LECTURE 3
TYPICAL SEQUENCES AND ASYMPTOTIC EQUIPARTITION PROPERTY (AEP)

Information Theory

Thomas Debris-Alazard

Inria, École Polytechnique

MOTIVATION

Lecture 2: how to compress i.i.d sources

by crucially using the informal concept of ***typical sequences***

- What is the essence of typical sequences?
- i.i.d. sources are not realistic *e.g* with languages, are there more general sources that we can compress by using typical sequences?

- To define formally the concept of typical sequences
- Showing that sources admitting typical sequences are those for which Shannon's source coding theorem holds (Lecture 2)
- Exhibiting more general sources that i.i.d. distributions verifying Shannon's theorem

COURSE OUTLINE

1. Entropy Rate of Stochastic Processes

2. **A**symptotic **E**quipartition **P**roperty (AEP)

   $\longrightarrow$ To define formally **typical sequences** and showing how to reach optimal compression!

   - Independent and Identically Distributed sources admit typical sequences
   - More general kind of sources verifying the AEP: **Markov chains**

SOME REMINDERS

SOME ABUSE OF NOTATION

Given random variables $\mathsf{X}_1, \ldots, \mathsf{X}_L$,

$$p(x_1, \ldots, x_L) \stackrel{\text{def}}{=} \mathbb{P}(\mathsf{X}_1 = x_1, \ldots, \mathsf{X}_L = x_L)$$

If it is clear from the context, given X and Y,

$$p(x) \stackrel{\text{def}}{=} \mathbb{P}(\mathsf{X} = x) \quad \text{and} \quad p(y) \stackrel{\text{def}}{=} \mathbb{P}(\mathsf{Y} = y)$$

Given a random variable X

$\log_2 \mathbb{P}(\mathsf{X})$ is a random variable: it is equal to $\log_2 p(x)$ when x as been picked according to the distribution $(p(y))_{y \in \mathcal{X}}$

Entropy:

Following average quantity of the source X,

$$H(\mathsf{X}) = \mathbb{E}_{\mathsf{X}}\Big(-\log_2 \mathbb{P}(\mathsf{X})\Big) = -\sum_{x} p(x)\log_2 p(x)$$

$$\longrightarrow \sharp\Big\{\text{typical set of } \mathsf{X}\Big\} \approx 2^{H(\mathsf{X})}$$

Entropy and Compression:

Optimal compression rate $H(\mathsf{X})$ when compressing symbols per symbols draw according to X

Conditional entropy:

$$H(\mathsf{Y} \mid \mathsf{X}_1, \ldots, \mathsf{X}_L) = - \sum_{y, x_1, \ldots, x_L} p(y, x_1, \ldots, x_L) \; \log_2 p(y \mid x_1, \ldots, x_L)$$

Proposition: conditioning reduces entropy

$$H(\mathsf{Y} \mid \mathsf{X}_1, \ldots, \mathsf{X}_L) \leq H(\mathsf{Y} \mid \mathsf{X}_1, \ldots, \mathsf{X}_{L-1})$$

$\longrightarrow$ To remember: if we admit that entropy is the optimal compression rate, conditioning can only help you, *i.e.*, decreasing the needed size to compress

Chain rule:

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

ENTROPY AND STOCHASTIC PROCESSES

Stochastic process:

A stochastic process is a discrete indexed sequence of random variables $\{X_i\}_i$ where the X_i's take their value in the same discrete alphabet $\mathcal{X}$

$\longrightarrow$ It is characterized by the joint **probability mass functions**

$$\mathbb{P}\Big((X_1, \dots, X_n) = (x_1, \dots, x_n) \Big) = p(x_1, \dots, x_n)$$

for all $n \in \mathbb{N}$ and $(x_1, \dots, x_n) \in \mathcal{X}^n$

- The X_i's can be dependent $\Big($**memory process**$\Big)$
- The X_i's don't have necessary the same distribution

Reminder: the entropy of a source outputting L symbols in $\mathcal{X}$ is defined as

$$H(X_1, \ldots, X_L) = \sum_{(x_1,\ldots,x_L)\in\mathcal{X}^L} -p(x_1, \ldots, x_L)\log_2 p(x_1, \cdots, x_L)$$

where $0 \cdot \log_2 0 = 0$

Entropy rate/Entropy per symbol:

The entropy of a stochastic process $\{X_i\}_i$ is defined by

$$H(\mathcal{X}) \stackrel{\text{def}}{=} \lim_{L\to+\infty} \frac{1}{L} H(X_1, \ldots, X_L)$$

when the limit exists

An important quantity:

For instance, informally, compressing with a Huffman code L symbols for L large enough can be done with $\approx LH(\mathcal{X})$ bits

- Typewriter: m equally likely output letters, hence m^L equally distributed sequences,

$$\frac{1}{L} H(X_1, \dots, X_L) = \frac{1}{L} \log_2 m^L = \log_2 m$$

- Independent **and** equally distributed,

$$\frac{1}{L} H(X_1, \dots, X_L) = \frac{1}{L} \sum_{i=1}^{L} H(X_i) = H(X)$$

Be careful:

$H(\mathcal{X})$ may not be defined when the X_i's are independent!

