Representing Information Losslessly

I-Hsiang Wang

Department of Electrical Engineering
National Taiwan University

ihwang@ntu.edu.tw

September 21, 2023

The information processing task motivating the study of this lecture:

> *For a (random) source sequence of length n, design an encoding scheme (mapping) to describe it using k bits, so that the decoder can reconstruct the source sequence at the destination from these k bits.*

Note: how the encoding scheme works is known by the decoder *a priori*.

Fundamental Questions:

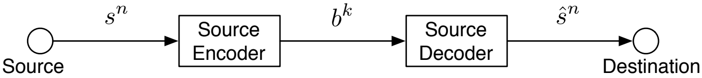
- What is the minimum possible ratio $\frac{k}{n}$ (compression ratio/rate) ?
- How to achieve that fundamental limit?

In this lecture, we will show that, for i.i.d. random sources, the fundamental limit is the entropy of the source when we require lossless reconstruction.

The proof is simple and only requires basic probability tools.

The result can be extended to “well-behaved” random sources.

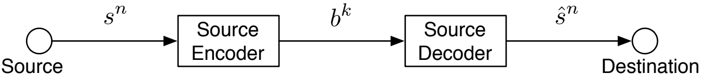
The source coding problem (Shannon's abstraction)



Meta Description

1. **Encoder**: Represent the source sequence s^n by a binary source codeword $w := b^k \in \{0, 1, \ldots, 2^k - 1\}$, with k as small as possible.
2. **Decoder**: From the source codeword w, reconstruct the source sequence either losslessly or within a certain distortion.
3. **Efficiency**: Determined by the code rate $\mathrm{R} := \frac{k}{n}$ bits/symbol time

Criteria of recovery: lossless vs. lossy



Two natural criteria of recovery for the source coding problem:

1. Exact: the reconstructed sequence $\hat{s}^n = s^n$.
2. Lossy: the reconstructed sequence $\hat{s}^n \neq s^n$ but within some distortion.

Source coding is all about **efficient representation** of the source data s^n.

Let's do some simple back-of-envelope analysis of the system with the exact recovery criterion to get some intuition about how small k can be.

If the decoder would like to reconstruct s^n exactly for all possible $s^n \in \mathcal{S}^n$, then it is simple to see that the smallest k must satisfy

$$2^{k-1} < |\mathcal{S}|^n \le 2^k \implies k = \lceil n \log |\mathcal{S}| \rceil.$$

Why?

Because every possible sequence has to be uniquely represented by k bits!

Seems impossible to achieve compression if we require exact reconstruction.

What is going wrong?

Redundancy in a random data source

Compression is possible since there is redundancy in the source sequence.

One of the simplest ways to capture redundancy is to model the data source as a random process. (Another reason to use a random source model is due to engineering reasons, as mentioned before.)

Redundancy comes from the fact that *different symbols in $\mathcal{S}$ take different probabilities to be drawn*.

With a random source model, immediately there are two approaches one can take to demonstrate data compression:

- Allow variable codeword length for different symbols with different probabilities, rather than fixing it to be k.
- Allow (almost) lossless reconstruction rather than exact recovery.

Focus of this lecture: (almost) lossless reconstruction.

(Almost) Lossless decoding criterion

Let the randomness kick in: allow non-exact recovery.

To be precise, turn the focus to finding the smallest possible $\mathrm{R} = \frac{k}{n}$ given that the *error probability*

$$\mathsf{P}_{\mathsf{e}}^{(n)} := \mathsf{Pr}\{S^n \neq \hat{S}^n\} \to 0 \text{ as } n \to \infty.$$

Key features of this approach:

- Focus on the asymptotic regime where $n \to \infty$:
  Instead of error-free reconstruction, relaxed to vanishing error probability.
- The analysis is mainly *probabilistic*:
  The law of large numbers suffices for the basic version!

Outline

In this lecture, the focus is on Shannon's lossless source coding theorem and the corresponding information measure – (Shannon) entropy.

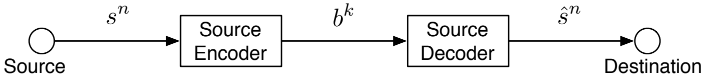
1. Lossless source coding theorem via typicality
2. Entropy and its properties
3. Extension to sources with memory

1. Representing an i.i.d. sequence almost losslessly
    - Typicality and AEP
    - Lossless Source Coding Theorem

2. Entropy: definition and properties
    - Definitions
    - Properties

3. Representing a sequence with memory almost losslessly

Finding a good fixed-length representation



Given a **discrete memoryless source** (DMS) $S \sim \mathsf{P}_S$, that is,

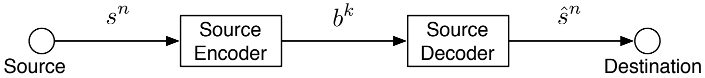
$$S_i \overset{\text{i.i.d.}}{\sim} \mathsf{P}_S,\ \forall\, i = 1, 2, \ldots$$

we ask for **lossless recovery** defined as follows: for a given $\epsilon \in (0, 1)$,

$$\Pr\{\hat{S}^n \neq S^n\} \leq \epsilon.$$

Equivalently, $\Pr\{\hat{S}^n = S^n\} \geq 1 - \epsilon$.

Warm-up: design of the encoder and the decoder



Encoder: a function

$$\mathrm{enc} : \mathcal{S}^n \to \{0,1\}^k$$

that maps each source sequence $s^n \in \mathcal{S}^n$ to a bit sequence $b^k \in \{0,1\}^k$.

Since we want to achieve compression, there exists $b^k \in \{0,1\}^k$ such that multiple s^n's are mapped to it.

Decoder: a function

$$\mathrm{dec} : \{0,1\}^k \to \mathcal{S}^n$$

that maps each bit sequence $b^k \in \{0,1\}^k$ to a source sequence $s^n \in \mathcal{S}^n$ such that $\mathrm{enc}(s^n) = b^k$.

Let $\mathcal{B}^{(n)} \subsetneq \mathcal{S}^n$ denote the range of the decoding function. The probability of successful reconstruction turns out to be

$$\mathsf{Pr}\{\hat{S}^n = S^n\} = \mathsf{Pr}\{S^n \in \mathcal{B}^{(n)}\}.$$

If our concern is on the probability of successful reconstruction, the design of the encoder-decoder pair is equivalent to design the set $\mathcal{B}^{(n)}$, that is, to select a set of 2^k source sequences and give them unique identifiers.

The rest of source sequences do not matter.

The overall design problem can be stated as

> *Given $\epsilon > 0$ and n, find the smallest $\mathcal{B}^{(n)} \subseteq \mathcal{S}^n$ such that*

$$\mathsf{Pr}\{S^n \in \mathcal{B}^{(n)}\} \ge 1 - \epsilon.$$

High-probability set

For a given $\epsilon \in (0,1)$, we say $\mathcal{B}(n,\epsilon) \subseteq \mathcal{S}^n$ is an ϵ-high-probability set iff

$$\mathsf{Pr}\{S^n \in \mathcal{B}(n,\epsilon)\} = \sum_{s^n \in \mathcal{B}(n,\epsilon)} \left(\prod_{i=1}^{n} \mathsf{P}_S(s_i) \right) \geq 1 - \epsilon.$$

The goal is to find the smallest size of such sets $\mathcal{B}(n,\epsilon)$.

An optimal way is to greedily choose the high-probability sequences (how?).

However, it remains difficult to analytically find the optimal compression rate

$$\frac{k}{n} = \frac{1}{n} \log_2 |\mathcal{B}(n,\epsilon)|.$$

The greedy algorithm does not tell much insight as $n \to \infty$ either.

Weak law of large numbers hints a simple way to construct these sets.

Recall the weak law of large numbers:

Weak Law of Large Numbers (Khinchin)

For a sequence of i.i.d. RVs $X_1, X_2, \ldots$ with $\mathsf{E}[|X_i|] < \infty$,

$$\lim_{n\to\infty} \mathsf{Pr}\{|\bar{X}_n - \mu| \ge \delta\} = 0 \quad \forall\, \delta > 0,$$

where $\mu = \mathsf{E}[X_i]$ denotes the mean.

