Contents

1. Entropy

1.1. Introduction

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

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

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

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

Example. 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. 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. 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.

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

Notation. 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. 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. 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. 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. 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.

- 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. 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 (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

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

  satisfies

  $$\begin{aligned} |B_n^*(\varepsilon)| &\le 2^{n(H+\varepsilon)} \quad \forall n \in \mathbb{N}, \quad \text{and} \\ P^n(B_n^*(\varepsilon)) = \Pr(X_1^n \in B_n^*(\varepsilon)) &\longrightarrow 1 \quad \text{as } n \to \infty \end{aligned}$$
- ($\Longleftarrow$): 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 $\forall \varepsilon > 0$,

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

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 1.2.8, 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.
- 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 1.2.8 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. 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. 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. 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. 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 (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 1.2.10). 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.

□