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 $p \in (0,1)$,

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

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

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

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

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

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

□

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

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

- ($\Longleftarrow$): 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 (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, $R_n = \frac{1}{n}(1 + \lceil \log|B_n^*| \rceil) \leq \frac{1}{n}\log|B_n^*| + \frac{2}{n} \leq H + \varepsilon + \frac{2}{n}$.
- ($\Longleftarrow$):
  - 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.

□

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 $\frac{1}{2} = 2^{-n(1/n)}$.

$\square$

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,

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

$\square$

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**

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

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

Proof (Hints). Consider the inequality

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

(justify why this holds). □

Proof.

- Consider the obvious inequality

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

- Then, summing over all x_1^n,

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

□

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

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

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 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 \ 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,

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

with equality iff X and Y are independent.

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

Proof. We have

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

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 $\boldsymbol{Y}$** 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 $H(X_1^n) \leq \sum_{i=1}^n H(X_i)$ with equality iff all X_1^n are independent.

Proof. Straightforward. □

Proof. $H(X_1^n) = \sum_{i=1}^n H(X_i \mid X_1^{i-1}) \leq \sum_{i=1}^n H(X_i)$ 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 $H(Y \mid X = x) \leq H(Y)$. Since $0 \leq H(Y \mid X = x) \leq \log|A_Y|$, we have $0 \leq H(Y \mid X) \leq \log|A_Y|$ 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. $H(f(X)) \leq H(X)$ with equality iff f is injective.

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

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. $\square$

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. $\square$

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) $\square$

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 $E \sim \mathrm{Bern}(P_e)$ 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 $H(X \mid E = 1, Y) \leq \log(|A| - 1)$. $\square$