Contents

1. Entropy

1.1. Introduction

Notation 1.1 Write $x_1^n := (x_1, ..., x_n) \in \{0, 1\}^n$ for an length n bit string.

Notation 1.2 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 1.3 A random variable X has a **Bernoulli distribution**, $X \sim \text{Bern}(p)$, if for some fixed $p \in (0, 1)$,

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

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

Notation 1.4 Throughout, we take log to be the base-2 logarithm, $\log_2$.

Definition 1.5 The **binary entropy function** $h : (0, 1) \to [0, 1]$ is defined as

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

Example 1.6 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 1.7 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 1.8 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 1.9

- 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.

1.2. Asymptotic equipartition property

Notation 1.10 We denote a finite alphabet by $A = \{a_1, ..., a_m\}$.

Notation 1.11 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 1.12 For $i \le j$, 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 1.13 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 1.14 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 1.15 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 1.16

- 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 1.17 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 1.18 (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 1.17, 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

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

$\square$

Remark 1.19
- 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 1.17 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".

1.3. Fixed-rate lossless data compression

Definition 1.20 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 1.21 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 1.22 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 1.23 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 1.24 (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

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

- ($\Longleftarrow$): 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 (Corollary 1.18). Then $P_e^{(n)} = 1 - \Pr(X_1^n \in B_n^*) \to 0$ as $n \to \infty$ by AEP.
  - Also by AEP, $R_n = \frac{1}{n}(1 + \lceil \log|B_n^*| \rceil) \le \frac{1}{n}\log|B_n^*| + \frac{2}{n} \le H + \varepsilon + \frac{2}{n}$.
- ($\Longleftarrow$):
  - WLOG let $0 < \varepsilon < 1/2$. By AEP,

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

  eventually.

□

2. Relative entropy

Definition 2.1 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 2.2 The associated **error probabilities** for a hypothesis test are

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

Definition 2.3 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 2.4

- 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 2.5 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.

2.1. Asymptotically optimal hypothesis testing

Theorem 2.6 (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 2.7

- 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.

2.2. Relative entropy and optimal hypothesis testing

Theorem 2.8 (Neyman-Pearson Lemma) For a hypothesis test between P and Q based on n data samples, the **likelihood ratio decision regions**

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

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

Proof (Hints). Consider the inequality

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

(justify why this holds). □

Proof.

- Consider the obvious inequality

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

- Then, summing over all x_1^n,

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

□

Remark 2.9 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 2.10 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 2.11 The Neyman-Pearson decision region B_{NP} can be expressed in information-theoretic form as

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

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 2.12 (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 2.13 (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 2.14 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, $\frac{A}{B} \log \frac{A}{B} \leq \frac{1}{B} \sum_i a_i \log \frac{a_i}{b_i}$.

$\square$

Proposition 2.15

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

Remark 2.16 This also holds for countably infinite A.

Proof (Hints).

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

$\square$

Proof.

- ‣ By the log-sum inequality,
$$D(P \parallel Q) = \sum_{x \in A} P(x) \log \frac{P(x)}{Q(x)} \geq \left(\sum_{x \in A} P(x) \right) \log \frac{\sum_{x \in A} P(x)}{\sum_{x \in A} Q(x)} = 0$$
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 $0 \leq D(P \parallel Q) = \sum_{x \in A} P(x) \log \frac{P(x)}{1/|A|} = \log|A| - H(X)$ 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.

$\square$

Remark 2.17 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.

3. Properties of entropy and relative entropy

3.1. Joint entropy and conditional entropy

Definition 3.1 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 3.2 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 3.3 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 3.4 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 3.5 $P_{Y \mid X}$ 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 3.6 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

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

□

3.2. Properties of entropy, joint entropy and conditional entropy

Proposition 3.7 (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 3.8 (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 3.9 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 $X - Z - Y$.

Example 3.10 For any function g on Y, we have $X - Y - g(Y)$.

Corollary 3.11 $H(X_1^n) \le \sum_{i=1}^n H(X_i)$ 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 3.12 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 3.13 (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 3.14 (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 3.15 (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 $H(X \mid E = 1, Y) \leq \log(|A| - 1)$. □

3.3. Properties of relative entropy

Theorem 3.16 (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

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

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

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

□

Remark 3.17 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 3.18 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 3.19 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 3.20 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 3.21 (Pinsker's Inequality) Let P and Q be PMFs on the same alphabet A. Then

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

Proof (Hints).

