} < 1.$$

Show that there are arbitrarily long binary sequences that cannot be decoded as having been produced by repeated uses of C_n.

Exercise 12.45 (Optimal codes that require one bit above the entropy)**.** Let X be a discrete RV with entropy H. We saw in Theorem 9.6 that there is always a prefix free code (C, L) with expected description length $\mathbb{E}[L(X)]$ that is strictly less than $H + 1$ bit. For any $\epsilon > 0$, give an example of a random variable X for which $\mathbb{E}[L(X)] > H + 1 - \epsilon$ for all prefix-free codes (C, L).

Exercise 12.46 (Functional submodularity of entropy)**.** A number of sumset bounds for the entropy were established by Tao in 2010 [20], using the following property of the entropy. Suppose X_1, X_2 are two arbitrary discrete RVs, and suppose that the RVs X_0 and X_{12} can be expressed as $X_0 = F(X_1) = G(X_2)$ and $X_{12} = R(X_1, X_2)$, for some functions F, G and R. Show that:

$$H(X_{12}) + H(X_0) \leq H(X_1) + H(X_2).$$

Exercise 12.47 (Ruzsa covering lemma for entropy)**.** Suppose X, Y are independent discrete RVs, and let (X_1, Y_1), (X_2, Y_2) be conditionally independent versions of $[(X, Y)$ conditional on $X + Y]$. Show that:

$$H(X_1, X_2, Y_1|Y_2) = 2H(X) + H(Y) - H(X + Y).$$

Exercise 12.48 (Ruzsa triangle for dependent RVs)**.** Show that (64) remains valid if Y is independent of (X, Z), but X and Z are not necessarily independent.

Exercise 12.49 (Properties of mutual information for continuous RVs)**.** Conditional entropy and mutual information between two continuous random variables X and Y are defined in the same way as in the discrete case:

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

(*i*) Verify that the basic properties of mutual information (Propositions 5.2 and 5.3) also hold in the continuous setting. Show also that the analogous bound to Proposition 5.4 (a) remains true. Does the analogous inequality to (b) hold in the continuous setting?

(*ii*) Prove that the analogous lower bound of Proposition 10.1 remains true for continuous RVs. Does the upper bound hold?

Exercise 12.50 (Stam's inequality)**.** Optimize in a to complete the proof of Lemma 11.4.

Exercise 12.51 (EPI equivalence)**.** Show that the two formulations of the EPI are equivalent to each other and to the following formulation:

$$h(X + Y) \geq h(X^* + Y^*)$$

where X^* and Y^* are independent Gaussians with the same entropies as X and Y respectively. *Hint.* Take $\lambda = \frac{N(X)}{N(X)+N(Y)}$.

Exercise 12.52 (Equality in the EPI)**.** Let $X \sim \mathcal{N}(\mu_1, \sigma_1^2), Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Verify that there is equality in (90). What does equality in (91) imply for σ_1 and σ_2?

Remark. *By examining the equality case in the application of Cauchy-Schwarz in Stam's lemma, it is possible to show that equality in the EPI forces the score functions to be linear. Thus, there is equality in the EPI* if and only if *X and Y are Gaussian.*

References

[1] P.H. Algoet and T.M. Cover. A sandwich proof of the Shannon-Mcmillan-Breiman theorem. *Ann. Probab.*, 16(2):876–898, April 1988.

[2] A.D. Barbour, L. Holst, and S. Janson. *Poisson approximation*. Oxford University Press, Oxford, U.K., 1992.

[3] A.R. Barron. Monotonic central limit theorem for densities. NSF Technical Report no. 50, Department of Statistics, Stanford University, March 1984. Available at `statistics.stanford.edu/resources/technical-reports`.

[4] A.R. Barron. Entropy and the central limit theorem. *Ann. Probab.*, 14(1):336–342, January 1986.

[5] P. Billingsley. *Probability and measure*. John Wiley & Sons, New York, NY, third edition, 1995.

[6] T.M. Cover and J.A. Thomas. *Elements of information theory*. John Wiley & Sons, New York, NY, second edition, 2006.

[7] I. Csiszár and J. Körner. *Information theory: Coding theorems for discrete memoryless systems*. Cambridge University Press, Cambridge, U.K., second edition, 2011.

[8] I. Csiszár and P. Shields. Information theory and statistics: A tutorial. *Foundations and Trends in Communications and Information Theory*, 1(4):417–528, December 2004.

[9] W. Feller. *An introduction to probability theory and its applications. Vol. I*. John Wiley & Sons, New York, NY, 1950.

[10] W.T. Gowers, B. Green, F. Manners, and T. Tao. Marton's conjecture in abelian groups with bounded torsion. *arXiv e-prints*, `2404.02244 [math.NT]`, April 2024.

[11] W.T. Gowers, B. Green, F. Manners, and T. Tao. On a conjecture of Marton. *Ann. of Math.*, 201(2):515–549, March 2025.

[12] B. Green, F. Manners, and T. Tao. Sumsets and entropy revisited. *Random Struct. Algorithms*, 66(1):e21252, 2025.

[13] P. Harremoës. Binomial and Poisson distributions as maximum entropy distributions. *IEEE Trans. Inform. Theory*, 47(5):2039–2041, July 2001.

[14] I. Kontoyiannis, P. Harremoës, and O. Johnson. Entropy and the law of small numbers. *IEEE Trans. Inform. Theory*, 51(2):466–472, February 2005.

[15] I. Kontoyiannis and B. Lucena. Mutual information, synergy and some curious phenomena for simple channels. In *2005 IEEE International Symposium on Information Theory (ISIT)*, pages 1651–1655, Adelaide, Australia, September 2005.

[16] I. Kontoyiannis and M. Madiman. Sumset and inverse sumset inequalities for differential entropy and mutual information. *IEEE Trans. Inform. Theory*, 60(8):4503–4514, August 2014.

[17] D.J.C. MacKay. *Information theory, inference and learning algorithms*. Cambridge University Press, New York, NY, 2003. Available at `www.inference.org.uk/itprnn`.

[18] I.Z. Ruzsa. Sums of finite sets. In G.V. Chudnovsky D.V. Chudnovsky and M.B. Nathanson, editors, *Number Theory: New York Seminar*. Springer-Verlag, 1996.

[19] P.C. Shields. *The ergodic theory of discrete sample paths*, volume 13 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 1996.

[20] T. Tao. Sumset and inverse sumset theory for Shannon entropy. *Comb., Probab. Comput.*, 19(4):603–639, July 2010.

[21] D. Williams. *Probability with martingales*. Cambridge University Press, Cambridge, U.K., 1991.