Contents

1. Probability basics

TODO: weak and strong laws of large numbers, Markov chains, Cesaro lemma, … probably others.

2. Entropy

2.1. Introduction

Notation 2.1 Write $x_1^n := (x_1, ..., x_n) \in \{0, 1\}^n$ for an length n bit string.

Notation 2.2 We use P to denote a probability mass function. Write P_1^n for the joint proability mass function of a sequence of n random variables $X_1^n = (X_1, ..., X_n)$.

Definition 2.3 A random variable X has a **Bernoulli distribution**, $X \sim \text{Bern}(p)$, if for some fixed $p \in (0, 1)$,

$$X = \begin{cases} 1 \text{ with probability } p \\ 0 \text{ with probability } 1 - p \end{cases}$$

i.e. the probability mass function (PMF) of X is $P : \{0, 1\} \to \mathbb{R}$, $P(0) = 1 - p$, $P(1) = p$.

Notation 2.4 Throughout, we take log to be the base-2 logarithm, $\log_2$.

Definition 2.5 The **binary entropy function** $h : (0, 1) \to [0, 1]$ is defined as

$$h(p) := -p \log p - (1 - p) \log(1 - p)$$

Example 2.6 Let $x_1^n \in \{0, 1\}^n$ be an n bit string which is the realisation of binary random variables (RVs) $X_1^n = (X_1, ..., X_n)$, where the X_i are independent and identically distributed (IID), with common distribution $X_i \sim \text{Bern}(p)$. Let $k = |\{i \in [n] : x_i = 1\}|$ be the number of ones in x_1^n. We have

$$\Pr(X_1^n = x_1^n) := P^n(x_1^n) = \prod_{i=1}^{n} P(x_i) = p^k (1 - p)^{n-k}.$$

Now by the law of large numbers, the probability of ones in a random x_1^n is $k/n \approx p$ with high probability for large n. Hence,

$$P^n(x_1^n) \approx p^{np} (1 - p)^{n(1-p)} = 2^{-nh(p)}.$$

Note that this reveals an amazing fact: this approximation is independent of x_1^n, so any message we are likely to encounter has roughly the same probability $\approx 2^{-nh(p)}$ of occurring.

Remark 2.7 By the above example, we can split the set of all possible n-bit messages, $\{0, 1\}^n$, into two parts: the set B_n of **typical** messages which are approximately uniformly distributed with probability $\approx 2^{-nh(p)}$ each, and the non-typical messages that occur with negligible probability. Since all but a very small amount of the probability is concentrated in B_n, we have $|B_n| \approx 2^{nh(p)}$.

Remark 2.8 Suppose an encoder and decoder both already know B_n and agree on an ordering of its elements: $B_n = \{x_1^n(1), ..., x_1^n(b)\}$, where $b = |B_n|$. Then instead of transmitting the actual message, the encoder can transmit its index $j \in [b]$, which can be described with

$$\lceil \log b \rceil = \lceil \log |B_n| \rceil \approx nh(p)$$

bits.

Remark 2.9

- The closer p is to $\frac{1}{2}$ (intuitively, the more random the messages are), the larger the entropy $h(p)$, and the larger the number of typical strings $|B_n|$.
- Assuing we ignore non-typical strings, which have vanishingly small probability for large n, the "compression rate" of the above method is $h(p)$, since we encode n bit strings using $nh(p)$ strings. $h(p) < 1$ unless the message is uniformly distributed over all of $\{0, 1\}^n$.
- So the closer p is to 0 or 1 (intuitively, the less random the messages are), the smaller the entropy $h(p)$, so the greater the compression rate we can achieve.

2.2. Asymptotic equipartition property

Notation 2.10 We denote a finite alphabet by $A = \{a_1, ..., a_m\}$.

Notation 2.11 If $X_1, ..., X_n$ are IID RVs with values in A, with common distribution described by a PMF $P : A \to [0, 1]$ (i.e. $P(x) = \Pr(X_i = x)$ for all $x \in A$), then write $X \sim P$, and we say "X has distribution P on A".

Notation 2.12 For $i \leq j$, write X_i^j for the block of random variables $\left(X_i, ..., X_j\right)$, and similarly write x_i^j for the length $j - i + 1$ string $\left(x_i, ..., x_j\right) \in A^{i-j+1}$.

Notation 2.13 For IID RVs $X_1, ..., X_n$ with each $X_i \sim P$, denote their joint PMF by $P^n : A^n \to [0, 1]$:

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

and we say that "the RVs X_1^n have the product distribution P^n".

Definition 2.14 A sequence of RVs $(Y_n)_{n \in \mathbb{N}}$ **converges in probability** to an RV Y if $\forall \varepsilon > 0$,

$$\Pr(|Y_n - Y| > \varepsilon) \to 0 \quad \text{as } n \to \infty.$$

Definition 2.15 Let $X \sim P$ be a discrete RV on a countable alphabet A. The **entropy** of X is

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

Remark 2.16

- We use the convention $0 \log 0 = 0$ (this is natural due to continuity: $x \log x \to 0$ as $x \downarrow 0$, and also can be derived measure-theoretically).

- Entropy is technically a functional the probability distribution P and not of X, but we use the notation $H(X)$ as well as $H(P)$.
- $H(X)$ only depends on the probabilities $P(x)$, not on the values $x \in A$. Hence for any bijective $f : A \to A$, we have $H(f(X)) = H(X)$.
- All summands of $H(X)$ are non-negative, so the sum always exists and is in $[0, \infty]$, even if A is countable infinite.
- $H(X) = 0$ iff all summands are 0, i.e. if $P(x) \in \{0, 1\}$ for all $x \in A$, i.e. X is **deterministic** (constant, so equal to a fixed $x_0 \in A$ with probability 1).

Theorem 2.17 Let $X = \{X_n : n \in \mathbb{N}\}$ be IID RVs with common distribution P on a finite alphabet A. Then

$$-\frac{1}{n} \log P^n(X_1^n) \longrightarrow H(X_1) \quad \text{in probability} \quad \text{as } n \to \infty$$

Proof (Hints). Straightforward. □

Proof. We have

$$\begin{aligned} P^n(X_1^n) &= \prod_{i=1}^n P(X_i) \\ \implies \frac{1}{n} \log P^n(X_1^n) &= \frac{1}{n} \sum_{i=1}^n \log P(X_i) \to \mathbb{E}[-\log P(X_1)] \quad \text{in probability} \end{aligned}$$

by the weak law of large numbers (WLLN) for the IID RVs $Y_i = -\log P(X_i)$. □

Corollary 2.18 (Asymptotic Equipartition Property (AEP)) Let $\{X_n : n \in \mathbb{N}\}$ be IID RVs on a finite alphabet A with common distribution P and common entropy $H = H(X_i)$. Then

- ($\Longrightarrow$): for all $\varepsilon > 0$, the set of **typical strings** $B_n^*(\varepsilon) \subseteq A^n$ defined by

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

  satisfies

  $$\begin{aligned} |B_n^*(\varepsilon)| &\leq 2^{n(H+\varepsilon)} \quad \forall n \in \mathbb{N}, \quad \text{and} \\ P^n(B_n^*(\varepsilon)) = \Pr(X_1^n \in B_n^*(\varepsilon)) &\longrightarrow 1 \quad \text{as } n \to \infty \end{aligned}$$

- ($\Longleftarrow$): for any sequence $(B_n)_{n \in \mathbb{N}}$ of subsets of A^n, if $P(X_1^n \in B_n) \to 1$ as $n \to \infty$, then $\forall \varepsilon > 0$,

  $$\begin{aligned} |B_n| &\geq (1 - \varepsilon) 2^{n(H-\varepsilon)} \quad \text{eventually} \\ \text{i.e. } \exists N \in \mathbb{N} : \forall n \geq N, \quad |B_n| &\geq (1 - \varepsilon) 2^{n(H-\varepsilon)}. \end{aligned}$$

Proof (Hints).

- ($\Longrightarrow$): straightforward.
- ($\Longleftarrow$): show that $P^n(B_n \cap B_n^*(\varepsilon)) \to 1$ as $n \to \infty$.

□

Proof.

- ($\Longrightarrow$):
  - Let $\varepsilon > 0$. By Theorem 2.17, we have

$$\Pr(X_1^n \notin B_n^*(\varepsilon)) = \Pr\left(\left|-\frac{1}{n}\log P^n(X_1^n) - H\right| > \varepsilon\right) \to 0 \quad \text{as } n \to \infty.$$

  - By definition of $B_n^*(\varepsilon)$,

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

- ($\Longleftarrow$):
  - We have $P^n(B_n \cap B_n^*(\varepsilon)) = P^n(B_n) + P^n(B_n^*(\varepsilon)) - P^n(B_n \cup B_n^*(\varepsilon)) \geq P^n(B_n) + P^n(B_n^*(\varepsilon)) - 1$, so $P^n(B_n \cap B_n^*(\varepsilon)) \to 1$.
  - So $P^n(B_n \cap B_n^*(\varepsilon)) \geq 1 - \varepsilon$ eventually, and so

$$\begin{aligned} 1 - \varepsilon \leq P^n(B_n \cap B_n^*(\varepsilon)) &= \sum_{x_1^n \in B_n \cap B_n^*(\varepsilon)} P^n(x_1^n) \\ &\leq |B_n \cap B_n^*(\varepsilon)| 2^{-n(H-\varepsilon)} \leq |B_n| 2^{-n(H-\varepsilon)}. \end{aligned}$$

□

Remark 2.19

- The $\Longrightarrow$ part of AEP states that a specific object (in this case, the $B_n^*(\varepsilon)$) can achieve a certain performance, while the $\Longleftarrow$ part states that no other object of this type can significantly perform better. This is common type of result in information theory.
- Theorem 2.17 gives a mathematical interpretation of entropy: the probability of a random string X_1^n generally decays exponentially with n ($P^n(X_1^n) \approx 2^{-nH}$ with high probability for large n). The AEP gives a more "operational interpretation": the smallest set of strings that can carry almost all the probability of P^n has size $\approx 2^{nH}$.
- The AEP tells us that higher entropy means more typical strings, and so the possible values of X_1^n are more unpredictable. So we consider "high entropy" RVs to be "more random" and "less predictable".

2.3. Fixed-rate lossless data compression

Definition 2.20 A **memoryless source** $X = \{X_n : n \in \mathbb{N}\}$ is a sequence of IID RVs with a common PMF P on the same alphabet A.