- First prove for case that P and Q are PMFs of Bern(p) and Bern(q) (explain why we can assume $q \leq p$ WLOG), by definining $\Delta(p,q) = 2D_e(P \parallel Q) - \|P - Q\|_{\mathrm{TV}}^2$, and showing that $\frac{\partial \Delta(p,q)}{\partial q} \leq 0$.
- 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 Bern(p) and Bern(q) for some $0 \leq q \leq p \leq 1$ ($q \leq p$ WLOG since we can simultaneously interchange both P with $1 - P$ and Q with $1 - Q$ if necessary). Let

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

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

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

since $q(1-q) \leq \frac{1}{4}$ 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 $f(X) \sim$ Bern(p), $f(X') \sim$ Bern(q) where $p = P(B)$ and $q = Q(B)$. Then

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

□

Theorem 3.22 (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 $0 < \lambda < 1$,

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

Proof. Exercise. □

Corollary 3.23 (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. □

4. Poisson approximation

4.1. Poisson approximation via entropy

Theorem 4.1 Let $X_1, ..., X_n$ be IID RVs with each $X_i \sim \mathrm{Bern}(\lambda/n)$, let $S_n = X_1 + \cdots + X_n$. Then $P_{S_n} \to \mathrm{Pois}(\lambda)$ 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 4.2 Using information theory, we can derive stronger and more general statements than the one above.

Theorem 4.3 Let $X_1, ..., X_n$ be (not necessarily independent) RVs with each $X_i \sim \mathrm{Bern}(p_i)$. Let $S_n = \sum_{i=1}^n X_i$ and $\lambda = \sum_{i=1}^n p_i = \mathbb{E}[S_n]$. Then

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

Proof (Hints).

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

□

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

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

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 $D_e\left(P_{X_i} \parallel P_{Z_i}\right) = D_e(\mathrm{Bern}(p_i) \parallel \mathrm{Pois}(p_i))$, and for all $p \in [0, 1]$,

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

since $1 - p \leq e^{-p}$ for all $p \in [0, 1]$. □

Corollary 4.4 Let $X_1, ..., X_n$ be independent, with each $X_i \sim \mathrm{Bern}(p_i)$. Then

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

Corollary 4.5 Theorem 4.1 follows directly from Theorem 4.3.

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

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

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 4.6 Theorem 4.3 is stronger than Theorem 4.1 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 4.7 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 4.3 is the best possible.

4.2. What is the Poisson distribution?

Lemma 4.8 (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 4.9 (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. □

5. Mutual information

Definition 5.1 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 5.2 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:

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

Proof (Hints). Straightforward. □

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

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

□

Remark 5.3
- $I(X;Y)$ 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 Theorem 2.6, 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 5.4
- $0 \leq I(X;Y) \leq H(X)$ 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 5.2 and non-negativity of conditional entropy, second is an exercise. □

Proposition 5.5 (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 5.6 (Data Processing Inequalities for Mutual Information) If $X - Y - Z$ (so X and Z are conditionally independent given Y), then

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

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

5.1. Synergy and redundancy

Note 5.8 $I(X;Y_1,Y_2)$ can greater than, equal to, or less than $I(X;Y_1) + (X;Y_2)$.

Definition 5.9 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 5.10 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 5.11 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. $\square$

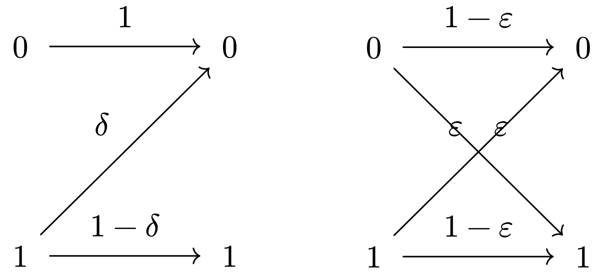
Proof. As in Definition 5.9, we have $I(X; W_2 \mid W_1) > I(X; W_2)$. $I(X; W_2) = I(X; Z_2)$ since (X, W_2) has the same joint distribution as (X, Z_2). By the data processing inequality, we have $I(X; Z_2 \mid Z_1) = I(Z_2; X \mid Z_1) \le I(Z_2; X) = I(X; Z_2)$, since Z_1 and Z_2 are conditionally 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. $\square$