By the WLLN, for a mapping $f : \mathcal{S} \to \mathbb{R}$ with $\mathsf{E}_{S\sim \mathsf{P}_S}[|f(S)|] < \infty$, for a memoryless source $S_i \overset{\text{i.i.d.}}{\sim} \mathsf{P}_S$,

$$\lim_{n\to\infty} \mathsf{Pr}\left\{\left|\tfrac{1}{n}\textstyle\sum_{i=1}^{n} f(S_i) - \mathsf{E}[f(S)]\right| > \delta\right\} = 0 \quad \forall\, \delta > 0.$$

Or equivalently, $\forall\, \delta > 0$, $\forall\, \epsilon \in (0,1)$, $\exists\, n_0(\epsilon) \in \mathbb{N}$ such that $\forall\, n \ge n_0(\epsilon)$,

$$\mathsf{Pr}\left\{\tfrac{1}{n}\textstyle\sum_{i=1}^{n} f(S_i) \in \big[\mathsf{E}[f(S)] - \delta, \mathsf{E}[f(S)] + \delta\big]\right\} \ge 1 - \epsilon.$$

In other words, if we can collect all the sequences s^n satisfying

$$\tfrac{1}{n}\textstyle\sum_{i=1}^{n} f(s_i) \in \big[\mathsf{E}[f(S)] - \delta, \mathsf{E}[f(S)] + \delta\big],$$

then for n large enough, it is an ϵ-high-probability set.

One additional thing we want from the ϵ-high-probability set defined by f:

A good estimate of its size.

If we can control the probability of each $s^n \in \mathcal{B}(n, \epsilon)$, we will be able to control its size (cardinality).

Why? Because its total probability is sandwiched between $1 - \epsilon$ and 1.

This gives some guidance on how to choose f:

We should choose f related to the probability of a sequence:

$$\Pr\{S^n = s^n\} = \textstyle\prod_{i=1}^{n} \mathsf{P}_S(s_i).$$

The definition of typical sequence emerges.

1 Representing an i.i.d. sequence almost losslessly
- Typicality and AEP
- Lossless Source Coding Theorem

2 Entropy: definition and properties
- Definitions
- Properties

3 Representing a sequence with memory almost losslessly

Overview of typicality methods

Goal: Understand and exploit the probabilistic asymptotic properties of an i.i.d. randomly generated sequence S^n, so as to learn how to represent it in the most efficient way.

Key Observation: When $n \to \infty$, one often observe that a substantially small set of sequences become "typical", which contribute almost the whole probability, while others become "atypical".

For lossless reconstruction with vanishing error probability, we can use shorter codewords to label "typical" sequences and ignore "atypical" ones.

We will show that the set of typical sequences (defined later), called the typical set, is "essentially" the smallest high-probability set.

For lossless source coding, this asymptotic size determines the optimal compression rate.

Question: how to choose $f : \mathcal{S} \to \mathbb{R}$ so that for a sequence $s^n = \{s_i\}_{i=1}^n$, as long as we know that

$$\left| \frac{1}{n} \sum_{i=1}^{n} f(s_i) - \mathsf{E}[f(S)] \right| \le \delta,$$

we can “control” the probability of s^n, $\mathsf{Pr}\{S^n = s^n\} = \prod_{i=1}^n \mathsf{P}_S(s_i)$?

Answer: quite intuitively, we can choose

$$f : s \mapsto \log \mathsf{P}_S(s).$$

For notational convenience, we define for a random variable $S \sim \mathsf{P}_S$,

$$\mathrm{H}(S) := -\mathsf{E}_{S \sim \mathsf{P}_S} \left[\log \mathsf{P}_S(S) \right].$$

The reason why there is a negative sign will become clear soon.

Typical sequence

Definition 1 (Typical Sequence)

For $\delta > 0$, a sequence s^n is called δ-typical with respect to r.v. $S \sim \mathsf{P}_S$ if

$$\left| \frac{1}{n} \sum_{i=1}^{n} \log \mathsf{P}_S(s_i) + \mathrm{H}(S) \right| \le \delta,$$

The δ-typical set $\mathcal{A}_\delta^{(n)}(S) := \{s^n \in \mathcal{S}^n \mid s^n$ is δ-typical with respect to $S\}$.

By definition, $\forall\, s^n \in \mathcal{A}_\delta^{(n)}(S)$,

$$2^{-n(\mathrm{H}(S)+\delta)} \le \underbrace{\mathsf{Pr}\{S^n = s^n\}}_{=\prod_{i=1}^{n} \mathsf{P}_S(s_i)} \le 2^{-n(\mathrm{H}(S)-\delta)}.$$

Properties of typical sequences

Proposition 1 (Properties of Typical Sequences and Typical Set)

1. $\forall\, s^n \in \mathcal{A}_\delta^{(n)}(S)$, $2^{-n(\mathrm{H}(S)+\delta)} \le \mathsf{Pr}\{S^n = s^n\} \le 2^{-n(\mathrm{H}(S)-\delta)}$.

   (by definition of typical sequences and entropy)

2. $\mathsf{Pr}\left\{S^n \in \mathcal{A}_\delta^{(n)}(S)\right\} \ge 1-\epsilon$ *for n large enough.*

   (by the weak law of large numbers (WLLN))

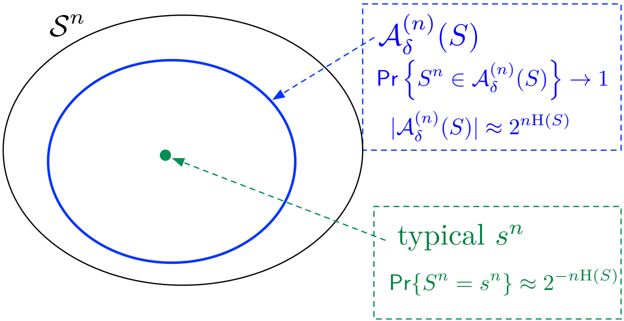
3. $|\mathcal{A}_\delta^{(n)}(S)| \le 2^{n(\mathrm{H}(S)+\delta)}$.

   (by summing up the lower bound in property 1 over the typical set)

4. $|\mathcal{A}_\delta^{(n)}(S)| \ge (1-\epsilon)2^{n(\mathrm{H}(S)-\delta)}$ *for n large enough.*

   (by the upper bound in property 1, and property 2)

Asymptotic equipartition property (AEP)



$\mathcal{S}^n$

$\mathcal{A}_\delta^{(n)}(S)$

$\mathsf{Pr}\left\{S^n \in \mathcal{A}_\delta^{(n)}(S)\right\} \to 1$

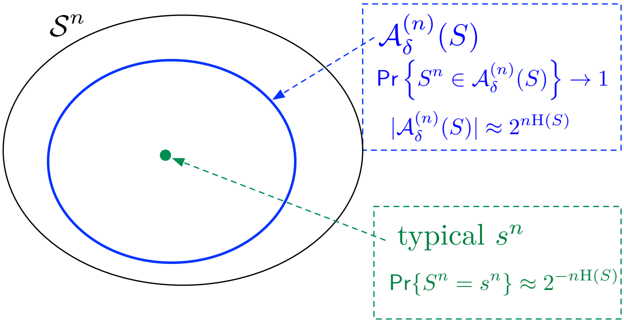
$|\mathcal{A}_\delta^{(n)}(S)| \approx 2^{n\mathrm{H}(S)}$

typical s^n

$\mathsf{Pr}\{S^n = s^n\} \approx 2^{-n\mathrm{H}(S)}$

- The typical set has probability approaching 1 as $n \to \infty$, while its size is roughly equal to $2^{n\mathrm{H}(S)}$, significantly smaller than $|\mathcal{S}^n| = 2^{n\log|\mathcal{S}|}$.
- All typical sequences have roughly the same probability $2^{-n\mathrm{H}(S)}$.

Application to data compression


$\mathcal{S}^n$
$\mathcal{A}_\delta^{(n)}(S)$
Pr $\left\{S^n \in \mathcal{A}_\delta^{(n)}(S)\right\} \to 1$
$|\mathcal{A}_\delta^{(n)}(S)| \approx 2^{n\mathrm{H}(S)}$
typical s^n
Pr$\{S^n = s^n\} \approx 2^{-n\mathrm{H}(S)}$

As $n \to \infty$: (1) the realization of the DMS is δ-typical with probability $\to 1$, (2) typical sequences are roughly uniformly distributed over the typical set, and (3) there are roughly $2^{n\mathrm{H}(S)}$ of them.

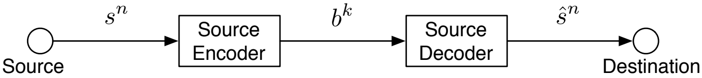
$\Longrightarrow$ Use roughly $n\mathrm{H}(S)$ bits to uniquely describe each typical sequence!