Definition 2.21 A **fixed-rate lossless compression code** for a source X consists of a sequence of **codebooks** $\{B_n : n \in \mathbb{N}\}$, where each $B_n \subseteq A^n$ is a set of source strings of length n.

Assume the encoder and decoder share the codebooks, each of which is sorted. To send x_1^n, an encoder checks with $x_1^n \in B_n$; if so, they send the index of x_1^n in B_n, along with a flag bit 1, which requires $1 + \lceil \log |B_n| \rceil$ bits. Otherwise, they send x_1^n

uncompressed, along with a flag bit 0 to indicate an "error", which requires $1 + \lceil \log|A|\rceil = 1 + \lceil n \log|A|\rceil$ bits.

Definition 2.22 For each $n \in \mathbb{N}$, the **rate** of a fixed-rate code $\{B_n : n \in \mathbb{N}\}$ for a source X is

$$R_n := \frac{1}{n}(1 + \lceil \log|B_n|\rceil) \approx \frac{1}{n}\log|B_n| \quad \text{bits/symbol.}$$

Definition 2.23 For each $n \in \mathbb{N}$, the **error probability** of a fixed-rate code $\{B_n : n \in \mathbb{N}\}$ for a source X is

$$P_e^{(n)} := \Pr(X_1^n \notin B_n).$$

Theorem 2.24 (Fixed-rate coding theorem) Let $X = \{X_n : n \in \mathbb{N}\}$ be a memoryless source with distribution P and entropy $H = H(X_i)$.

- ($\Longrightarrow$): $\forall \varepsilon > 0$, there is a fixed-rate code $\{B_n^*(\varepsilon) : n \in \mathbb{N}\}$ with vanishing error probability ($P_e^{(n)} \to 0$ as $n \to \infty$) and with rate

$$R_n \leq H + \varepsilon + \frac{2}{n} \quad \forall n \in \mathbb{N}.$$

- ($\Longleftarrow$): let $\{B_n : n \in \mathbb{N}\}$ be a fixed-rate with vanishing error probabilit. Then $\forall \varepsilon > 0$, its rate R_n satisfies

$$R_n > H - \varepsilon \quad \text{eventually.}$$

Proof (Hints). ($\Longrightarrow$): straightforward. ($\Longleftarrow$): straightforward. □

Proof.

- ($\Longrightarrow$):
  - Let $B_n^*(\varepsilon)$ be the sets of typical strings defined in AEP (Asymptotic Equipartition Property (AEP)). Then $P_e^{(n)} = 1 - \Pr(X_1^n \in B_n^*) \to 0$ as $n \to \infty$ by AEP.
  - Also by AEP, $R_n = \frac{1}{n}(1 + \lceil \log|B_n^*|\rceil) \leq \frac{1}{n}\log|B_n^*| + \frac{2}{n} \leq H + \varepsilon + \frac{2}{n}$.
- ($\Longleftarrow$):
  - WLOG let $0 < \varepsilon < 1/2$. By AEP,

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

eventually.

□

3. Relative entropy

Definition 3.1 Suppose $x_1^n \in A^n$ are observations generated by IID RVs X_1^n and we want to decide whether $X_1^n \sim P^n$ or Q^n, for two distinct candidate PMFs P, Q on A. A **hypothesis test** is described by a **decision region** $B_n \subseteq A^n$ such that

- If $x_1^n \in B_n$, then we declare that $X_1^n \sim P^n$.
- Otherwise, if $x_1^n \notin B_n$, then we declare that $X_1^n \sim Q^n$.

Definition 3.2 The associated **error probabilities** for a hypothesis test are

$$\begin{aligned} e_1^{(n)} &= e_1^{(n)}(B_n) := \Pr(\text{declare } P \mid \text{data} \sim Q) = Q^n(B_n) \\ e_2^{(n)} &= e_2^{(n)}(B_n) := \Pr(\text{declare } Q \mid \text{data} \sim P) = P^n(B_n^c). \end{aligned}$$

Definition 3.3 The **relative entropy** between PMFs P and Q on the same countable alphabet A is

$$D(P \parallel Q) := \sum_{x \in A} P(x) \log \frac{P(x)}{Q(x)} = \mathbb{E}\left[\log \frac{P(X)}{Q(X)}\right], \quad \text{where } X \sim P.$$

Remark 3.4

- We use the convention that $0 \log \frac{0}{0} = 0$ (this can be avoided by defining relative entropy measure-theoretically).
- $D(P \parallel Q)$ always exists and $D(P \parallel Q) \geq 0$ with equality iff $P = Q$.
- Relative entropy is not symmetric: $D(P \parallel Q) \neq D(Q \parallel P)$ in general, and does not satisfy the triangle inequality.
- Despite this, it is reasonable and natural to think of $D(P \parallel Q)$ as a statistical "distance" between P and Q.

Remark 3.5 Let $X \sim P$. We have, by WLLN,

$$\begin{aligned} \frac{1}{n} \log\left(\frac{P^n(X_1^n)}{Q^n(X_1^n)}\right) &= \frac{1}{n} \log \prod_{i=1}^{n} \frac{P(X_i)}{Q(X_i)} \\ &= \frac{1}{n} \sum_{i=1}^{n} \log \frac{P(X_i)}{Q(X_i)} \\ &\longrightarrow D(P \parallel Q) \text{ in probability} \quad \text{as } n \to \infty. \end{aligned}$$

So for large n, $\frac{P^n(X_1^n)}{Q^n(X_1^n)} \approx 2^{nD(P \parallel Q)}$ with high probability. Hence, the random string X_1^n is exponentially more likely under its true distribution P than under Q.

3.1. Asymptotically optimal hypothesis testing

Theorem 3.6 (Stein's Lemma) Let P, Q be PMFs on a finite alphabet A, with $D = D(P \parallel Q) \in (0, \infty)$. Let $X = \{X_n : n \in \mathbb{N}\}$ be a memoryless source on A, with either each $X_i \sim P$ or each $X_i \sim Q$.

- ($\Longrightarrow$): for all $\varepsilon > 0$, there is a hypothesis test with decision regions $\{B_n^*(\varepsilon) : n \in \mathbb{N}\}$ such that

$$\forall n \in \mathbb{N}, \quad e_1^{(n)}(B_n^*(\varepsilon)) \leq 2^{-n(D-\varepsilon)}$$

  and $e_2^{(n)} \to 0$ as $n \to \infty$.
- ($\Longleftarrow$): for any hypothesis test with decision regions $\{B_n : n \in \mathbb{N}\}$ such that $e_2^{(n)}(B_n) \to 0$ as $n \to \infty$, we have $\forall \varepsilon > 0$,

$$e_1^{(n)}(B_n) \geq 2^{-n(D+\varepsilon+\frac{1}{n})} \quad \text{eventually}.$$

Proof (Hints).

- ($\Longrightarrow$):
  - Let $B_n^*(\varepsilon) = \left\{x_1^n \in A^n : 2^{n(D-\varepsilon)} \leq \frac{P^n(x_1^n)}{Q^n(x_1^n)} \leq 2^{n(D+\varepsilon)}\right\}$. The rest is straightforward (use above remark).
- ($\Longleftarrow$):
  - Show that $P^n(B_n^*(\varepsilon) \cap B_n) \to 1$ as $n \to \infty$, use that $\frac{1}{2} = 2^{-n(1/n)}$.

□

Proof.

- ($\Longrightarrow$):
  - Let $B_n^*(\varepsilon) = \left\{x_1^n \in A^n : 2^{n(D-\varepsilon)} \leq \frac{P^n(x_1^n)}{Q^n(x_1^n)} \leq 2^{n(D+\varepsilon)}\right\}$.
  - Then the convergence in probability of $\frac{1}{n}\sum_{i=1}^n \log \frac{P(X_i)}{Q(X_i)}$ is equivalent to $\Pr(X_1^n \notin B_n^*) = P^n(B_n^*(\varepsilon)) = e_2^{(n)} \to 0$ as $n \to \infty$, when $X_1^n \sim P^n$.
  - Also, $1 \geq P^n(B_n^*) = \sum_{x_1^n \in B_n^*(\varepsilon)} Q^n(x_1^n)\frac{P^n(x_1^n)}{Q^n(x_1^n)} \geq 2^{n(D-\varepsilon)} \sum_{x_1^n \in B_n^*(\varepsilon)} Q^n(x_1^n) = 2^{n(D-\varepsilon)} Q^n(B_n^*(\varepsilon))$.
- ($\Longleftarrow$):
  - We havee $e_2^{(n)}(B_n^*(\varepsilon)) = P^n(B_n^*(\varepsilon)) \to 0$ as $n \to \infty$. Suppose $e_2^{(n)}(B_n) = P^n(B_n^c) \to 0$. Then $P^n(B_n \cap B_n^*(\varepsilon)) \to 1$. So eventually,

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

□

Remark 3.7

- The decision regions B_n^* are asymptotically optimal in that, among all tests that have $e_2^{(n)} \to 0$, they achieve the asymptotically smallest possible $e_1^{(n)} \approx 2^{-nD}$. However, they are not the most optimal decision regions for finite n. For finite regions, the optimal regions are given by the Neyman-Pearson Lemma.
- Assuming $D \neq 0$ is a trivial assumption, as otherwise $P = Q$ on A, so any test would give the correct answer.
- Assuming $D < \infty$ is a reasonable assumption, as otherwise there is some $a \in A$ such that $P(a) > 0$ but $Q(a) = 0$. In that case, we check whether any such a appear in x_1^n or not.
- In Stein's Lemma, we assume one error vanishes at possibly an arbitrarily slow rate, while the other decays exponentially. This is a natural asymmetry in many applications, e.g. in diagnosing disease.
- Stein's Lemma shows why the relative entropy is a natural measure of "distance" between two distributions, as large D means a smaller error probability (one vanishes exponentially at rate D), so easier to tell apart the distributions from the data.

3.2. Relative entropy and optimal hypothesis testing

Theorem 3.8 (Neyman-Pearson Lemma) For a hypothesis test between P and Q based on n data samples, the **likelihood ratio decision regions**

$$B_{\mathrm{NP}} = \left\{ x_1^n \in A^n : \frac{P^n(x_1^n)}{Q^n(x_1^n)} \geq T \right\}, \quad \text{for some threshold } T > 0,$$

are optimal in that, for any decision region $B_n \subseteq A^n$, if $e_1^{(n)}(B_n) \leq e_1^{(n)}(B_{\mathrm{NP}})$, then $e_2^{(n)}(B_n) \geq e_2^{(n)}(B_{\mathrm{NP}})$, and vice versa.