Example 5.12 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. $I(X; Z_1, Z_2) \ge I(X; W_1, W_2)$) is **false**. We can show diagramatically as

$$\begin{tikzcd}
& X \arrow[dl] \arrow[dr] & & & X \arrow[d] & \\
\mathrm{P}(y \mid x) \arrow[d] & & P(y \mid x) \arrow[d] & & \mathrm{P}(y \mid x) \arrow[d] & \\
Y_1 \arrow[d] & & Y_2 \arrow[d] & & Y \arrow[dl] \arrow[dr] & \\
Q(z \mid y_1) \arrow[d] & & Q(z \mid y_2) \arrow[d] & Q(z \mid y) \arrow[d] & & Q(z \mid y) \arrow[d] \\
Z_1 & & Z_2 & W_1 & & W_2
\end{tikzcd}$$

Example 5.13 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.

6. Entropy and additive combinatorics

6.1. Simple sumset entropy bounds

Definition 6.1 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 6.2 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 6.3 Let $A, B \subseteq \mathbb{Z}$ be finite. Then

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

Proof (Hints). Trivial. □

Proof. Trivial. □

Proposition 6.4 (Ruzsa Triangle Inequality) Let $A, B, C \subseteq \mathbb{Z}$ be finite. Then

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

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

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

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 6.5 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. Since $2^{nH} = \left(2^H\right)^n$ is typically much smaller than all $|A|^n = \left(2^{\log|A|}\right)^n$ strings, we can think of 2^H as the effective support size of the X_i.

Proposition 6.6 Let X and Y are independent RVs on alphabet $\mathbb{Z}$, then

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

Proof. For the lower bound,

$$\begin{aligned}
H(X) + H(Y) &= H(X, Y) && \text{by independence} \\
&= H(Y, X+Y) && \text{by data processing} \\
&= H(X+Y) + H(Y \mid X+Y) && \text{by chain rule} \\
&\leq H(X+Y) + H(Y) && \text{by conditioning reduces entropy}
\end{aligned}$$

Hence $H(X+Y) \geq H(X)$, and the same argument shows that $H(X+Y) \geq H(Y)$.

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

Theorem 6.7 (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. By the data processing inequality for mutual information, we have $I(X; (X-Y, Y-Z)) \geq I(X; X-Z)$. So $H(X) + H(X-Y, Y-Z) - H(X, X-Y, Y-Z) \geq H(X) + H(X-Z)$. So $H(X-Y, Y-Z) - H(X,Y,Z) \geq H(X-Z) - H(X,Z)$. Hence $H(X-Y, Y-Z) - H(Y) \geq H(X-Z)$, and $H(X-Y, Y-Z) \geq H(X-Y) + H(Y-Z)$. □

6.2. The doubling-difference inequality for entropy

Definition 6.8 For IID RVs X_1, X_2 on alphabet $\mathbb{Z}$, the **entropy-increase** due to addition (Δ^+) or subtraction (Δ^-) is

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

Lemma 6.9 Let X, Y, Z be independent RVs on alphabet $\mathbb{Z}$. Then

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

In particular, $H(X+Z) + H(Y) \leq H(X+Y) + H(Y+Z)$.

Proof. By the data processing inequality for mutual information, since $X - (X+Y) - (X+Y+Z)$, we have $I(X; X+Y) \geq I(X; X+Y+Z)$, i.e. $H(X) + H(X+Y) - H(X, X+Y) \geq H(X) + H(X+Y+Z) - H(X, X+Y+Z)$. $H(X, X+Y) = H(X,Y) = H(X) + H(Y)$ by independence. So we have $H(X+Y) - H(Y) \geq H(X+Y+Z) - H(Y+Z)$. □

Theorem 6.10 (Doubling-difference Inequality) Let X_1 and X_2 be IID RVs on $\mathbb{Z}$. Then

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

Equivalently,

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

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