1 Representing an i.i.d. sequence almost losslessly
- Typicality and AEP
- Lossless Source Coding Theorem

2 Entropy: definition and properties
- Definitions
- Properties

3 Representing a sequence with memory almost losslessly

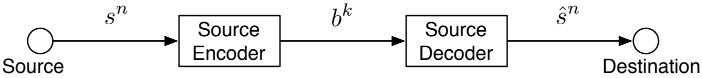
Lossless source coding: problem setup



1. An (n, k) source code consists of an encoding and a decoding function
   - an encoding function (encoder) $\text{enc}_n : \mathcal{S}^n \to \{0, 1\}^k$ that maps each source sequence s^n to a bit sequence b^k.
   - a decoding function (decoder) $\text{dec}_n : \{0, 1\}^k \to \mathcal{S}^n$ that maps each bit sequence b^k to a reconstructed source sequence $\hat{s}^n$.
2. An (n, k) code with $\mathsf{P}_\mathsf{e}^{(n)} := \mathsf{Pr}\{S^n \neq \hat{S}^n\} \leq \epsilon$ is called an (n, k, ϵ) code.
3. Let $k^*(n, \epsilon)$ denote the smallest k such that there exists an (n, k, ϵ) code. The optimal ϵ-achievable compression rate

$$\mathrm{R}^*(\epsilon) := \lim_{n \to \infty} \frac{k^*(n, \epsilon)}{n} \quad \text{if the limit exists.}$$

A lossless source coding theorem



Theorem 1 (A Lossless Source Coding Theorem for DMS)

For a DMS S,

$$\mathrm{R}^*(\epsilon) = \mathrm{H}(S) \quad \forall\, \epsilon \in (0,1).$$

To establish the coding theorem, since it is difficult to get a direct handle on $k^*(n,\epsilon)$, one can split the proof into two directions:

- Direct part (achievability): show that $\exists$ a sequence of (n,k,ϵ) codes such that for sufficiently large n, $\frac{k}{n} \le \mathrm{H}(S) + \delta$ for every $\delta > 0$.
- Converse part (converse): show that $\forall$ sequence of (n,k,ϵ) codes, for sufficiently large n, $\frac{k}{n} \ge \mathrm{H}(S) - \delta$ for every $\delta > 0$.

Achievability

pf: Here we provide a simple proof based on typical sequences (typicality).

As discussed earlier, the design of an (n, k, ϵ) code boils down to choosing a subset of length-n source sequences (the range of the decoding function) such that it has at least $1 - \epsilon$ probability.

AEP hints an obvious choice: pick any $\delta > 0$ and choose this subset as the δ-typical set $\mathcal{A}_\delta^{(n)}$.

- This choice gives a valid (n, k, ϵ) code for sufficiently large n by property 2 of Proposition 1.
- By property 3 of Proposition 1, the code rate

$$\tfrac{k}{n} = \tfrac{1}{n} \log_2 |\mathcal{A}_\delta^{(n)}| \le \mathrm{H}(S) + \delta.$$

This completes the proof of the direct part. □

Converse

pf: For a given (n, k, ϵ) code, let $\mathcal{B}^{(n)}$ denote the range of the decoding function. By definition,

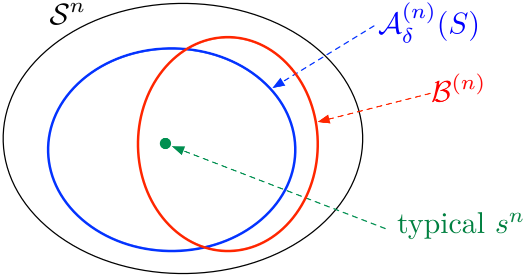
$$\mathsf{Pr}\{S^n \in \mathcal{B}^{(n)}\} \ge 1 - \epsilon.$$

To get an idea about the size of this set, one needs to get an idea about the probability of each sequence in this set.

We don't know how to do it for an *arbitrary* sequence, but we do know how to do it for a *typical* sequence.

So we turn our attention to those δ'-typical sequences in $\mathcal{B}^{(n)}$ and see if we can lower bound its probability

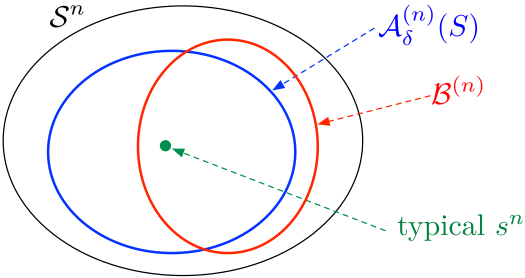
$$\mathsf{Pr}\{S^n \in \mathcal{B}^{(n)} \cap \mathcal{A}_{\delta'}^{(n)}\}.$$



It is not very difficult because for sufficiently large n, $\mathcal{A}_{\delta'}^{(n)}$ contains almost the whole probability: for a given $\epsilon' \in (0,1)$,

$$\begin{aligned}&\Pr\{S^n \in \mathcal{B}^{(n)} \cap \mathcal{A}_{\delta'}^{(n)}\}\\ &= \Pr\{S^n \in \mathcal{B}^{(n)}\} + \Pr\{S^n \in \mathcal{A}_{\delta'}^{(n)}\} - \Pr\{S^n \in \mathcal{B}^{(n)} \cup \mathcal{A}_{\delta'}^{(n)}\}\\ &\geq 1-\epsilon+1-\epsilon'-1 = 1-\epsilon-\epsilon'\end{aligned}$$

for sufficiently large n by property 2 of Proposition 1.



Meanwhile,

$$\Pr\{S^n \in \mathcal{B}^{(n)} \cap \mathcal{A}_{\delta'}^{(n)}\} \le |\mathcal{B}^{(n)} \cap \mathcal{A}_{\delta'}^{(n)}| 2^{-n\left(\mathrm{H}(S)-\delta'\right)} \le |\mathcal{B}^{(n)}| 2^{-n\left(\mathrm{H}(S)-\delta'\right)}$$

by property 1 of Proposition 1. Hence, for sufficiently large n,

$$\tfrac{k}{n} = \tfrac{1}{n}\log_2|\mathcal{B}^{(n)}| \ge \mathrm{H}(S) - \delta' - \tfrac{1}{n}\log_2 \tfrac{1}{1-\epsilon-\epsilon'}.$$

Proof of the converse part is complete by choosing δ' and ϵ' properly. □

Summary

- Shannon's lossless source coding theorem says:

  The **entropy** of a discrete random variable (or discrete probability) is the fundamental rate of efficient representation of sequences generated by a DMS following that probability law:

$$\mathrm{R}^*(\epsilon) = \mathrm{H}(S) \quad \forall\, \epsilon \in (0,1).$$

- Asymptotic Equipartition Property (AEP):

  Entropy determines the asymptotic size of a typical set, and determines the probability of a typical sequence asymptotically.

Next: an intuitive way to understand why entropy can measure the amount of information, formal definitions, and important properties.

1 Representing an i.i.d. sequence almost losslessly
- Typicality and AEP
- Lossless Source Coding Theorem

2 Entropy: definition and properties
- Definitions
- Properties

3 Representing a sequence with memory almost losslessly

How to measure information?

Before this, we should ask:

What is information?

Information is often obtained by learning something unknown before.

Examples: result of a ball game, score of an exam, weather, …

In other words, one gets some information by learning something about which that he/she was **uncertain** before.

> Shannon: "**Information is the resolution of uncertainty.**"

Motivating example

Let us take the following example:

- Suppose there is a professional basketball (NBA) final and a tennis tournament (the French Open quarterfinals) happening right now.
- An enthusiastic sports fan is interested in who will win the NBA final and who will win the Men's single.
- However, due to work, he/she cannot access any news in 10 days.
- How much information can the sports fan get after 10 days when he/she learns the two pieces of news (the two messages)?

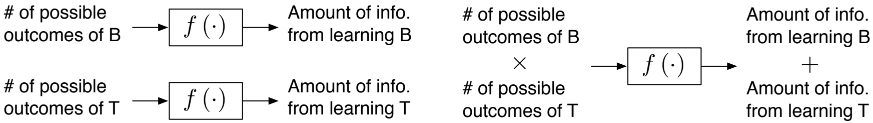
For the NBA final, the fan will learn that **one of the two** teams eventually wins the final (message B).

For the French Open quarterfinals, the fan will learn that **one of the eight** players eventually wins the gold medal (message T).

Observations

1. The amount of information is related to the number of possible outcomes: message B is a result of two possible outcomes, while message T is a result of eight possible outcomes.
2. The amount of information obtained in learning the two messages should be additive, while the number of possible outcomes is multiplicative.