Proof (Hints). Consider the inequality

$$(P^n(x_1^n) - TQ^n(x_1^n))\big(\mathbb{1}_{B_{\mathrm{NP}}}(x_1^n) - \mathbb{1}_{B_n}(x_1^n)\big) \geq 0$$

(justify why this holds). □

Proof.

- Consider the obvious inequality

$$(P^n(x_1^n) - TQ^n(x_1^n))\big(\mathbb{1}_{B_{\mathrm{NP}}}(x_1^n) - \mathbb{1}_{B_n}(x_1^n)\big) \geq 0$$

- Then, summing over all x_1^n,

$$\begin{aligned} 0 &\leq P^n(B_{\mathrm{NP}}) - P^n(B_n) - TQ^n(B_{\mathrm{NP}}) + TQ^n(B_n) \\ &= 1 - e_2^{(n)}(B_{\mathrm{NP}}) - \Big(1 - e_2^{(n)}(B_n)\Big) - T\Big(e_1^{(n)}(B_{\mathrm{NP}}) - e_1^{(n)}(B_n)\Big) \\ &\implies e_2^{(n)}(B_n) - e_2^{(n)}(B_{\mathrm{NP}}) \geq T\Big(e_1^{(n)}(B_{\mathrm{NP}}) - e_1^{(n)}(B_n)\Big) \end{aligned}$$

□

Remark 3.9 Neyman-Pearson says that if any decision region has an error as small as that of B_{NP}, then its other error must be larger than that of B_{NP}.

Notation 3.10 Let $\hat{P}_n$ denote the empirical distribution (or **type**) induced by x_1^n on A^n (the frequency with which $a \in A$ occurs in x_1^n):

$$\forall a \in A, \quad \hat{P}_n(a) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{x_i = a\}}$$

Proposition 3.11 The Neyman-Pearson decision region B_{NP} can be expressed in information-theoretic form as

$$B_{\mathrm{NP}} = \Big\{ x_1^n \in A^n : D\Big(\hat{P}_n \parallel Q\Big) \geq D\Big(\hat{P}_n \parallel P\Big) + T' \Big\}$$

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

Proof (Hints). Rewrite the expression $\frac{1}{n}\log\frac{P^n(x_1^n)}{Q^n(x_1^n)}$. □

Proof. We have

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

□

Theorem 3.12 (Jensen's Inequality) Let I be an interval, $f : I \to \mathbb{R}$ be convex and X be an RV with values in I. Then

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

Moreover, if f is strictly convex, then equality holds iff X is almost surely constant.

Theorem 3.13 (Log-sum Inequality) Let $a_1, ..., a_n$, $b_1, ..., b_n$ be non-negative constants. Then

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

with equality iff $\frac{a_i}{b_i} = c$ for all i, for some constant c. We use the convention that $0\log 0 = 0\log\frac{0}{0} = 0$.

Remark 3.14 This also holds for countably many a_i and b_i.

Proof (Hints). Use Jensen's inequality with X the RV such that $\Pr\left(X = \frac{a_i}{b_i}\right) = \frac{b_i}{\sum_{j=1}^n b_j}$ for all $i \in [n]$, and a suitable f. □

Proof.

- Define

$$f(x) = \begin{cases} x\log x & \text{if } x > 0\\ 0 & \text{otherwise}\end{cases}.$$

  f is strictly convex.
- Let $A = \sum_i a_i$, $B = \sum_i b_i$. Let X be the RV with $\Pr\left(X = \frac{a_i}{b_i}\right) = \frac{b_i}{B}$ for all $i \in [n]$.
- Then $\mathbb{E}[f(X)] = \sum_i \frac{b_i}{B}\frac{a_i}{b_i}\log\frac{a_i}{b_i} = \frac{1}{B}\sum_i a_i \log\frac{a_i}{b_i}$.
- $f(\mathbb{E}[X]) = \mathbb{E}[X]\log\mathbb{E}[X] = \sum_i \frac{a_i}{b_i}\frac{b_i}{B}\log\sum_i \frac{a_i}{b_i}\frac{b_i}{B} = \frac{A}{B}\log\frac{A}{B}$.

- So by Jensen's inequality, $\frac{A}{B}\log\frac{A}{B} \le \frac{1}{B}\sum_i a_i \log\frac{a_i}{b_i}$.

□

Proposition 3.15

1. If P and Q are PMFs on the same finite alphabet A, then
$$D(P \parallel Q) \ge 0$$
with equality iff $P = Q$.
2. If $X \sim P$ on a finite alphabet A, then
$$0 \le H(X) \le \log|A|$$
with equality to 0 iff X is a constant, and equality to $\log|A|$ iff X is uniformly distributed on A.

Remark 3.16 This also holds for countably infinite A.

Proof (Hints).

1. Straightforward.
2. For $\le \log|A|$, consider $D(P \parallel Q)$ where Q is the uniform distribution on A. ≥ 0 is straightforward.

□

Proof.

- ‣ By the log-sum inequality,
$$D(P \parallel Q) = \sum_{x\in A} P(x)\log\frac{P(x)}{Q(x)} \ge \left(\sum_{x\in A} P(x)\right)\log\frac{\sum_{x\in A} P(x)}{\sum_{x\in A} Q(x)} = 0$$
with equality if $\frac{P(x)}{Q(x)}$ is the same constant for all $x \in A$, i.e. $P = Q$.
- ‣ Let Q be the uniform distribution on A, so $H(Q) = \sum_{x\in A}\frac{1}{|A|}\log\frac{1}{1/|A|} = \log|A|$.
  ‣ Now $0 \le D(P \parallel Q) = \sum_{x\in A} P(x)\log\frac{P(x)}{1/|A|} = \log|A| - H(X)$ with equality iff $P = Q$, i.e. P is uniform.
  ‣ Each term in $-H(X)$ is ≤ 0, with equality iff each $P(x)\log P(x)$ is 0, i.e. $P(x) = 0$ or 1.

□

Remark 3.17 If $X = \{X_n : n \in \mathbb{N}\}$ is a memoryless source with PMF P on A, then we have shown that it can be at best compressed to $\approx H(P)$ bits/symbol. This means that we can always achieve non-trivial compression, i.e. a description using $\approx H(P) < \log|A|$ bits/symbol, unless the source X is completely random (i.e. IID and uniformly distribute), in which case we cannot do better than simply describing each x_1^n uncompressed using $\frac{\lceil\log|A^n|\rceil}{n} \approx \log|A|$ bits/symbol.

4. Properties of entropy and relative entropy

4.1. Joint entropy and conditional entropy

Definition 4.1 Let X_1^n be an arbitrary finite collection of discrete RVs on corresponding alphabets $A_1, ..., A_n$. Note we can think of X_1^n itself a discrete RV on alphabet $A_1 \times \cdots \times A_n$. Let X_1^n have PMF P_n, then the **joint entropy** of X_1^n is

$$H(X_1^n) = H(P_n) = H(X_1, ..., X_n) := \mathbb{E}[-\log P_n(X_1^n)] = -\sum_{x_1^n \in A^n} P_n(x_1^n) \log P_n(x_1^n).$$

Example 4.2 Note that if X and Y are independent, then $P_{X,Y}(x,y) = P_X(x)P_Y(y)$, so

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

Example 4.3 Let X and Y have joint PMF given by

$Y \backslash X$	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	

Note that X and Y are not independent. We have

$$\begin{aligned} H(X) &= -\frac{3}{10}\log\frac{3}{10} - \frac{1}{4}\log\frac{1}{4} - \frac{9}{20}\log\frac{9}{20} \approx 1.539, \\ H(Y) &= -\frac{11}{20}\log\frac{11}{20} - \frac{9}{20}\log\frac{9}{20} \approx 0.993, \\ H(X,Y) &= -\frac{1}{10}\log\frac{1}{10} - \cdots - \frac{1}{5}\log\frac{1}{5} \approx 2.441 < H(X) + H(Y). \end{aligned}$$

In general, if X and Y are not independent, then $P_{XY}(x,y) = P_X(x)P_{Y \mid X}(y \mid x)$, so

$$H(X,Y) = \mathbb{E}[-\log P_{XY}(x,y)] = \mathbb{E}[-\log P_X(x)] + \mathbb{E}\left[-\log P_{Y \mid X}(y \mid x)\right].$$

Definition 4.4 Let X and Y be discrete random variables with joint PMF $P_{X,Y}$, then the **conditional entropy** of Y given X is

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

Note 4.5 $P_{Y \mid X}$ is a function of $(x,y) \in X$, and so for the expected value we multiply the log by the probability that $X = x$ and $Y = y$.

Proposition 4.6 For discrete RVs X and Y, we have

$$H(Y \mid X) = H(X,Y) - H(X).$$

Proof (Hints). Straightforward. □

Proof. Note that $P_{Y \mid X}(y \mid x) = \Pr(Y = y \mid X = x) = \frac{\mathbb{P}(Y=y, X=x)}{\mathbb{P}(X=x)} = P_{X,Y}(x, y) P_X(x)$. Hence

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

□

4.2. Properties of entropy, joint entropy and conditional entropy

Proposition 4.7 (Chain Rule for Entropy) Let X_1^n be a collection of discrete RVs. Then

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

In particular, if the X_1^n are independent, then

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

Proof (Hints). By induction. □

Proof. We can write

$$\begin{aligned} P_{X_1^n}(x_1^n) &= P_{X_1}(x_1) P_{X_2|X_1}(x_2 \mid x_1) \cdots P_{X_n \mid X_1, \ldots, x_{n-1}}(x_n \mid x_1, \ldots, x_{n-1}) \\ &= \prod_{i=1}^{n} P_{X_i \mid X_1^{i-1}}(x_i \mid x_1^{i-1}). \end{aligned}$$

Then the result follows by inductively using the above proposition. □

Proposition 4.8 (Conditioning Reduces Entropy) For discrete RVs X and Y,

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

with equality iff X and Y are independent.