$\longrightarrow$ The following example is "technical" but is insightful

PROCESS FOR WHICH ENTROPY RATE IS NOT DEFINED (I)

$$\mathbb{N}\backslash\{0,1\} = \bigsqcup_{k=0}^{+\infty} [\![2^{2^{2k}}, 2^{2^{2k+2}}[\![\quad \text{and} \quad [\![2^{2^{2k}}, 2^{2^{2k+2}}[\![= \underbrace{[\![2^{2^{2k}}, 2^{2^{2k+1}}[\![}_{\text{length: } 2^{2^{2k}}(2^{2^{2k}}-1)} \bigcup \underbrace{[\![2^{2^{2k+1}}, 2^{2^{2k+2}}}_{\text{length: } 2^{2^{2k+1}}(2^{2^{2k+1}}-1)} [\![$$

- Each interval $[\![2^{2^{2k}}, 2^{2^{2k+2}}[\![$ has **exponential size**
- Each interval is decomposed into two exponential size intervals with one **exponentially bigger** than the other one

$$\frac{2^{2^{2k+1}}(2^{2^{2k+1}}-1)}{2^{2^{2k}}(2^{2^{2k}}-1)} \approx 2^{2^{k+1}}$$

$\{X_i\}_{i\geq 2}$ independent with $X_i \in \{0,1\}$ and $p_i \overset{\text{def}}{=} \mathbb{P}(X_i = 1)$ where

$$p_i = \begin{cases} \frac{1}{2} & \text{if } 2^{2^{2k}} \leq i < 2^{2^{2k+1}} \\ 0 & \text{if } 2^{2^{k+1}} \leq i < 2^{2^{2k+2}} \end{cases}$$

By definition:

$$H(X_i) = \begin{cases} 1 & \text{if } p_i = 1/2 \\ 0 & \text{otherwise} \end{cases}$$

By independence:

$$\frac{1}{L}H(X_1,\ldots,X_L) = \frac{1}{L}\sum_{i=1}^{L} \underbrace{H(X_i)}_{\in\{0,1\}}$$

Define:

$$u_{2k} \stackrel{\text{def}}{=} \sum_{i\leq 2^{2^{2k}}} H(X_i) \quad \text{and} \quad u_{2k+1} \stackrel{\text{def}}{=} \sum_{i<2^{2^{2k+1}}} H(X_i)$$

We obtain:

$$u_{2k+1} - u_{2k} = 2^{2^{2k}}\left(2^{2^{2k}} - 1\right) \quad \text{and} \quad u_{2k} - u_{2k-1} = 0$$

In particular,

$$0 \leq \frac{u_{2k}}{2^{2^{2k}}} = \frac{u_{2k-1}}{2^{2^{2k}}} \leq \frac{2^{2^{2(k-1)}}}{2^{2^{2k}}} = 2^{2^{2(k-1)}-2^{2k}} = 2^{2^{2(k-1)}(1-2^2)} \xrightarrow[k\to+\infty]{} 0$$

$$1 \geq \frac{u_{2k+1}}{2^{2^{2k+1}}} \geq \frac{2^{2^{2k}}\left(2^{2^{2k}}-1\right)}{2^{2^{2k+1}}} \xrightarrow[k\to+\infty]{} 1$$

Conclusion:

$\frac{1}{L}H(X_1,\ldots,X_L)$ cannot have a limit as two sub-series have different limits

The stochastic process we exhibited is highly dependent of the time of observation

Particularly: the process "is not defined" even after a sequence as long as we wish

An important class of processes/sources

Stationary process:

A stochastic process is said to be stationary if its behaviour is invariant by time observation,

$$\mathbb{P}\Big(X_1 = x_1, \dots, X_n = x_n\Big) = \mathbb{P}\Big(X_{1+\ell} = x_1, \dots, X_{n+\ell} = x_n\Big)$$

for any $n, \ell > 0$ and $(x_1, \dots, x_n) \in \mathcal{X}^n$

Exercise:

Show that a stochastic process independent and identically distributed is stationary

Be careful:

Stationary process is a very strong condition: it implies that X_i's are identically distributed

Stationary processes are important as their entropy per symbol is defined

Theorem:

For a stationary stochastic process, the following limits exist and are equal,

$$H(\mathcal{X}) = \lim_{L \to +\infty} \frac{1}{L} H(X_1, \ldots, X_L) = \lim_{L \to +\infty} H\left(X_L \mid X_1, \ldots, X_{L-1}\right)$$

PROOF(I): CONDITIONING DECREASES UNCERTAINTY

Proof:

For all L, using results of Slide 7,

$$\begin{aligned} H(X_L \mid X_1, \ldots, X_{L-1}) &\leq H(X_L \mid X_2, \ldots, X_{L-1}) \\ &= H(X_{L-1} \mid X_1, \ldots, X_{L-2}) \end{aligned}$$

where in the equality we used that the process is stationary,

$$\longrightarrow \lim_{L \to +\infty} H\left(X_L \mid X_1, \ldots, X_{L-1}\right) \text{ exists as decreasing} \geq 0 \text{ series}$$

PROOF(II): USE THE CHAIN RULE

Proof:

By the chain rule,

$$\frac{1}{L} H(X_1, \ldots, X_L) = \frac{1}{L} \sum_{i=1}^{L} H(X_i \mid X_1, \ldots, X_{i-1})$$

Cesaro's theorem:

Let $(a_L) \in \mathbb{C}^{\mathbb{N}}$ s.t $\lim\limits_{L \to +\infty} a_L = \ell$, then

$$\frac{1}{L} \sum_{i=1}^{L} a_i \xrightarrow[L \to +\infty]{} \ell$$

Proof:

To conclude, combine the result of the previous slide and the equation given by the chain rule

It is tempting to conclude that any source $X_1, \ldots, X_L, \ldots$ for which $H(\mathcal{X})$ is defined can be compressed at rate tending to $H(\mathcal{X})$

What do you think?

It is tempting to conclude that any source $X_1, \dots, X_L, \dots$ for which $H(\mathcal{X})$ is defined can be compressed at rate tending to $H(\mathcal{X})$

What do you think?

The work is not finished! Is it true that such sources concentrate over some subset as we used in Lecture 2?

$\longrightarrow$ No reason to be true!

ASYMPTOTIC EQUIPARTITION PROPERTY

ASSUMPTION:

In this section: only stochastic processes $\{X_i\}$ for which the entropy per symbol is defined!

$$\mathcal{H} \stackrel{\text{def}}{=} H(\mathcal{X}) = \lim_{L \to +\infty} \frac{1}{L} H(X_1, \ldots, X_L)$$

$\Big($in particular stationary processes$\Big)$

Remember the following anecdote:

At the police station, is it easier to answer the following questions: what were you doing three Monday ago? or what were you doing a typical Monday?

$\longrightarrow$ Typical realisations: simple mean to answer hard questions!

Typical sentences are those concentrating close to the entropy rate

Typical sequences:

The ε-typical set $A_\varepsilon^{(n)}$ is defined as $\Big($be careful: source for which $\mathcal{H}$ is defined$\Big)$,

$$A_\varepsilon^{(n)} \stackrel{\text{def}}{=} \left\{(x_1, \ldots, x_n) \in \mathcal{X}^n \,:\, 2^{-n(\mathcal{H}+\varepsilon)} \leq p(x_1, \ldots, x_n) \leq 2^{-n(\mathcal{H}-\varepsilon)}\right\}$$
$$= \left\{(x_1, \ldots, x_n) \in \mathcal{X}^n \,:\, \left|\frac{1}{n} \log_2 \frac{1}{p(x_1, \ldots, x_n)} - \mathcal{H}\right| \leq \varepsilon\right\}$$

Typical sequences are not the most probable!

$\longrightarrow$ Two dimensions in typical sequences, probability to fall in some set!