Let $f(\cdot)$ be a function that measures the amount of information:



# of possible outcomes of B → $f(\cdot)$ → Amount of info. from learning B

# of possible outcomes of T → $f(\cdot)$ → Amount of info. from learning T

# of possible outcomes of B × # of possible outcomes of T → $f(\cdot)$ → Amount of info. from learning B + Amount of info. from learning T

What function produces additive outputs with multiplicative inputs?

Logarithmic Function

Logarithm as the measure of information

Initial guess of the measure of information: $\log($# of possible outcomes$)$.

[Hartley, 1928]

However, this measure does not take the likeliness into account – if some outcome occurs with very high probability, the amount of information of that outcome should be very little.

For example, suppose the fan knows the Warriors was leading the Cavs 3:1

- The probability that the Cavs win the final: $\frac{1}{2} \to \frac{1}{8}$.

  The Cavs win the final (w.p. $\frac{1}{8}$): it is like out of 8 times there is only 1 time that will generate this outcome

  $\Longrightarrow$ the amount of information $= \log 8 = 3$ bits.

- The probability that the Warriors win the final: $\frac{1}{2} \to \frac{7}{8}$.

  The Warriors win the final (w.p. $\frac{7}{8}$): it is like out of $\frac{8}{7}$ times there is only 1 time that will generate this outcome

  $\Longrightarrow$ the amount of information $= \log \frac{8}{7} = 3 - \log 7$ bits.

Information and uncertainty

From the motivation, we collect the following intuitions:

1. The amount of information is related to the # of possible outcomes
2. The measure of information should be additive
3. The measure of information should take the likeliness into account
4. measure of information = amount of uncertainty of an unknown outcome

Hence, a *plausible* measure of information of a realization x drawn from a random outcome X is

$$\boxed{f(x) := \log \frac{1}{\Pr\{X=x\}}}.$$

Correspondingly, the measure of information of a random outcome X is the averaged value of $f(x)$:

$$\boxed{\mathsf{E}_X\left[f(X)\right]}.$$

(The logarithms are of base 2 if not specified.)

Measure of uncertainty of a random variable

$$\boxed{\log \frac{1}{\mathsf{Pr}\left\{X = x\right\}} : \text{ measure of information/uncertainty of an outcome } x.}$$

If the outcome has small probability, it contains higher uncertainty. However, on the average, it happens rarely.

Hence, to measure the uncertainty of a *random variable*, we should take the expectation of the self information over all possible realizations. This leads to the following definition.

Definition 2 (Entropy for a Random Variable)

The entropy of a (discrete) random variable $X \in \mathcal{X}$ with probability mass function $\mathsf{P}_X(\cdot)$ is defined as

$$\mathrm{H}(X) := \mathsf{E}_X\left[\log\frac{1}{\mathsf{P}_X(X)}\right] = \sum_{x\in\mathcal{X}} \mathsf{P}_X(x)\log\frac{1}{\mathsf{P}_X(x)}.$$

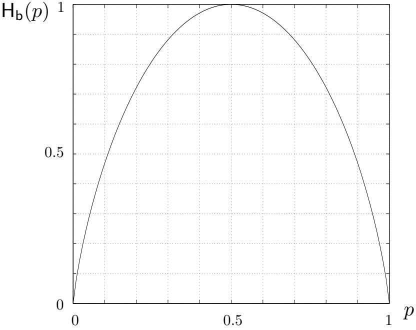
(by convention we set $0\log(1/0) = 0$ since $\lim_{t\to 0} t\log t = 0$.)

Note: Entropy can be understood as the (average) amount of information when one learns the actual outcome/realization of r.v. X.

Example 1 (Binary entropy function)

Let $X \sim \mathrm{Ber}(p)$ be a Bernoulli r.v., that is, $X \in \{0, 1\}$, $\mathsf{P}_X(1) = 1 - \mathsf{P}_X(0) = p$. Then, the entropy of X is called the binary entropy function $\mathsf{H}_\mathsf{b}(p)$, where

$$\mathsf{H}_\mathsf{b}(p) := \mathrm{H}(X) = -p \log p - (1-p)\log(1-p),\ p \in [0, 1].$$



Exercise 1

1. Analytically check that

$$\max_{p \in [0,1]} \mathsf{H}_\mathsf{b}(p) = 1,$$

$$\arg\max_{p \in [0,1]} \mathsf{H}_\mathsf{b}(p) = 1/2.$$

2. Show that $\mathsf{H}_\mathsf{b}(p)$ is concave in p.

Example 2

Consider a random variable $X \in \{0, 1, 2, 3\}$ with PMF defined as follows:

x	0	1	2	3
$\mathsf{P}_X(x)$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{6}$

Compute $\mathrm{H}(X)$ and $\mathrm{H}(Y)$, where $Y := X \mod 2$.

sol:

$$\mathrm{H}(X) = 2 \times \tfrac{1}{6} \times \log 6 + 2 \times \tfrac{1}{3} \times \log 3 = \tfrac{1}{3} + \log 3.$$

$$\mathrm{H}(Y) = 2 \times \tfrac{1}{2} \times \log 2 = 1.$$

1. Representing an i.i.d. sequence almost losslessly
   - Typicality and AEP
   - Lossless Source Coding Theorem
2. Entropy: definition and properties
   - Definitions
   - Properties
3. Representing a sequence with memory almost losslessly

Entropy: definition

Initially we define entropy for a random variable; it is straightforward to extend this definition to a sequence of random variables, or, a random vector.

The entropy of a random vector is also called the joint entropy of the component random variables.

Definition 3 (Entropy)

The entropy of a d-dimensional random vector $\boldsymbol{X} := (X_1, ..., X_d)$ is defined by the expectation of the self information

$$\mathrm{H}(\boldsymbol{X}) \equiv \mathrm{H}(X_1, \ldots, X_d) := \mathsf{E}_{\boldsymbol{X}}\left[\log \frac{1}{\mathsf{P}_{\boldsymbol{X}}(\boldsymbol{X})}\right] = \sum_{\boldsymbol{x} \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_d} \mathsf{P}_{\boldsymbol{X}}(\boldsymbol{x}) \log \frac{1}{\mathsf{P}_{\boldsymbol{X}}(\boldsymbol{x})}.$$

Remark: Entropy of a r.v. is a function of the distribution of the r.v.. Hence, we occasionally write $\mathrm{H}(P)$ and $\mathrm{H}(X)$ interchangeably for a discrete r.v. $X \sim P$.

Example 3

Consider two random variables $X_1, X_2 \in \{0,1\}$ with joint PMF

(x_1, x_2)	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
$\mathsf{P}_{X_1,X_2}(x_1,x_2)$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{6}$

Compute $\mathrm{H}(X_1)$, $\mathrm{H}(X_2)$, and $\mathrm{H}(X_1, X_2)$.

sol:

$$\mathrm{H}(X_1, X_2) = 2 \times \tfrac{1}{6} \times \log 6 + 2 \times \tfrac{1}{3} \times \log 3 = \tfrac{1}{3} + \log 3.$$

$$\mathrm{H}(X_1) = 2 \times \left(\tfrac{1}{3} + \tfrac{1}{6}\right) \times \log \frac{1}{\frac{1}{3}+\frac{1}{6}} = 1 = \mathrm{H}(X_2)\,.$$

Compared to Example 2, it can be understood that the value of entropy only depends on the distribution of the random variable/vector, not on the actual values it may take.

Conditional entropy

For two random variables with conditional PMF $\mathsf{P}_{X|Y}(x|y)$, we are able to define "the entropy of X given $Y = y$" according to $\mathsf{P}_{X|Y}(\cdot|y)$:

$$\mathrm{H}(X|Y = y) := \sum_{x \in \mathcal{X}} \mathsf{P}_{X|Y}(x|y) \log \frac{1}{\mathsf{P}_{X|Y}(x|y)}.$$

$\mathrm{H}(X|Y = y)$: the amount of uncertainty of X when we know that $Y = y$.