Proof (Hints). Express $H(Y) - H(Y \mid X)$ as a relative entropy. □

Proof. We have

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

This is non-negative iff $P_{X,Y} = P_X P_Y$, i.e. X and Y are independent. □

Definition 4.9 Discrete RVs X and Z are **conditionally independent given Y** if:

- $P_{X,Z \mid Y}(x, z \mid y) = P_{X \mid Y}(x \mid y) P_{Z \mid Y}(z \mid y)$,
- or equivalently, $P_{X \mid Z,Y}(x \mid z, y) = P_{X \mid Y}(x \mid y)$,
- or equivalently, $P_{Z \mid X,Y}(z \mid x, y) = P_{Z \mid Y}(z \mid y)$.

We denote this by writing $X - Y - Z$ and we say that X, Y, Z form a Markov chain. Note that $X - Y - Z$ is equivalent to $Z - Y - X$, but not to $X - Z - Y$.

Note 4.10 For any function g on Y, we have $X - Y - g(Y)$.

Corollary 4.11 $H(X_1^n) \le \sum_{i=1}^n H(X_i)$ with equality iff all X_1^n are independent.

Proof. Straightforward. □

Proof. $H(X_1^n) = \sum_{i=1}^n H(X_i \mid X_1^{i-1}) \le \sum_{i=1}^n H(X_i)$ by the chain rule and conditioning reducing entropy. □

Remark 4.12 We can write

$$
\begin{aligned}
H(Y \mid X) &= -\sum_{x,y} \big(P_{X,Y}(x, y)\big) \log P_{Y \mid X}(y \mid x) \\
&= \sum_x P_X(x) \left(-\sum_y P_{Y \mid X}(y \mid x) \log P_{Y \mid X}(y \mid x)\right) \\
&=: \sum_x P_X(x) H(Y \mid X = x)
\end{aligned}
$$

Note $H(Y \mid X = x)$ is **not** a conditional entropy, and in particular, we do not always have $H(Y \mid X = x) \le H(Y)$. Since $0 \le H(Y \mid X = x) \le \log|A_Y|$, we have $0 \le H(Y \mid X) \le \log|A_Y|$ with equality to 0 iff Y is a function of X (i.e. $H(Y \mid X = x) = 0$ for all x).

Proposition 4.13 (Data Processing Inequality for Entropy) Let X be discrete RV on alphabet A and f be function on A. Then

1. $H(f(X)|X) = 0$.
2. $H(f(X)) \le H(X)$ with equality iff f is injective.

Proof (Hints). Use that $x \mapsto (x, f(x))$ is injective and the chain rule. □

Proof. We have already shown the "if" direction of 2. We have $H(X) = H(X, f(X)) = H(f(X)|X) + H(X)$, since $x \mapsto (x, f(x))$ is injective. Also, $H(X) = H(X, f(X)) = H(X \mid f(X)) + H(f(X)) \geq H(f(X))$. So $H(X) \geq H(f(X))$ with equality iff $H(X \mid f(X)) = 0$, i.e. X is a deterministic function of $f(X)$, i.e. f is invertible. □

Proposition 4.14 (Properties of Conditional Entropy) For discrete RVs X, Y, Z:

- Chain rule: $H(X, Z \mid Y) = H(X \mid Y) + H(Z \mid X, Y)$.
- Subadditivity: $H(X, Z \mid Y) \leq H(X \mid Y) + H(Z \mid Y)$ with equality iff X and Z are conditionally independent given Y.
- Conditioning reduces entropy: $H(X \mid Y, Z) \leq H(X \mid Y)$ with equality iff X and Z are conditionally independent given Y.

Proof. Exercise. □

Theorem 4.15 (Fano's Inequality) Let X and Y be RVs on respective alphabets A and B. Suppose we are interested in the RV X but only are allowed to observe the possibly correlated RV Y. Consider the estimate $\widehat{X} = f(Y)$, with probability of error $P_e := \Pr(\widehat{X} \neq X)$. Then

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

where h is the binary entropy function.

Proof (Hints). Consider an "error" Bernoulli RV E which depends on X and Y. Use the chain rule in two directions on $H(X, E \mid Y)$. Merge these and split up into the cases when $E = 0$ and $E = 1$ (using) □

Proof. Let E be the binary RV taking value 1 when there is an error (i.e. $\widehat{X} \neq X$), and taking value 0 otherwise. So $E \sim \text{Bern}(P_e)$ and $H(E) = h(P_e)$. Then

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

since E is function of (X, Y). Using the chain rule in the other direction,

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

Now

$$\begin{aligned} H(X \mid Y) - h(P_e) &\leq H(X \mid E, Y) \\ &= P_e H(X \mid E = 1, Y) + (1 - P_e) H(X \mid E = 0, Y) \end{aligned}$$

When $E = 0$, given Y, we can determine $X = f(Y)$ as a function of Y, so $H(X \mid E = 0, Y) = 0$. When $E = 1$, given Y, we know X doesn't take value $f(Y)$, so there are $|A| - 1$ possible values that it takes, so $H(X \mid E = 1, Y) \leq \log(|A| - 1)$. □

4.3. Properties of relative entropy

Theorem 4.16 (Data Processing Inequality for Relative Entropy) Let $X \sim P_X$ and $X' \sim Q_X$ be RVs on the same alphabet A, and $f : A \to B$ be an arbitrary function. Let $P_{f(X)}$ and $Q_{f(X)}$ be the PMFs of $f(X)$ and $f(X')$ respectively. Then

$$D\big(P_{f(X)} \parallel Q_{f(X)}\big) \le D(P_X \parallel Q_X).$$

Proof (Hints). Use that $P_{f(X)}(y) = \sum_{x \in f^{-1}(\{y\})} P_X(x)$. □

Proof. For each $y \in B$, let $A_y = \{x \in A : f(x) = y\} = f^{-1}(\{y\})$. Then

$$\begin{aligned}
D\big(P_{f(X)} \parallel Q_{f(X)}\big) &= \sum_{y \in B} P_{f(X)}(y) \log \frac{P_{f(X)}(y)}{Q_{f(X)}(y)} \\
&= \sum_{y \in B} \left(\sum_{x \in A_y} P_X(x) \right) \log \frac{\sum_{x \in A_y} P_X(x)}{\sum_{x \in A_y} Q_X(x)} \\
&\le \sum_{y \in B} \sum_{x \in A_y} P_X(x) \log \frac{P_X(x)}{Q_X(x)} \quad \text{by log-sum inequality} \\
&= \sum_{x \in A} P_X(x) \log \frac{P_X(x)}{Q_X(x)} = D(P_X \parallel Q_X).
\end{aligned}$$

□

Remark 4.17 The data processing inequality for relative entropy shows that we cannot make two distributions more "distinguishable" by first "processing" the data (by applying f).

Definition 4.18 The **total variation distance** between PMFs P and Q on the same alphabet A is

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

Remark 4.19 Let $B = \{x \in A : P(x) > Q(x)\}$, then

$$\begin{aligned}
\|P - Q\|_{\mathrm{TV}} &= \sum_{x \in A} |P(x) - Q(x)| \\
&= \sum_{x \in B} (P(x) - Q(x)) + \sum_{x \in B^c} (Q(x) - P(x)) \\
&= P(B) - Q(B) + Q(B^c) - P(B^c) \\
&= P(B) - Q(B) + (1 - Q(B)) + (1 - P(B)) \\
&= 2(P(B) - Q(B)).
\end{aligned}$$

Notation 4.20 Write

$$D_e(P \parallel Q) = (\ln 2) P(D \parallel Q) = \sum_{x \in A} P(x) \log_e \frac{P(x)}{Q(x)}$$

and more generally, write

$$D_c(P \parallel Q) = (\log_c 2)P(D \parallel Q) = \sum_{x\in A} P(x)\log_c \frac{P(x)}{Q(x)}.$$

Theorem 4.21 (Pinsker's Inequality) Let P and Q be PMFs on the same alphabet A. Then

$$\|P-Q\|_{\text{TV}}^2 \le (2\ln 2)D(P \parallel Q) = 2D_e(P \parallel Q).$$

Proof (Hints).

- First prove for case that P and Q are PMFs of $\text{Bern}(p)$ and $\text{Bern}(q)$ (explain why we can assume $q \le p$ WLOG), by definining $\Delta(p,q) = 2D_e(P \parallel Q) - \|P-Q\|_{\text{TV}}^2$, and showing that $\frac{\partial\Delta(p,q)}{\partial q} \le 0$.
- Then show for general PMFs by using data processing, where $f = \mathbb{1}_B$ for $B = \{x \in A : P(x) > Q(x)\}$.

□

Proof. First, assume that P and Q are the PMFs of the distributions $\text{Bern}(p)$ and $\text{Bern}(q)$ for some $0 \le q \le p \le 1$ ($q \le p$ WLOG since we can simultaneously interchange both P with $1-P$ and Q with $1-Q$ if necessary). Let

$$\Delta(p,q) = (2\ln 2)D(P \parallel Q) - \|P-Q\|_{\text{TV}}^2 = 2p\ln\frac{p}{q} + 2(1-p)\ln\frac{1-p}{1-q} - (2(p-q))^2.$$

Since $\Delta(p,p) = 0$ for all p, it suffices to show that $\frac{\partial\Delta(p,q)}{\partial q} \le 0$. Indeed,

$$\frac{\partial\Delta(p,q)}{\partial q} = -2\frac{p}{q} + 2\frac{1-p}{1-q} + 8(p-q) = 2(q-p)\left(\frac{1}{q(1-q)} - 4\right) \le 0$$

since $q(1-q) \le \frac{1}{4}$ for all $q \in [0,1]$.

Now, assume P and Q are general PMFs and let $B = \{x \in A : P(x) > Q(x)\}$ and $f = \mathbb{1}_B$. Define the RVs $X \sim P$ and $X' \sim Q$, and let P_f and Q_f be the respective PMFs of the RVs $f(X)$ and $f(X')$. Note that $f(X) \sim \text{Bern}(p)$, $f(X') \sim \text{Bern}(q)$ where $p = P(B)$ and $q = Q(B)$. Then

$$\begin{aligned}
2D_e(P \parallel Q) &\ge 2D_e(P_f \parallel Q_f) && \text{by data-processing} \\
&\ge \left\|P_f - Q_f\right\|_{\text{TV}}^2 && \text{by above} \\
&= (2(p-q))^2 \\
&= (2(P(B)-Q(B)))^2 \\
&= \|P-Q\|_{\text{TV}}^2.
\end{aligned}$$

□

Theorem 4.22 (Convexity of Relative Entropy) The relative entropy $D(P \parallel Q)$ is jointly convex in P, Q: for all PMFs P, P', Q, Q' on the same alphabet and for all $0 < \lambda < 1$,

$$D(\lambda P + (1-\lambda)P' \parallel \lambda Q + (1-\lambda)Q') \leq \lambda D(P \parallel Q) + (1-\lambda)D(P' \parallel Q').$$

Proof. Exercise. □

Corollary 4.23 (Concavity of Entropy) The entropy of $H(P)$ is a concave function on all PMFs P on a finite alphabet.