Averaging over Y, we obtain the amount of uncertainty of X given Y:

Definition 4 (Conditional Entropy)

The conditional entropy of X given Y is defined by

$$\begin{aligned}\mathrm{H}(X|Y) &:= \textstyle\sum_{y \in \mathcal{Y}} \mathsf{P}_Y(y) \mathrm{H}(X|Y = y) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \mathsf{P}_{X,Y}(x, y) \log \frac{1}{\mathsf{P}_{X|Y}(x|y)} \\ &= \mathsf{E}_{X,Y}\left[\log \frac{1}{\mathsf{P}_{X|Y}(X|Y)}\right].\end{aligned}$$

Example 4

Consider two random variables $X_1, X_2 \in \{0,1\}$ with joint PMF

(x_1, x_2)	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
$\mathsf{P}_{X_1,X_2}(x_1,x_2)$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{6}$

Compute $\mathrm{H}(X_1|X_2=0)$, $\mathrm{H}(X_1|X_2=1)$, $\mathrm{H}(X_1|X_2)$, and $\mathrm{H}(X_2|X_1)$.

sol:

(x_1, x_2)	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
$\mathsf{P}_{X_1\|X_2}(x_1\|x_2)$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{1}{3}$
$\mathsf{P}_{X_2\|X_1}(x_2\|x_1)$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{1}{3}$

$$\mathrm{H}(X_1|X_2=0) = \frac{1}{3}\log 3 + \frac{2}{3}\log\frac{3}{2} = \mathsf{H_b}\left(\frac{1}{3}\right),$$
$$\mathrm{H}(X_1|X_2=1) = \frac{2}{3}\log\frac{3}{2} + \frac{1}{3}\log 3 = \mathsf{H_b}\left(\frac{1}{3}\right).$$
$$\mathrm{H}(X_1|X_2) = 2\times\frac{1}{6}\times\log 3 + 2\times\frac{1}{3}\times\log\frac{3}{2} = \mathsf{H_b}\left(\frac{1}{3}\right) = \log 3 - \frac{2}{3} = \mathrm{H}(X_2|X_1)$$

1 Representing an i.i.d. sequence almost losslessly

- Typicality and AEP
- Lossless Source Coding Theorem

2 Entropy: definition and properties

- Definitions
- Properties

3 Representing a sequence with memory almost losslessly

Properties of entropy

Theorem 2 (Properties of (Joint) Entropy)

1. $\mathrm{H}(X) \geq 0$, *with equality iff X is deterministic.*
2. $\mathrm{H}(X) \leq \log|\mathcal{X}|$, *with equality iff X is uniformly distributed over $\mathcal{X}$.*
3. $\mathrm{H}(\boldsymbol{X}) \geq 0$, *with equality iff $\boldsymbol{X}$ is deterministic.*
4. $\mathrm{H}(\boldsymbol{X}) \leq \sum_{i=1}^{d} \log|\mathcal{X}_i|$, *with equality iff $\boldsymbol{X} \sim \mathrm{Unif}(\mathcal{X}_1 \times \cdots \times \mathcal{X}_d)$.*

Interpretation: Quite natural:

- Amount of uncertainty in $X = 0 \iff X$ is deterministic.
- Amount of uncertainty in X is maximized
  $\iff X$ is equally likely to take every value in $\mathcal{X}$.

Lemma 1 (Jensen's inequality)

$f : \mathbb{R} \to \mathbb{R}$ be a strictly concave function, and X be a real-valued r.v.. Then, $\mathsf{E}\left[f(X)\right] \le f\left(\mathsf{E}\left[X\right]\right)$, with equality iff X is deterministic.

We are ready to prove that $\mathrm{H}(X) \le \log|\mathcal{X}|$, with equality iff $X \sim \mathrm{Unif}(\mathcal{X})$.

pf: Recall the support of X, supp_X, denote the subset of $\mathcal{X}$ where X takes non-zero probability. Define a new r.v. $U := \frac{1}{\mathsf{P}_X(X)}$. Note that $\mathsf{E}\left[U\right] = |\mathsf{supp}_X|$. Hence,

$$\mathrm{H}(X) = \mathsf{E}\left[\log U\right] \overset{(\text{Jensen})}{\le} \log\left(\mathsf{E}\left[U\right]\right) = \log|\mathsf{supp}_X| \le \log|\mathcal{X}|.$$

The first inequality holds with equality iff U is deterministic iff $\forall\, x \in \mathsf{supp}_X$, $\mathsf{P}_X(x)$ are equal. The second inequality holds with equality iff $\mathsf{supp}_X = \mathcal{X}$. □

Exercise 2

For jointly distributed (X, Y), show that $\mathrm{H}(X|Y) \ge 0$ with equality if and only if X is a function of Y.

Chain rule

Theorem 3 (Chain Rule)

$$\mathrm{H}(X,Y) = \mathrm{H}(Y) + \mathrm{H}(X|Y) = \mathrm{H}(X) + \mathrm{H}(Y|X).$$

Interpretation:

Amount of uncertainty of (X,Y) =
Amount of uncertainty of Y + Amount of uncertainty of X after knowing Y.

pf: By definition,

$$\begin{aligned}\mathrm{H}(X,Y) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \mathsf{P}(x,y)\log\tfrac{1}{\mathsf{P}(x,y)} = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \mathsf{P}(x,y)\log\tfrac{1}{\mathsf{P}(y)\mathsf{P}(x|y)}\\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \mathsf{P}(x,y)\log\tfrac{1}{\mathsf{P}(y)} + \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \mathsf{P}(x,y)\log\tfrac{1}{\mathsf{P}(x|y)}\\ &= \mathrm{H}(Y) + \mathrm{H}(X|Y) \qquad \square\end{aligned}$$

(when the context is clear, we drop the subscripts in P_X, P_Y, $\mathsf{P}_{Y|X}$, etc.)

Conditioning reduces entropy

Theorem 4 (Conditioning Reduces Entropy)

$\mathrm{H}(X|Y) \le \mathrm{H}(X)$, *with equality iff* X *is independent of* Y*.*

Interpretation: The more one learns, the less the uncertainty is.

The amount of uncertainty of your target remains the same if and only if what you have learned is independent of your target.

Exercise 3

While it is always true that $\mathrm{H}(X|Y) \le \mathrm{H}(X)$, for $y \in \mathcal{Y}$, the following two are both possible:

- $\mathrm{H}(X|Y=y) < \mathrm{H}(X)$, or
- $\mathrm{H}(X|Y=y) > \mathrm{H}(X)$.

Please construct examples for the above two cases respectively.

pf: By definition and Jensen's inequality, we have

$$\begin{aligned}
&\mathrm{H}(X|Y) - \mathrm{H}(X) \\
&= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \mathsf{P}(x,y) \log \tfrac{\mathsf{P}(x)}{\mathsf{P}(x|y)} = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \mathsf{P}(x,y) \log \tfrac{\mathsf{P}(x)\mathsf{P}(y)}{\mathsf{P}(x,y)} \\
&\le \log\left(\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \mathsf{P}(x,y) \tfrac{\mathsf{P}(x)\mathsf{P}(y)}{\mathsf{P}(x,y)} \right) = \log\left(\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \mathsf{P}(x)\mathsf{P}(y) \right) \\
&= \log(1) = 0
\end{aligned}$$

□

Example 5

Consider two random variables $X_1, X_2 \in \{0,1\}$ with joint PMF

(x_1, x_2)	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
$\mathsf{P}_{X_1,X_2}(x_1,x_2)$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{6}$

In the previously examples, we have

$$\begin{aligned} &\mathrm{H}(X_1, X_2) = \log 3 + \tfrac{1}{3},\ \mathrm{H}(X_1) = \mathrm{H}(X_2) = 1, \\ &\mathrm{H}(X_1|X_2) = \mathrm{H}(X_2|X_1) = \log 3 - \tfrac{2}{3}. \end{aligned}$$

It is straightforward to check that the chain rule holds. Besides, it can be easily seen that conditioning reduces entropy.

Generalization

Proofs of the more general “Chain Rule” and “Conditioning Reduces Entropy” are left as exercises.

Theorem 5 (Chain Rule)

The chain rule can be generalized to more than two r.v.’s:

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

Theorem 6 (Conditioning Reduces Entropy)

Conditioning reduces entropy can be generalized to more than two r.v.’s:

$$\mathrm{H}(X|Y, Z) \le \mathrm{H}(X|Y).$$

Upper bound on joint entropy

Corollary 1 (Joint Entropy $\leq$ Sum of Marginal Entropies)

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

Proof is left as exercise (chain rule of entropy + conditioning reduces entropy).