Proof (Hints). Use convexity of relative entropy of P and a suitable distribution. □

Proof. Let P be a PMF on finite alphabet A and U be the uniform PMF on A. Then by convexity of relative entropy, $D(P \parallel U) = \sum_{x \in A} p(x) \log \frac{P(x)}{1/|A|} = \log m - H(P)$ is convex in P, so $H(P)$ is concave in P. □

5. Poisson approximation

5.1. Poisson approximation via entropy

Theorem 5.1 Let $X_1, ..., X_n$ be IID RVs with each $X_i \sim \text{Bern}(\lambda/n)$, let $S_n = X_1 + \cdots + X_n$. Then $P_{S_n} \to \text{Pois}(\lambda)$ in distribution as $n \to \infty$, i.e. $\forall k \in \mathbb{N}$,

$$\Pr(S_n = k) \to e^{-\lambda} \frac{\lambda^k}{k!} \quad \text{as } n \to \infty$$

Remark 5.2 Using information theory, we can derive stronger and more general statements than the one above.

Theorem 5.3 Let $X_1, ..., X_n$ be (not necessarily independent) RVs with each $X_i \sim \text{Bern}(p_i)$. Let $S_n = \sum_{i=1}^n X_i$ and $\lambda = \sum_{i=1}^n p_i = \mathbb{E}[S_n]$. Then

$$D_e\left(P_{S_n} \parallel \text{Pois}(\lambda)\right) \leq \sum_{i=1}^n p_i^2 + \left(\sum_{i=1}^n H_e(X_i) - H_e(X_1^n)\right).$$

Proof (Hints).

- Let $Z_i = \text{Pois}(p_i)$ for each $i \in [n]$ be independent Poisson RVs so that $T_n = \sum_{i=1}^n Z_i \sim \text{Pois}(\lambda)$.
- Use data processing inequality for relative entropy, and prove the fact that $D_e(\text{Bern}(p) \parallel \text{Pois}(p)) \leq p^2$ for all $p \in [0,1]$ (use that $1 - p \leq e^{-p}$).

□

Proof. Let $Z_i = \text{Pois}(p_i)$ for each $i \in [n]$ be independent Poisson RVs so that $T_n = \sum_{i=1}^n Z_i \sim \text{Pois}(\lambda)$. Then

$$\begin{aligned}
D_e\left(P_{S_n} \,\|\, \mathrm{Pois}(\lambda)\right) &= D_e\left(P_{S_n} \,\|\, P_{T_n}\right) \\
&\leq D_e\left(P_{X_1^n} \,\|\, P_{Z_1^n}\right) \quad \text{by data-processing with } f(x_1^n) = x_1 + \cdots + x_n \\
&= \mathbb{E}\left[\ln \frac{P_{X_1^n}(X_1^n)}{P_{Z_1^n}(X_1^n)}\right] \\
&= \mathbb{E}\left[\ln\left(\frac{P_{X_1^n}(x_1^n)}{\prod_{i=1}^n P_{Z_1^n}(X_i)} \cdot \frac{\prod_{i=1}^n P_{X_i}(X_i)}{\prod_{i=1}^n P_{X_i}(X_i)}\right)\right] \\
&= \mathbb{E}\left[\ln\left(\prod_{i=1}^n \frac{P_{X_i}(x_i)}{P_{Z_i}(x_i)}\right)\right] + \sum_{x_1^n \in A^n} P_{X_1^n}(x_1^n) \ln \frac{1}{\prod_{i=1}^n P_{X_i}(x_i)} - H_e(X_1^n) \\
&= \sum_{i=1}^n D_e\left(P_{X_i} \,\|\, P_{Z_i}\right) + \sum_{i=1}^n H_e(X_i) - H_e(X_1^n)
\end{aligned}$$

since for given $x_1 \in A$, $\sum_{x_2^n \in A^n} P_{X_1^n}(x_1^n) = P_{X_1}(x_1)$ (and similarly for each x_j, $j = 2, ..., n$). Now note that $D_e\left(P_{X_i} \,\|\, P_{Z_i}\right) = D_e(\mathrm{Bern}(p_i) \,\|\, \mathrm{Pois}(p_i))$, and for all $p \in (0, 1)$,

$$\begin{aligned}
D_e(\mathrm{Bern}(p) \,\|\, \mathrm{Pois}(p)) &= (1-p)\ln\frac{1-p}{e^{-p}} + p\ln\frac{p}{pe^{-p}} \\
&= (1-p)\ln(1-p) + (1-p)p + p^2 \\
&\leq (1-p)\ln(e^{-p}) + p \\
&= p^2
\end{aligned}$$

since $1 - p \leq e^{-p}$ for all $p \in [0, 1]$. Similarly, if $p = 0$ or 1, then $D_e(\mathrm{Bern}(p) \,\|\, \mathrm{Pois}(p)) = 0 \leq p^2$. □

Corollary 5.4 Let $X_1, ..., X_n$ be independent, with each $X_i \sim \mathrm{Bern}(p_i)$. Then

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

Corollary 5.5 Theorem 5.1 follows directly from Theorem 5.3.

Proof. Let P_λ be the PMF of the $\mathrm{Pois}(\lambda)$ distribution. Then by Pinsker's inequality,

$$\left\|P_{S_n} - P_\lambda\right\|_{\mathrm{TV}}^2 \leq 2D_e\left(P_{S_n} \,\|\, \mathrm{Pois}(\lambda)\right) \leq 2\sum_{i=1}^n \frac{\lambda^2}{n^2} = 2\frac{\lambda^2}{n}.$$

So for each $k \in \mathbb{N}$, $\left|P_{S_n}(k) - P_\lambda(k)\right| \leq \left\|P_{S_n} - P_\lambda\right\|_{\mathrm{TV}} \leq \sqrt{\frac{2}{n}}\lambda \to 0$ as $n \to \infty$. □

Remark 5.6 Theorem 5.3 is stronger than Theorem 5.1 in that it holds for all n rather than being asymptotic. It also provides an easily computable bound on the difference between P_{S_n} and $\mathrm{Pois}(\lambda)$, and does not assume the p_i are equal, or that the RVs $X_1, ..., X_n$ are independent.

Remark 5.7 It is known that for independent $X_1, ..., X_n$, $P_{S_n} \to \text{Pois}(\lambda)$ iff $\sum_{i=1}^n p_i^2 \to 0$. So the bound in Theorem 5.3 is the best possible.

5.2. What is the Poisson distribution?

Lemma 5.8 (Binomial Maximum Entropy) Let $B_n(\lambda)$ be set of distributions on $\mathbb{N}_0$ that arise from sums $\sum_{i=1}^n X_i$ where $X_i \sim \text{Bern}(p_i)$ are independent and $\sum_{i=1}^n p_i = \lambda$. For all $n \geq \lambda$,

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

Proof. Exercise. □

Theorem 5.9 (Poisson Maximum Entropy) We have

$$\begin{aligned} &H_e(\text{Pois}(\lambda)) \\ &= \sup\left\{H_e(S_n) : S_n = \sum_{i=1}^n X_i, X_i \sim \text{Bern}(p_i) \text{ independent} \wedge \sum_{i=1}^n p_i = \lambda, n \geq 1\right\} \\ &= \sup_{n \in \mathbb{N}} \sup\left\{H_{e(P)} : P \in B_n(\lambda)\right\}. \end{aligned}$$

Proof. Let $H^* = \sup_{n \in \mathbb{N}} \sup\{H_e(P) : P \in B_n(\lambda)\}$. Note that $B_n(\lambda) \subseteq B_{n+1}(\lambda)$, hence $H^* = \lim_{n\to\infty} \sup\left\{H_{e(P)} : P \in B_n(\lambda)\right\} = \lim_{n\to\infty} H_e(\text{Bin}(n, \lambda/n))$.

Let P_n and Q be respective PMFs of $\text{Bin}(n, \lambda/n)$ and $\text{Pois}(\lambda)$. Using that $k! \leq k^k \leq e^{k^2}$, we have

$$\begin{aligned} H_e(Q) &= \sum_{k=0}^{\infty} Q(k) \ln \frac{k!}{e^{-\lambda}\lambda^k} \\ &\leq \sum_{k=0}^{\infty} Q(k)(\lambda - k \ln \lambda + k^2) \\ &= \lambda^2 + 2\lambda - \lambda \ln \lambda < \infty \end{aligned}$$

since $\mathbb{E}[X] = \lambda$ and $\mathbb{E}[X^2] = \lambda + \lambda^2$ for $X \sim \text{Pois}(\lambda)$. So $H_e(Q)$ is finite. The convergence is left as an exercise. □

6. Mutual information

Definition 6.1 The **mutual information** between discrete RVs X and Y is

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

The **conditional mutual information** between X and Y given a discrete RV Z is

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

Proposition 6.2 Let X and Y be discrete RVs with marginal PMFs P_X and P_Y respectively, and joint PMF $P_{X,Y}$, then the mutual information can be expressed as:

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

Proof (Hints). Straightforward. □

Proof. The first two lines are by the chain rule. For the third, we have

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

□

Remark 6.3

- $I(X;Y)$ is symmetric in X and Y.
- The sum of the information contain in X and Y separately minus the information contained in the pair indeed is the amount of mutual information shared by both.
- Considering Stein's Lemma, we can consider $I(X;Y)$ as a measure of how well data generated from $P_{X,Y}$ can be distinguished from independent pairs (X', Y') generated by the product distribution $P_X P_Y$, so is a measure of how far X and Y are from being independent.

Proposition 6.4

- $0 \leq I(X;Y) \leq H(X)$ with equality to 0 iff X and Y are independent.
- Similarly, $I(X;Z \mid Y) \geq 0$ with equality iff $X - Y - Z$, i.e. X and Z are conditionally independent given Y.

Proof. First is by Proposition 6.2 and non-negativity of conditional entropy, second is an exercise. □

Proposition 6.5 (Chain Rule for Mutual Information) For all discrete RVs $X_1, ..., X_n, Y$,

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

Proof (Hints). Straighforward. □

Proof. By the chain rule for entropy,

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

□

Theorem 6.6 (Data Processing Inequalities for Mutual Information) If $X - Y - Z$ (so X and Z are conditionally independent given Y), then

$$I(X;Z), I(X;Y \mid Z) \le I(X;Y).$$

Proof (Hints). Use chain rule for mutual information twice on the same expression. □

Proof. By the chain rule, we have

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

Now $I(X;Z \mid Y) = 0$ by conditional independence, so $I(X;Y) = I(X;Z) + I(X;Y \mid Z)$. □

Example 6.7 We always have $X - Y - f(Y)$, hence $I(X;f(Y)) \le I(X;Y)$, so applying a function to Y cannot make X and Y "less independent".

6.1. Synergy and redundancy

Note 6.8 $I(X;Y_1,Y_2)$ can greater than, equal to, or less than $I(X;Y_1) + (X;Y_2)$.

Definition 6.9 The **synergy** of Y_1, Y_2 about X is

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

So the synergy can be < 0, > 0 or $= 0$.

Definition 6.10 If $S(X;Y_1,Y_2)$ is:

- negative, then Y_1 and Y_2 contain **redundant** information about X;
- zero, then Y_1 and Y_2 are **orthogonal**;
- positive, then Y_1 and Y_2 are **synergistic**. Intuitively, knowing Y_1 already makes the information in Y_2 more valuable (in that it gives more information about X).

Theorem 6.11 Let RVs Y_1, Y_2 be conditionally independent given X, each with distribution $P_{Y \mid X}$, and RVs Z_1, Z_2 be distributed according to $Q_{Z \mid Y}(\cdot \mid Y_1)$, $Q_{Z \mid Y}(\cdot \mid Y_2)$ respectively. Let RV Y have distribution $P_{Y \mid X}$, and W_1, W_2 be conditionally independent given Y, distributed according to $Q_{Z \mid Y}(\cdot \mid Y)$.

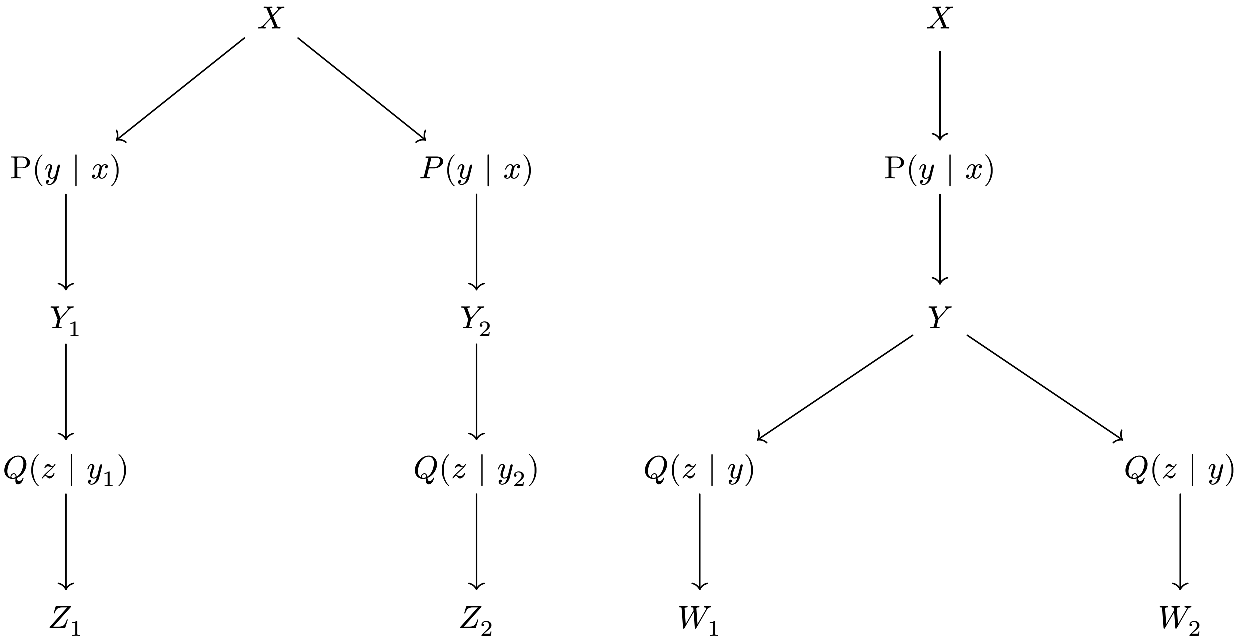
If $S(X;W_1,W_2) > 0$, then $I(X;W_1,W_2) > I(X;Z_1,Z_2)$, for independent Z_1 and Z_2, i.e. correlated observations are better than independent ones.

Proof (Hints). Use data processing for mutual information. □

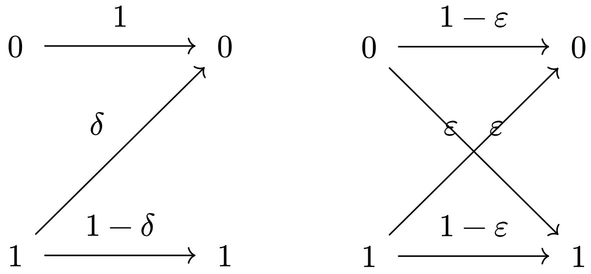
Proof. As in Definition 6.9, we have $I(X;W_2 \mid W_1) > I(X;W_2)$. $I(X;W_2) = I(X;Z_2)$ since (X,W_2) has the same joint distribution as (X,Z_2). By the data processing inequality, we have $I(X;Z_2 \mid Z_1) = I(Z_2;X \mid Z_1) \leq I(Z_2;X) = I(X;Z_2)$, since Z_1 and Z_2 are conditionally independent given X. Hence $I(X;W_2 \mid W_1) > I(X;Z_2 \mid Z_1)$, so $I(X;W_2 \mid W_1) + I(X;W_1) > I(X;Z_2 \mid Z_1) + I(X;Z_1)$, and the result follows by the chain rule. □

Example 6.12 Given two equally noisy channels of a signal X, we want to decide whether it is better (gives more information about X) for the channels to be independent (this corresponds with choosing the Y_1,Y_2,Z_1,Z_2) or correlated (this corresponds with choosing the Y,W_1,W_2).

The natural assumption that the conditionally independent observations Z_1,Z_2 would be "better" than W_1,W_2 (i.e. $I(X;Z_1,Z_2) \geq I(X;W_1,W_2)$) is **false**. We can show diagramatically as



Example 6.13 For example, let $P_{Y \mid X}$ be the Z-channel: if $X = 0$, then $Y = 0$ with probability 1, and if $X = 1$, then $Y \sim \mathrm{Bern}(1-\delta)$ for some $\delta \in (0,1)$. Let $Q_{Z \mid Y}$ be a binary symmetric channel: given Y taking values in $0,1$, $Z = Y$ with probability $1-\varepsilon$, and $Z = 1-Y$ with probability ε for some $\varepsilon \in (0,1)$. We can represent this as



If $X \sim \mathrm{Bern}(1/2)$, $\delta = 0.85$ and $\varepsilon = 0.1$, then $I(X;W_1,W_2) \approx 0.047 > I(X;Z_1,Z_2) \approx 0.039$. So the correlated observations W_1, W_2 are better than the independent observations Z_1, Z_2.

7. Entropy and additive combinatorics

7.1. Simple sumset entropy bounds

Definition 7.1 For $A, B \subseteq \mathbb{Z}$ the **sumset** of A and B is

$$A + B \coloneqq \{a + b : a \in A, b \in B\}.$$

Definition 7.2 For $A, B \subseteq \mathbb{Z}$ the **difference set** of A and B is

$$A - B \coloneqq \{a - b : a \in A, b \in B\}.$$

Proposition 7.3 Let $A, B \subseteq \mathbb{Z}$ be finite. Then

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

Proof (Hints). Trivial. □

Proof. Trivial. □

Proposition 7.4 (Ruzsa Triangle Inequality) Let $A, B, C \subseteq \mathbb{Z}$ be finite. Then

$$|A - C| \cdot |B| \le (|A - B||B - C|).$$

Proof (Hints). Show that an appropriate function is injective. □

Proof. Fix a presentation $y = a_y - c_y$ (where $a_y \in A, c_y \in C$) for each $y \in A - C$. Let

$$\begin{aligned} f : B \times (A - C) &\to (A - B) \times (B - C) \\ (b, y) &\mapsto \big(a_y - b, b - c_y\big). \end{aligned}$$

If $f(b, y) = f(b', y')$, then $a_{y'} - b' = a_y - b$ and $b' - c_{y'} = b - c_y$. So $a_y - a_{y'} = b - b' = c_y - c_{y'}$. So $y = a_y - c_y = a_{y'} - c_{y'} = y'$. Hence $a_y = a_{y'}$, and so $b = b'$. So f is injective, so $|B \times (A - C)| \le |(A - B) \times (B - C)|$. □

Remark 7.5 If X_1^n is a large collection of IID RVs with common PMF P on alphabet A, then the AEP tells us that we can concentrate on the 2^{nH} typical strings. $2^{nH} = (2^H)^n$ is typically much smaller than all $|A|^n = \big(2^{\log|A|}\big)^n$ strings. We can think of $(2^H)^n$ as the effective support size of P^n, and can of 2^H as the effective support size of a single RV with entropy H.

Remark 7.6 We can use the above interpretation to obtain useful conjectures about bounds for the entropy of discrete RVs, from corresponding results on bounds on sumsets. We start with a sumset bound, then replace subsets of $\mathbb{Z}$ by independent RVs on $\mathbb{Z}$, and replace $\log|A|$ of each set A by the entropy of the corresponding RV.

Proposition 7.7 Let X and Y are independent RVs on alphabet $\mathbb{Z}$, then

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

Proof (Hints).

- For lower bound, show that $H(X) \le H(X+Y)$ using data processing and similarly for $H(Y)$. The upper bound should follow directly from this calculation.

□

Proof. For the lower bound,

$$\begin{aligned} H(X) + H(Y) &= H(X,Y) && \text{by Chain Rule for Entropy} \\ &= H(Y, X+Y) && \text{by Data Processing} \\ &= H(X+Y) + H(Y \mid X+Y) && \text{by Chain Rule for Entropy} \\ &\le H(X+Y) + H(Y) && \text{by Conditioning Reduces Entropy.} \end{aligned}$$

Note we have equality for data processing, since $(x, y) \mapsto (x, x+y)$ is injective. Hence $H(X+Y) \ge H(X)$, and the same argument shows that $H(X+Y) \ge H(Y)$.