Exercise 4

Show that

$$\mathrm{H}(X, Y, Z) \leq \mathrm{H}(X, Y) + \mathrm{H}(X, Z) - \mathrm{H}(X) \,.$$

Concavity of entropy

Theorem 7 (Concavity of Entropy)

Let $\boldsymbol{p} := (p_1, ..., p_d)$ denote the PMF vector of a random variable X. Then, the entropy of X, $\mathrm{H}(\boldsymbol{p})$, is concave in $\boldsymbol{p}$, where $\mathrm{H}(\boldsymbol{p}) := -\sum_{i=1}^{d} p_i \log p_i$. *(written as $\mathrm{H}(\boldsymbol{p})$ since it is a function of $\boldsymbol{p}$)*

pf: We would like to show that for any $\lambda \in [0,1]$, $\overline{\lambda} = 1 - \lambda$,

$$\mathrm{H}\left(\lambda \boldsymbol{p}_1 + \overline{\lambda} \boldsymbol{p}_2\right) \ge \lambda \mathrm{H}(\boldsymbol{p}_1) + \overline{\lambda} \mathrm{H}(\boldsymbol{p}_2).$$

Setting $X_1 \sim \boldsymbol{p}_1$, $X_2 \sim \boldsymbol{p}_2$, and Θ with $\Pr\{\Theta = 1\} = \lambda = 1 - \Pr\{\Theta = 2\}$, since conditioning reduces entropy, we have

$$\mathrm{H}(X_\Theta) \ge \mathrm{H}(X_\Theta | \Theta).$$

Done by $X_\Theta \sim \boldsymbol{p}_\lambda := \lambda \boldsymbol{p}_1 + \overline{\lambda} \boldsymbol{p}_2$ and $\mathrm{H}(X_\Theta | \Theta = i) = \mathrm{H}(\boldsymbol{p}_i)$, $i = 1, 2$. □

(Recall: we often use $\boldsymbol{p}$ and $\mathsf{P}(\cdot)$ interchangeably to denote a PMF (vectors))

Fano's inequality

Lemma 2 (Fano's Inequality)

$\mathrm{H}(U|V) \le \mathsf{H}_\mathsf{b}(\mathsf{P}_\mathsf{e}) + \mathsf{P}_\mathsf{e} \log|\mathcal{U}|$, *where* $\mathsf{P}_\mathsf{e} \triangleq \mathsf{Pr}\{U \neq V\}$.

pf: Let $E := \mathbb{1}\{U \neq V\}$, the indicator function of $\{U \neq V\}$. $E \sim \mathrm{Ber}(\mathsf{P}_\mathsf{e})$.

Using chain rule and the non-negativity of conditional entropy, we have

$$\mathrm{H}(U|V) \le \mathrm{H}(U,E|V) = \mathrm{H}(E|V) + \mathrm{H}(U|V,E)\,.$$

Note that $\mathrm{H}(E|V) \le \mathrm{H}(E) = \mathsf{H}_\mathsf{b}(\mathsf{P}_\mathsf{e})$, and

$$\mathrm{H}(U|V,E) = \underbrace{\mathsf{Pr}\{E=1\}}_{=\mathsf{P}_\mathsf{e}}\underbrace{\mathrm{H}(U|V,E=1)}_{\le\log|\mathcal{U}|} + \mathsf{Pr}\{E=0\}\underbrace{\mathrm{H}(U|V,E=0)}_{=0,\ \because U=V}$$

Hence, $\mathrm{H}(U|V) \le \mathsf{H}_\mathsf{b}(\mathsf{P}_\mathsf{e}) + \mathsf{P}_\mathsf{e} \log|\mathcal{U}|$. □

Corollary 2 (Lower Bound on Error Probability)

$$\mathsf{P_e} \geq \frac{\mathrm{H}(U|V) - 1}{\log|\mathcal{U}|}.$$

pf: From Lemma 2 and $\mathsf{H_b}(\mathsf{P_e}) \leq 1$, we have

$$\mathrm{H}(U|V) \leq \mathsf{H_b}(\mathsf{P_e}) + \mathsf{P_e}\log|\mathcal{U}| \leq 1 + \mathsf{P_e}\log|\mathcal{U}|. \qquad \square$$

Note: If $\mathrm{H}(U|V)$ is close to $\log|\mathcal{U}|$, $\mathrm{H}(U|V)$ will also be close to $\mathrm{H}(U)$, and hence one can hardly determine U from V.

Exercise 5

Show that Lemma 2 can be sharpened as follows

$$\mathrm{H}(U|V) \leq \mathsf{H_b}(\mathsf{P_e}) + \mathsf{P_e}\log(|\mathcal{U}| - 1),$$

if U, V both take values in $\mathcal{U}$.

Summary

- Entropy $\mathrm{H}(X) := \mathsf{E}\left[\log \frac{1}{\mathsf{P}_X(X)}\right]$ measures the amount of uncertainty in X.
- Conditional entropy $\mathrm{H}(X|Y) := \mathsf{E}\left[\log \frac{1}{\mathsf{P}_{X|Y}(X|Y)}\right]$ measures the amount of uncertainty in X given Y.
- Conditioning reduces entropy: $\mathrm{H}(X|Y,Z) \le \mathrm{H}(X|Y)$.
- Chain rule: $\mathrm{H}(X,Y) = \mathrm{H}(Y) + \mathrm{H}(X|Y) = \mathrm{H}(X) + \mathrm{H}(Y|X)$.
- $0 \le \mathrm{H}(X) \le \log|\mathcal{X}|$ is maximized if X is uniformly distributed, and it is minimized if X is deterministic.
- Entropy, as a function of the discrete probability vector, is concave.
- Fano's inequality: $\mathsf{Pr}\{U \ne V\} \ge \frac{\mathrm{H}(U|V)-1}{\log|\mathcal{U}|}$.

1. Representing an i.i.d. sequence almost losslessly
   - Typicality and AEP
   - Lossless Source Coding Theorem

2. Entropy: definition and properties
   - Definitions
   - Properties

3. Representing a sequence with memory almost losslessly

Beyond memoryless sources

Recap: So far we have established the fundamental limit of representing a sequence generated by a discrete memoryless sources (DMS) almost losslessly in a fixed-to-fixed setting.

Key: a concentration property, AEP, of randomly generated sequences.

- Use typical sequences to construct a code for any rate $\mathrm{R} > \mathrm{H}(S)$ with probability of error arbitrarily close to 0.
- Use typical sequences to argue that if $\mathrm{R} < \mathrm{H}(S)$, the probability or error gets arbitrarily close to 1.
- Entropy $\mathrm{H}(S)$ is a measure of information.

Question: What if the source is not memoryless?

In other words, a single PMF P_S cannot describe the random process.

A source generates symbols that statistically depend on the past (memory).

It can be described as a (discrete-time) random process (r.p.) $\{S_i | i = 1, 2, ...\}$ consisting of an infinite sequence of r.v.'s.

Such a r.p. is characterized by all joint PMF's $\mathsf{P}_{S_1, S_2, \ldots, S_n}$, $\forall\, n = 1, 2, ...$.

AEP, definition of typical sequence, and information measure must change.

For sources with memory, we should develop the following two so that a lossless source coding theorem can be established:

1. A general AEP for *random processes*.
2. A measure of information for *random processes* called **entropy rate**.

Intuition:
sources with memory $\implies$ *more redundancy and can be compressed further.*

Typicality for sources with memory

Key to AEP: establish a WLLN-like property for general random processes: as $n \to \infty$,

$$\frac{1}{n} \log \frac{1}{\mathsf{P}(S^n)} \xrightarrow{\mathrm{p}} \mathcal{H}(\{S_i\}) \text{ (for now, some hypothetical quantity).}$$

$\mathsf{P}(s^n)$ is a short-hand notation for the joint PMF of S^n evaluated at s^n.

By defining typical sequences as those s^n sequences with "normalized self information"

$$\frac{1}{n} \log \frac{1}{\mathsf{P}(s^n)} \approx \mathcal{H}(\{S_i\}),$$

AEP immediately emerges, and consequently all the arguments in proving the DMS lossless source coding theorem can be applied to prove a corresponding one for sources with memory.

So, what should be this $\mathcal{H}(\{S_i\})$, if the above p-limit exists?

Guesses based on LLN-like structures

A guess:

$$\frac{1}{n}\log\frac{1}{\mathsf{P}(s^n)} = \frac{1}{n}\log\frac{1}{\mathsf{P}(S_1,S_2,\ldots,S_n)} \xrightarrow{\text{p}} \underbrace{\lim_{n\to\infty}\frac{1}{n}\mathsf{E}\left[\log\frac{1}{\mathsf{P}(S_1,S_2,\ldots,S_n)}\right]}_{\lim_{n\to\infty}\frac{1}{n}\mathrm{H}(S_1,S_2,\ldots,S_n)}$$

Interpretation: the *average* amount of uncertainty per symbol.