For the upper bound, we have $H(X) + H(Y) = H(X+Y) + H(Y \mid X+Y) \ge H(X+Y)$ by non-negativity of conditional entropy. □

Lemma 7.8 Let X, Y, Z be independent RVs on alphabet $\mathbb{Z}$. Then

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

Proof (Hints).

- Show that $I(X; X-Z) \le I(X; (X-Y, Y-Z))$.
- Rewrite both sides of the above inequality in terms of entropies, using Data Processing.

□

Proof. Since $X-Z = (X-Y) + (Y-Z)$, X and $X-Z$ are conditionally independent given $(X-Y, Y-Z)$ by Note 4.10. Thus by Data Processing for mutual information, we have $I(X; (X-Y, Y-Z)) \ge I(X; X-Z)$. Now

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

by Data Processing (since, given $X = x$, $x - z \mapsto z$ is injective), and independence of X and Z. Also,

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

by Data Processing (since $(x, x-y, y-z) \mapsto (x,y,z)$ is injective), and independence of X, Y and Z. □

Theorem 7.9 (Ruzsa Triangle Inequality for Entropy) Let X,Y,Z be independent RVs on alphabet $\mathbb{Z}$. Then

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

Proof (Hints). By above lemma. □

Proof. By the above lemma, we have

$$\begin{aligned} H(X-Z) + H(Y) &\leq H(X-Y,Y-Z) \\ &= H(X-Y) + H(Y-Z \mid X-Y) \quad \text{by Chain Rule for Entropy} \\ &\leq H(X-Y) + H(Y-Z). \end{aligned}$$

by Conditioning Reduces Entropy. □

7.2. The doubling-difference inequality for entropy

Definition 7.10 For IID RVs X_1, X_2 on alphabet $\mathbb{Z}$, the **entropy-increase** due to addition (Δ^+) or subtraction (Δ^-) is

$$\begin{aligned} \Delta^+ &:= H(X_1 + X_2) - H(X_1), \\ \Delta^- &:= H(X_1 - X_2) - H(X_1). \end{aligned}$$

Proposition 7.11 For IID X_1, X_2 on $\mathbb{Z}$, we have

$$\begin{aligned} \Delta^+ &= I(X_1 + X_2; X_2), \\ \Delta^- &= I(X_1 - X_2; X_2). \end{aligned}$$

Proof (Hints). Straightforward. □

Proof. We have

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

by Data Processing (since $(x_1 + x_2, x_2) \mapsto (x_1, x_2)$ is injective) and Chain Rule for Entropy. The proof is identical for Δ^-. □

Lemma 7.12 Let X,Y,Z be independent RVs on alphabet $\mathbb{Z}$. Then

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

Proof (Hints).

- Show that $I(X; X+Y+Z) \leq I(X+Y; X)$.
- Rewrite both sides in terms of entropies.

□

Proof. Since $X - (X+Y, Z) - (X+Y+Z)$ form a Markov chain by Note 4.10, we have, by Data Processing and Chain Rule for mutual information,

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

since Z is (conditionally) independent of X given $X+Y$. Now

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

since $(y, x) \mapsto (x+y, x)$ is injective and X and Y are independent. Also,

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

since, given $X = x$, $x+y+z \mapsto y+z$ is injective, and X and $Y+Z$ are independent.
□

Theorem 7.13 (Doubling-difference Inequality) Let X_1 and X_2 be IID RVs on $\mathbb{Z}$. Then

$$\frac{1}{2} \leq \frac{\Delta^+}{\Delta^-} \leq 2.$$

Proof (Hints).

- For lower bound, use Ruzsa Triangle Inequality for appropriate RVs.
- For upper bound,

□

Proof. For the lower bound, let $X, -Y, Z$ be IID with the same distribution as X_1. Then by the Ruzsa Triangle Inequality,

$$H(X_1 - X_2) + H(X_1) \leq H(X_1 + X_2) + H(X_1 + X_2).$$

So $2(H(X_1 + X_2) - H(X_1)) \geq H(X_1 - X_2) - H(X_1)$.

For the upper bound, let $X, -Y, Z$ be IID with the same distribution as X_1. Then by the above lemma and Proposition 7.7,

$$H(X_1 + X_2) + H(X_1) \leq H(X_1 - X_2) + H(X_1 - X_2)$$

so $H(X_1 + X_2) - H(X_1) \leq 2(H(X_1 - X_2) - H(X_1))$. □

8. Entropy rate

Definition 8.1 For an arbitrary source $\boldsymbol{X} = \{X_n : n \in \mathbb{N}\}$, the **entropy rate** $H(\boldsymbol{X})$ of $\boldsymbol{X}$ is the limit of the average number of bits per symbol:

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

whenever the limit exists.

Example 8.2 If $\boldsymbol{X}$ is memoryless (so a sequence of IID RVs) with common entropy $H = H(X_i)$, then the entropy rate is

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

Example 8.3 Let $\boldsymbol{X} = \{X_n : n \in \mathbb{N}\}$ be an irreducible, aperiodic Markov chain on a finite alphabet A with transition matrix Q, where

$$Q_{ab} = \Pr(X_{n+1} = b \mid X_n = a), \quad \forall a, b \in A$$

Let $X_1 \sim P_{X_1}$ be the initial distribution and π be the unique stationary distribution ($\Pr(X_n = x) \to \pi(x)$ as $n \to \infty$). $\boldsymbol{X}$ has a unique invariant distribution π to which it converges:

$$\forall x \in A, \quad \Pr(X_n = x) \to \pi(x) \quad \text{as } n \to \infty$$

and hence also

$$\Pr(X_{n-1} = x, X_n = y) = \Pr(X_n = x) Q_{xy} \to \pi(x) Q_{xy}.$$

Then by the Chain Rule for Entropy and conditional independence,

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

By the convergence theorem for Markov chains, we have $P_{X_n} \to \pi$ as $n \to \infty$. $H(X \mid Y)$ is a continuous function of the joint distribution $P_{X,Y}$, so $H(X_n \mid X_{n-1}) \to H\left(\overline{X_1} \mid \overline{X_0}\right)$ as $n \to \infty$, where $\overline{X_0} \sim \pi$ and $\Pr\left(\overline{X_1} = b \mid \overline{X_1} = a\right) = Q_{ab}$. We have

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

The first term tends to 0 since the numerator is bounded, and the summands in the second term tend to $H\left(\overline{X_1} \mid \overline{X_0}\right)$. So the entropy rate exists and is equal to $H(\boldsymbol{X}) = H\left(\overline{X_1} \mid \overline{X_0}\right)$.

Definition 8.4 A source $\boldsymbol{X}$ is **stationary** if for any block length $n \in \mathbb{N}$, the distribution of X_{k+1}^{k+n} is independent of k.

Remark 8.5 If $\boldsymbol{X} = \{X_n : n \in \mathbb{N}\}$ is one-sided stationary process, then by Kolmogorov's extension theorem, $\boldsymbol{X}$ admits a unique two-sided extension to $\boldsymbol{X} = \{X_n : n \in \mathbb{Z}\}$.

Theorem 8.6 If $\boldsymbol{X} = \{X_n : n \in \mathbb{N}\}$ is a stationary process on finite alphabet A, then its entropy rate exists and is equal to

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

Proof (Hints). Show that the sequence $\{H(X_n) \mid X_1^{n-1} : n \in \mathbb{N}\}$ is non-increasing and use the Cèsaro Lemma. □

Proof. The sequence $\{H(X_n) \mid X_1^{n-1} : n \in \mathbb{N}\}$ is non-negative by non-negativity of conditional entropy, and is non-increasing, since

$$\begin{aligned} H(X_{n+1} \mid X_1^n) &\le H(X_{n+1} \mid X_2^n) && \text{by Conditioning Reduces Entropy} \\ &= H\left(X_2^{n+1}\right) - H(X_2^n) && \text{by Chain Rule for Entropy} \\ &= H(X_1^n) - H(X_1^{n-1}) && \text{by stationarity} \\ &= H(X_{n-1} \mid X_1^{n-2}) && \text{by Chain Rule for Entropy.} \end{aligned}$$

Hence the limit $\lim_{n\to\infty} H(X_n \mid X_1^{n-1})$ exists, and so by the Cèsaro Lemma, the averages converge to the same limit. But by the Chain Rule for Entropy, the averages are

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

□

Theorem 8.7 For a stationary process $\boldsymbol{X} = \{X_n : n \in \mathbb{Z}\}$ on a finite alphabet A,

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

Proof (Hints). Uses measure-theoretic probability, so beyond the scope of the course. □

Proof. By Martingale convergence, we have that

$$P(x_0 \mid X_{-n}^{-1}) \to P(x_0 \mid X_{-\infty}^{-1}) \quad \text{almost surely} \quad \text{as } n \to \infty,$$

where $P(\cdot \mid x_{-n}^{-1})$ is the conditional distribution of X_0 given $X_{-n}^{-1} = x_{-n}^{-1}$, and $P(\cdot \mid x_{-\infty}^{-1})$ is the conditional distribution of X_0 given $X_{-\infty}^{-1} = x_{-\infty}^{-1}$. Now, we can take expectations to obtain that, by the bounded convergence theorem (since $p \mapsto p \log p$ is continuous and bounded for $p \in [0, 1]$),

$$\begin{aligned} H(X_0 \mid X_{-n}^{-1}) &= \mathbb{E}\left[-\sum_{x_0 \in A} P(x_0 \mid X_{-n}^{-1}) \log P(x_0 \mid X_{-n}^{-1})\right] \\ &\to \mathbb{E}\left[-\sum_{x_0 \in A} P(x_0 \mid X_{-\infty}^{-1}) \log P(x_0 \mid X_{-\infty}^{-1})\right] \\ &=: H(X_0 \mid X_{-\infty}^{-1}) \quad \text{almost surely} \quad \text{as } n \to \infty. \end{aligned}$$

Finally, $H(X_0 \mid X_{-n}^{-1}) = H(X_{n+1} \mid X_1^n)$ by stationarity, so we are done by Theorem 8.6. □

Definition 8.8 Let $\boldsymbol{X} = \{X_n : n \in \mathbb{Z}\}$ be a stationary source on finite alphabet A, and define the (left) **shift** operator $T : A^{\mathbb{Z}} \to A^{\mathbb{Z}}$ on sequences $A^{\mathbb{Z}}$ by

$$(Tx)_n = x_{n+1} \quad \forall n \in \mathbb{Z}.$$

$\boldsymbol{X}$ is **ergodic** if all shift invariant events are trivial, i.e. for any measurable $B \subseteq A^{\mathbb{Z}}$, we have

$$T^{-1}B = B \implies \Pr(X_{-\infty}^{\infty} \in B) = 0 \text{ or } 1.$$

Intuitively, an ergodic process is one which satisfies the general form of the strong law of large numbers.

It turns out that ergodicity is equivalent to the validity of the following:

Theorem 8.9 (Birkhoff's Ergodic Theorem) Let $\boldsymbol{X} = \{X_n : n \in \mathbb{Z}\}$ be a stationary ergodic source on alphabet A. Then for any measurable function $f : A^{\mathbb{Z}} \to \mathbb{R}$ such that

$$\mathbb{E}[|f(X_{-\infty}^{\infty})|] < \infty,$$

we have

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

Proof (Hints). Beyond the scope of this course. □

Proof. Omitted. □

Remark 8.10 The strong law of large numbers follows instantly from Birkhoff by setting $f(x_{-\infty}^{\infty}) = x_1$.

Example 8.11 Every IID source is ergodic.

Theorem 8.12 (Shannon-McMillan-Breiman) Let $\boldsymbol{X} = \{X_n : n \in \mathbb{N}\}$ be a stationary ergodic source on alphahbet A with entropy rate $H = H(\boldsymbol{X})$, then

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

where P_n is the PMF of X_1^n.

Proof. Idea: by Chain Rule for Entropy, we have

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

but we cannot directly apply the ergodic theorem to this, since $-\log P(X_i \mid X_1^{i-1})$ is not of the form $f(T^i x_{-\infty}^{\infty})$. Instead, note that by Birkhoff's Ergodic Theorem and Theorem 8.7,

$$\begin{aligned}-\frac{1}{n}\log P(X_1^n \mid X_{-\infty}^0) &= \frac{1}{n}\sum_{i=1}^{n}[-\log P(X_i \mid X_{-\infty}^{i-1})]\\ &\to \mathbb{E}[-\log P(X_0 \mid X_{-\infty}^{-1})]\\ &=: H(X_0 \mid X_{-\infty}^{-1}) = H \text{ almost surely} \quad \text{as } n \to \infty.\end{aligned}$$

Also, by Birkhoff's Ergodic Theorem, for each fixed $k \geq 1$,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(-\log P(X_i \mid X_{i-k}^{i-1})) &\to \mathbb{E}[-\log P(X_0 \mid X_{-k}^{-1})]\\ &=: H(X_0 \mid X_{-k}^{-1}) \text{ almost surely} \quad \text{as } n \to \infty.\end{aligned}$$

we show

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

We have

$$\begin{aligned}\Pr\left(-\frac{1}{n}\log P(X_1^n \mid X_{-\infty}^0) - \left(-\frac{1}{n}\log P_n(X_1^n)\right) > \varepsilon\right) &= \Pr\left(\frac{1}{n}\log\frac{P_n(X_1^n)}{P(X_1^n \mid X_{-\infty}^0)} > \varepsilon\right)\\ &= \Pr\left(\frac{P_n(X_1^n)}{P(X_1^n \mid X_{-\infty}^0)} > 2^{n\varepsilon}\right)\\ &\leq 2^{-n\varepsilon}\mathbb{E}\left[\frac{P_n(X_1^n)}{P(X_1^n \mid X_{-\infty}^0)}\right] \quad \text{by markov's inequality}\\ &\leq 2^{-n\varepsilon}\mathbb{E}\left[\mathbb{E}\left[\frac{P_n(X_1^n)}{P(X_1^n \mid X_{-\infty}^0)} \mid X_{-\infty}^0\right]\right]\\ &= 2^{-n\varepsilon}\mathbb{E}\left[\sum_{\substack{x_1^n\\ P(x_1^n \mid X_{-\infty}^0)>0}} P(x_1^n \mid X_{-\infty}^0)\frac{P_n(x_1^n)}{P(x_1^n \mid X_{-\infty}^0)}\right]\\ &\leq 2^{-n\varepsilon}\end{aligned}$$

which is summable, so by Borel-Cantelli,

$$\liminf_{n\to\infty} -\frac{1}{n}\log P(X_1^n \mid X_{-\infty}^0) \leq \liminf_{n\to\infty} -\frac{1}{n}\log P_n(X_1^n) \text{ almost surely.}$$

For each fixed k, consider the sequence of PMFs $Q_n^{(k)}(x_1^n) = P_k(x_1^k)\prod_{i=k+1}^n P(x_i \mid X_{i-k}^{i-1})$ for $x_1^n \in A^n$. Then

$$\begin{aligned}&-\frac{1}{n}\log Q_n^{(k)}(X_1^n) - \left[-\frac{1}{n}\sum_{i=1}^n \log P(x_i \mid x_{i-k}^{i-1})\right]\\ &= -\frac{1}{n}\left[\log P_k(x_1^k) - \sum_{i=1}^k \log P(X_i \mid X_{i-k}^{i-1})\right]\\ &\to 0 \text{ almost surely as } n\to\infty\end{aligned}$$

So suffices to show that $\limsup_{n\to\infty} -\frac{1}{n}\log P_n(X_1^n) \leq \limsup_{n\to\infty} -\frac{1}{n}\log Q_n^{(k)}(X_1^n)$ almost surely. So again, let $\varepsilon > 0$ be arbitrary, then

$$\begin{aligned}&\Pr\left(-\frac{1}{n}\log P_n(X_1^n) - \left(-\frac{1}{n}\log Q_n^{(k)}(X_1^n)\right) > \varepsilon\right)\\ &= \Pr\left(\frac{Q_n^{(k)}(X_1^n)}{P_n(X_1^n)} > 2^{n\varepsilon}\right) \leq 2^{-n\varepsilon}\mathbb{E}\left[\frac{Q_n^{(k)}(X_1^n)}{P_n(X_1^n)}\right] \text{ by markov's inequality}\\ &\leq 2^{-n\varepsilon}\sum_{x_1^n\in A^n} P_n(x_1^n)\frac{Q_n^{(k)}(x_1^n)}{P_n(x_1^n)} = 2^{-n\varepsilon}\end{aligned}$$

which is summable, so by Borel-Cantelli and the fact that $\varepsilon > 0$ was arbitrary, we have

$$\limsup_{n\to\infty} -\frac{1}{n}\log P_n(X_1^n) \leq \limsup_{n\to\infty} -\frac{1}{n}\sum_{i=1}^n \log P(X_i \mid X_{i-k}^{i-1})$$

□

9. Types and large deviations

9.1. The method of types

Definition 9.1 Let A be a finite alphabet and $x_1^n \in A^n$. The **type** of x_1^n is its empirical distribution $\hat{P}_n$:

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

Notation 9.2 For a finite alphabet $A = \{a_1, ..., a_m\}$, let $\mathcal{P}$ denote the set of all PMFs on A:

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

Note that $\mathcal{P}$ is an m-simplex.

Notation 9.3 We write $\mathcal{P}_n$ for the set of all ***n*-types**:

$$\mathcal{P}_n = \{P \in \mathcal{P} : nP(a) \in \mathbb{Z}\ \forall a \in A\}.$$

Note that $\mathcal{P}_n$ is finite.

Proposition 9.4 We have $|\mathcal{P}_n| \leq (n+1)^m$.

Proof (Hints). Straightforward. □

Proof. Each $P \in \mathcal{P}_n$ is of the form $(k_1/n, ..., k_m/n)$. There are at most $(n+1)$ choices $(0, ..., n)$ for each k_i. □

Proposition 9.5 Let $x_1^n \in A^n$ have type $\hat{P}_n$. Then for any PMF Q,

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

In particular, if $Q = \hat{P}_n$, then $Q^n(x_1^n) = 2^{-nH(Q)}$.

Proof (Hints). Rewrite $\log Q^n(x_1^n)$. □

Proof. We have

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

□

Definition 9.6 Given a n-type P, its **type class** is

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

Note that $A^n = \coprod_{P\in\mathcal{P}_n} T(P)$: since A^n has size $|A|^n$ exponential in n, and the union is over $|\mathcal{P}_n| \le (n+1)^m$ (polynomial in n) elements, at least one type class must contain exponentially many strings.

$T(P)$ consists of all possible arrangements of $nP(a_1)$ a_1's, …, $nP(a_m)$ a_m's, so

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

Lemma 9.7 Let $P \in \mathcal{P}_n$. Then

$$P^n(T(P)) = \max\{P^n(T(Q)) : Q \in \mathcal{P}_n\}.$$

i.e. the most likely type class under P^n is $T(P)$.

Proof (Hints).

□

Proof. Let $Q \in \mathcal{P}_n$ be arbitrary. Then

$$\begin{aligned}\frac{P^n(T(P))}{P^n(T(Q))} &= \frac{|T(P)| \cdot \prod_{i=1}^m P(a_i)^{nP(a_i)}}{|T(Q)| \cdot \prod_{i=1}^m P(a_i)^{nQ(a_i)}} \\ &= \frac{n!}{\prod_{i=1}^m (nP(a_i))!} \cdot \frac{\prod_{i=1}^m (nQ(a_i))!}{n!} \prod_{i=1}^m P(a_i)^{n(P(a_i)-Q(a_i))} \\ &= \prod_{i=1}^m P(a_i)^{n(P(a_i)-Q(a_i))} \prod_{i=1}^m \frac{(nQ(a_i))!}{(nP(a_i))!}\end{aligned}$$

Now since $k!/\ell! \geq \ell^{k-\ell}$ (to show this, consider $k \geq \ell$ and $k < \ell$ cases separately), this is

$$\geq \prod_{i=1}^{m} (nP(a_i))^{n(P(a_i)-Q(a_i))} \prod_{i=1}^{m} (n(P(a_i)))^{n(Q(a_i)-P(a_i))}$$

$$= \prod_{i=1}^{m} n^{n(Q(a_i)-P(a_i))}$$

$$= n^{n\sum_{i=1}^{m}(Q(a_i)-P(a_i))} = 1$$

since probabilities sum to 1. □