Another guess:

$$\frac{1}{n}\log\frac{1}{\mathsf{P}(S^n)} = \frac{1}{n}\sum_{i=1}^{n}\log\frac{1}{\mathsf{P}(S_i|S^{i-1})} \xrightarrow{\text{p}} \underbrace{\lim_{n\to\infty}\mathsf{E}\left[\log\frac{1}{\mathsf{P}(S_n|S^{n-1})}\right]}_{\lim_{n\to\infty}\mathrm{H}(S_n|S^{n-1})}$$

Interpretation: the *marginal* amount of uncertainty of the current symbol conditioned on all the past symbols.

Entropy rate

For a discrete r.p. $\{X_i | i \in \mathbb{N}\}$, how do we measure its uncertainty?

- For a single r.v. X, entropy $\mathrm{H}(X)$ measures its amount of uncertainty.
- Infinitely many r.v.'s in $\{X_i\}$, but $\mathrm{H}(X_1, X_2, ...)$ is meaningless. (likely to be ∞)
- We should measure the *average* amount of uncertainty per symbol!
- Or, we can measure the *marginal* amount of uncertainty of the current symbol conditioned on all the past symbols

Definition 5 (Entropy Rate)

Two definitions of the entropy rate of a random process $\{X_i\}$:

$$\mathcal{H}(\{X_i\}) := \lim_{n\to\infty} \frac{1}{n} \mathrm{H}(X_1, X_2, \ldots, X_n) \quad \text{if the limit exists} \tag{1}$$

$$\widetilde{\mathcal{H}}(\{X_i\}) := \lim_{n\to\infty} \mathrm{H}\left(X_n \middle| X^{n-1}\right) \quad \text{if the limit exists.} \tag{2}$$

Example 6 (Entropy Rate of i.i.d. Process)

Consider a random process $\{X_i\}$ where $X_1, X_2, \ldots$ are i.i.d. according to P_X. Does the entropy rate exist? If so, compute it.

sol: Since the r.v.'s are i.i.d., for all $n \in \mathbb{N}$,

$$\mathrm{H}(X_1, \ldots, X_n) = n\mathrm{H}(X_1)\,,\ \mathrm{H}\left(X_n \middle| X^{n-1}\right) = \mathrm{H}(X_n) = \mathrm{H}(X_1)\,.$$

Hence, $\mathcal{H}(\{X_i\}) = \widetilde{\mathcal{H}}(\{X_i\}) = \mathrm{H}(X_1) = \mathrm{H}(\mathsf{P}_X)$.

Exercise 6 ($\mathcal{H}$ and $\widetilde{\mathcal{H}}$ May be Different)

Consider a random process $\{X_i\}$ where $X_1, X_3, \ldots$ are i.i.d. and $X_{2k} = X_{2k-1}$ for all $k \in \mathbb{N}$. Show that $\mathcal{H}(\{X_i\})$ exists, but $\widetilde{\mathcal{H}}(\{X_i\})$ does not.

Two notions of entropy rate

In Definition 5, we have defined two notions of entropy rate: $\mathcal{H}$ and $\widetilde{\mathcal{H}}$.

In Exercise 6, we see that the two notions are not equivalent in general.

Question: when do they meet?

Lemma 3 (Cesàro Mean)

$\lim_{n\to\infty} b_n = c \implies \lim_{n\to\infty} a_n = c$, *where* $a_n := \frac{1}{n}\sum_{k=1}^{n} b_k$. *(the reverse is not true in general)*

As a corollary, if $\widetilde{\mathcal{H}}$ exists, so does $\mathcal{H}$ and $\mathcal{H} = \widetilde{\mathcal{H}}$.

Why? Let $a_n = \frac{1}{n}\mathrm{H}(X_1, \ldots, X_n)$ and $b_n = \mathrm{H}\left(X_n \middle| X^{n-1}\right)$:

- $a_n = \frac{1}{n}\sum_{k=1}^{n} b_k$ due to chain rule.
- $\mathcal{H}(\{X_i\}) = \lim_{n\to\infty} a_n$ and $\widetilde{\mathcal{H}}(\{X_i\}) = \lim_{n\to\infty} b_n$.

Next, we introduce two kinds of random processes where $\mathcal{H} = \widetilde{\mathcal{H}}$.

Entropy rate of a stationary process

Definition 6 (Stationary Random Process)

A random process $\{X_i\}$ is *stationary* if for all shift $l \in \mathbb{N}$,

$$\mathsf{P}_{X_1,X_2,\ldots,X_n} = \mathsf{P}_{X_{l+1},X_{l+2},\ldots,X_{l+n}},\ \forall\, n \in \mathbb{N}.$$

It turns out stationarity implies that the marginal amount of uncertainty conditioned on all the past is decreasing over time.

Lemma 4

For a stationary random process $\{X_i\}$, $\mathrm{H}\left(X_n \middle| X^{n-1}\right)$ is decreasing in n.

pf: Due to the fact that conditioning reduces entropy, we have

$$\mathrm{H}(X_{n+1}|X^n) = \mathrm{H}(X_{n+1}|X_2^n, X_1) \le \mathrm{H}(X_{n+1}|X_2^n)\,. \quad \text{(notation: } x_i^j \equiv (x_i, \ldots, x_j),\ i \le j\text{)}$$

Since $\{X_i\}$ is stationary, $\mathrm{H}(X_{n+1}|X_2^n) = \mathrm{H}\left(X_n \middle| X^{n-1}\right)$. □

Theorem 8

For a stationary random process $\{X_i\}$,

$$\mathcal{H}(\{X_i\}) = \widetilde{\mathcal{H}}(\{X_i\}).$$

pf: Since $b_n := \mathrm{H}\left(X_n \middle| X^{n-1}\right)$ is decreasing in n, and $b_n \geq 0$ is bounded from below, we conclude that b_n converges as $n \to \infty$.

Since $\frac{1}{n}\mathrm{H}(X_1, \ldots, X_n) = \frac{1}{n}\sum_{k=1}^{n} b_k$, by Lemma 3, proof complete. □

Exercise 7

Show that for a stationary random process $\{X_i\}$,

- $\frac{1}{n}\mathrm{H}(X_1, \ldots, X_n)$ is decreasing in n.
- $\mathrm{H}\left(X_n \middle| X^{n-1}\right) \leq \frac{1}{n}\mathrm{H}(X_1, \ldots, X_n)$.

LLN-like limiting behavior beyond sum of i.i.d. r.v.'s

Back to the attempt to get AEP for general r.p.'s beyond i.i.d. processes.

It turns out that the kind of random processes that have the LLN-like limiting behavior are stationary *ergodic* processes.

Roughly speaking, a stationary process $\{X_i\}$ is ergodic iff the time average (empirical average) converges to the ensemble average almost surely. More specifically, $\forall\, k_1, k_2, \ldots, k_m \in \mathbb{N}$, f measurable and absolutely integrable,

$$\frac{1}{n}\sum_{l=0}^{n-1} f\left(X_{k_1+l}, \ldots, X_{k_m+l}\right) \xrightarrow{\text{a.s.}, L^1} \mathsf{E}\left[f\left(X_{k_1}, \ldots, X_{k_m}\right)\right] \quad \text{as } n \to \infty.$$

This is the Birkhoff-Khinchin ergodic theorem, a main founding result in ergodic theory. It is essentially a law of large numbers for random processes.

Memoryless (i.i.d.) sources, stationary Markov sources, etc., are all special cases of stationary ergodic sources.

AEP for stationary ergodic processes

Theorem 9 (Shannon-McMillan-Breiman)

If $\mathcal{H}(\{S_i\})$ is the entropy rate of a stationary ergodic process $\{S_i\}$,

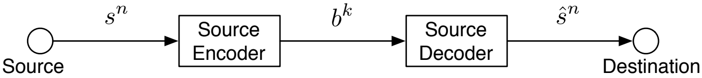
$$\frac{1}{n}\log\frac{1}{\mathsf{P}(S^n)} \xrightarrow{\text{a.s.},L^1} \mathcal{H}(\{S_i\}) \quad \textit{as } n\to\infty,$$

which implies convergence in probability.

The proof can be found in Chapter 16.8 of Cover and Thomas and other standard textbooks of probability theory. A main ingredient is a Markov approximation that allows the ergodic theorem to kick in.

Take-away: With the above theorem, we can re-define typical sequences as we did in the i.i.d. case, with the following substitution: $\mathrm{H}(S)\to\mathcal{H}(\{S_i\})$ and derive corresponding properties. As we discussed before, the four key properties in Proposition 1 remain the same and AEP remains to hold.

Lossless source coding theorem for ergodic DSS



Theorem 10 (A Lossless Source Coding Theorem for Ergodic DSS)

For a discrete stationary ergodic source $\{S_i\}$,

$$\mathrm{R}^*(\epsilon) = \mathcal{H}(\{S_i\}) \quad \forall\, \epsilon \in (0,1).$$

The proof is exactly the same as the DMS case except for the new definition of typical sequences.

Markov process

Markov process is one of the simplest random processes with memory.

Definition 7 (Markov Process)

$\{X_i \,|\, i = 1, 2, ...\}$ is a Markov process if $\forall\, n > 1$,

$$\mathsf{P}_{X_n | X_{n-1}, X_{n-2}, \ldots, X_1} = \mathsf{P}_{X_n | X_{n-1}}.$$

A typical convention: $X_1 - X_2 - \cdots - X_n - \cdots$.

The common alphabet $\mathcal{X}$ is called the *state space* of the Markov process.

Some further basic definitions of a Markov process are given below.

1. A Markov process is *irreducible* if $\forall\, x, y \in \mathcal{X}$, it is possible to reach to start at x and reach y in a finite number of steps.
2. The period of a state is the g.c.d. of the # of times that a state can return to itself. A Markov process is *aperiodic* if all states have period $= 1$.

3. A Markov process is *homogeneous* (or time-invariant) if $\forall\, n > 1$, $\mathsf{P}_{X_n|X_{n-1}} = \mathsf{P}_{X_2|X_1}$. Hence, a homogeneous Markov process is completely defined by its *initial state distribution* P_{X_1} and *transition probability* $\mathsf{P}_{X_2|X_1}$.

4. A *steady-state distribution* $\pi : \mathcal{X} \to [0,1]$ is one such that the distribution does not change after one transition:

   $$\pi(x) = \textstyle\sum_{y\in\mathcal{X}} \pi(y)\mathsf{P}_{X_{n+1}|X_n}(x|y),\ \forall\, x \in \mathcal{X},\ n \in \mathbb{N}.$$

   For a finite-alphabet homogeneous Markov process, steady-state distribution always exists, and it is unique if the process is irreducible.

5. For a finite-alphabet homogeneous Markov process that is both irreducible and aperiodic,

   $$\lim_{n\to\infty} \mathsf{Pr}\{X_{n+1} = y | X_1 = x\} = \pi(y),\ \forall\, x, y \in \mathcal{X},$$

   where $\pi(\cdot)$ is the unique steady-state distribution.

   If $\mathsf{P}_{X_1} = \pi$, the Markov process becomes a stationary process.

Entropy rate of a Markov process

Theorem 11

For a homogeneous, irreducible, and aperiodic Markov process $\{X_i\}$*,*

$$\mathcal{H}(\{X_i\}) = \widetilde{\mathcal{H}}(\{X_i\}) = \mathrm{H}(X_2|X_1)\,|_{\mathsf{P}_{X_1}=\pi} = \sum_{x\in\mathcal{X}} \pi(x)\mathrm{H}(X_2|X_1=x)\,.$$

where π *is the unique steady-state distribution.*

Remark: if $\{X_i\}$ is *stationary*, the entropy rate is simple to compute:

$$\mathcal{H}(\{X_i\}) = \widetilde{\mathcal{H}}(\{X_i\}) = \lim_{n\to\infty} \mathrm{H}\left(X_n\middle|X^{n-1}\right) \overset{\text{Markovity}}{=} \lim_{n\to\infty} \mathrm{H}(X_n|X_{n-1}) \overset{\text{Stationarity}}{=} \mathrm{H}(X_2|X_1)\,.$$

But in this theorem, we do not require the Markov process to be stationary.

pf: Again by Lemma 3 (Cesàro Mean), it suffices to show that $\widetilde{\mathcal{H}}(\{X_i\})$ exists and is equal to $\mathrm{H}(X_2|X_1)\,|_{\mathsf{P}_{X_1}=\pi}$, that is,

$$\lim_{n\to\infty} \mathrm{H}\left(X_n \middle| X^{n-1}\right) = \mathrm{H}(X_2|X_1)\,|_{\mathsf{P}_{X_1}=\pi}.$$

By Markovity, $\mathrm{H}\left(X_n \middle| X^{n-1}\right) = \mathrm{H}(X_n|X_{n-1})$, and we can expand it as

$$\begin{aligned} \mathrm{H}(X_n|X_{n-1}) &= \sum_{x\in\mathcal{X}} \mathsf{P}_{X_{n-1}}(x)\mathrm{H}(X_n|X_{n-1}=x) \\ &= \sum_{x\in\mathcal{X}} \mathsf{P}_{X_{n-1}}(x)\mathrm{H}(X_2|X_1=x) \qquad (\because \text{ homogeneity}) \end{aligned}$$

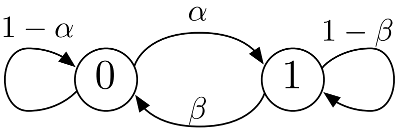
The proof is complete by observing $\lim_{n\to\infty} \mathsf{P}_{X_{n-1}}(x) = \pi(x)\ \forall\, x\in\mathcal{X}$. □

Example 7 (Two-State Markov Process)

Consider a stationary two-state Markov process $\{X_i \mid i \in \mathbb{N}\}$ taking values in $\{0, 1\}$ with probability transition matrix

$$\mathsf{P}_{X_2|X_1} = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix},$$

where $\alpha, \beta \in (0, 1)$. Find the marginal p.m.f. $\mathsf{P}_{X_n}(x)$ for all $n \in \mathbb{N}$ and the entropy rate $\mathrm{H}(\{X_i\})$.



$$\begin{array}{c|cc} & 0 & 1 \\ \hline 0 & 1-\alpha & \alpha \\ 1 & \beta & 1-\beta \end{array}$$

sol: The stationary distribution $\begin{bmatrix}\pi(0) & \pi(1)\end{bmatrix}$ of a Markov chain can be computed by solving the following linear equation:

$$\begin{bmatrix}\pi(0) & \pi(1)\end{bmatrix} = \begin{bmatrix}\pi(0) & \pi(1)\end{bmatrix}\begin{bmatrix}1-\alpha & \alpha \\ \beta & 1-\beta\end{bmatrix}$$

$$\implies \begin{bmatrix}\pi(0) & \pi(1)\end{bmatrix} = \begin{bmatrix}\frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta}\end{bmatrix} = \begin{bmatrix}\mathsf{P}_{X_n}(0) & \mathsf{P}_{X_n}(1)\end{bmatrix},\ \forall\, n \in \mathbb{N}.$$

Since $\mathcal{H}(\{X_i\})$ is equal to $\mathrm{H}(X_2|X_1)$, we can easily compute it as follows:

$$\begin{aligned}\mathcal{H}(\{X_i\}) &= \mathrm{H}(X_2|X_1) \\ &= \pi(0)\mathrm{H}(X_2|X_1=0) + \pi(1)\mathrm{H}(X_2|X_1=1) \\ &= \frac{\beta}{\alpha+\beta}\mathsf{H_b}(\alpha) + \frac{\alpha}{\alpha+\beta}\mathsf{H_b}(\beta).\end{aligned}$$

Summary

- Lossless source coding theorem: for ergodic DSS $\{S_i\}$,

$$\mathrm{R}^*(\epsilon) = \mathcal{H}(\{S_i\}) \quad \forall\, \epsilon \in (0,1).$$

- Asymptotic Equipartition Property (AEP) for ergodic DSS:

1. $\forall\, s^n \in \mathcal{A}_\delta^{(n)}(\{S_i\})$, $2^{-n(\mathcal{H}(\{S_i\})+\delta)} \le \mathsf{P}(s^n) \le 2^{-n(\mathcal{H}(\{S_i\})-\delta)}$.
2. $\mathsf{P}\left(\mathcal{A}_\delta^{(n)}(\{S_i\})\right) \ge 1-\epsilon$ for n large enough.
3. $|\mathcal{A}_\delta^{(n)}(\{S_i\})| \le 2^{n(\mathcal{H}(\{S_i\})+\delta)}$.
4. $|\mathcal{A}_\delta^{(n)}(\{S_i\})| \ge (1-\epsilon)2^{n(\mathcal{H}(\{S_i\})-\delta)}$ for n large enough.