Be careful:

The most probable sequence associated to $X_i \in \{0,1\}$ where $\mathbb{P}(X_i = 1) = p < 1/2$, is

$$(0, \ldots, 0)$$

while the typical sequences associated to $X_i \in \{0,1\}$ where $\mathbb{P}(X_i = 1) = p$, are

$$\left\{ \mathbf{x} \in \{0,1\}^n : \ |\mathbf{x}| \approx np \right\} \quad \text{where } |\mathbf{x}| \stackrel{\text{def}}{=} \sharp\left\{ i : x_i \neq 0 \right\} \ \left(\text{Hamming weight of } \mathbf{x}\right)$$

An important remark: one may say that considering $\{\mathbf{x} : \ |\mathbf{x}| \leq np\}$ can be useful as it contains typical sequences and most probable sequences! However, it is useless...

$$\sharp\{\mathbf{x} : \ |\mathbf{x}| \leq np\} \approx \sharp\left\{ \mathbf{x} \in \{0,1\}^n : \ |\mathbf{x}| \approx np \right\}$$

It does not increase the size of the set of interest, it only brings negligible quantities

Asymptotic Equipartition Property (AEP):

A stochastic process $\{X_i\}_i$ verifies the AEP if,

$$\forall \varepsilon > 0, \quad \lim_{n \to +\infty} \mathbb{P}\Big((X_i)_{1 \leq i \leq n} \in A_\varepsilon^{(n)} \Big) = 1 \iff \frac{1}{n} \log_2 \mathbb{P}(X_1, \ldots, X_n) \xrightarrow[n \to +\infty]{P} H(\mathcal{X})$$

$\longrightarrow$ The entropy per symbol is defined for stochastic processes verifying the AEP

Exercise:

Show that the i.i.d. stochastic process $X_i \in \{0, 1\}$ where $\mathbb{P}(X_i = 1) = p$ verifies the AEP

Proposition:

For any source verifying the AEP, for all $\varepsilon > 0$,

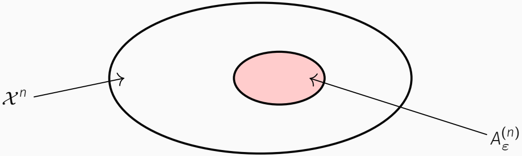
1. $\mathbb{P}\Big((X_i)_{1\leq i\leq n} \in A_\varepsilon^{(n)}\Big) \geq 1 - \varepsilon$ for n being sufficiently large
2. $\sharp A_\varepsilon^{(n)} \leq 2^{n(\mathcal{H}+\varepsilon)}$
3. $\sharp A_\varepsilon^{(n)} \geq (1-\varepsilon)2^{n(\mathcal{H}-\varepsilon)}$ for n being sufficiently large

Proof $\Big($same thing than in Lecture 2$\Big)$:

1. By definition
2. We have the following computation,

$$\begin{aligned} 1 &= \sum_{\mathsf{x}} p(\mathsf{x}) \\ &\geq \sum_{\mathsf{x}\in A_\varepsilon^{(n)}} p(\mathsf{x}) \\ &\geq \sum_{\mathsf{x}\in A_\varepsilon^{(n)}} 2^{-n(\mathcal{H}+\varepsilon)} \end{aligned}$$

   where we used the definition of typical sequences. It concludes the proof
3. Same reasoning but starting from $1-\varepsilon \leq \mathbb{P}\Big((\mathsf{X}_i)_{1\leq i\leq n} \in A_\varepsilon^{(n)}\Big)$ instead of $1 = \sum_{\mathsf{x}} p(\mathsf{x})$



Only an exponentially small fraction of sequences $\left(\sharp A_\varepsilon^{(n)} \ll \sharp \mathcal{X}^n\right)$ concentrates all the distribution mass of sequences verifying the AEP

$$\mathbb{P}\left(\mathsf{x} \in T_\varepsilon^{(n)}\right) \approx 1$$

Remember Lecture 2, set with high probability, i.e., δ-sufficient subset $\left(\text{for compression}\right)$

δ-sufficient subset $\mathcal{S}_\delta$:

$$\mathbb{P}\left(\mathbf{x} \in S_\delta^{(n)}\right) \geq 1 - \delta$$

Theorem:

For any δ-sufficient subset $\mathcal{S}_\delta$ and any source verifying the AEP, for all $\varepsilon > 0$, for n being sufficiently large,

1. $$\mathbb{P}\left(\mathbf{x} \in S_\delta^{(n)} \cap A_\varepsilon^{(n)}\right) \geq 1 - \varepsilon - \delta$$
2. $$\frac{1}{n} \log_2 \sharp S_\delta^{(n)} > \mathcal{H} - \varepsilon$$

$\longrightarrow$ Sufficient subsets cannot be smaller than typical sets as $\sharp A_\varepsilon^{(n)} \geq (1 - \varepsilon) 2^{n(\mathcal{H} - \varepsilon)}$

for n being sufficiently large

Proof:

First, let $\mathcal{E}$ and $\mathcal{F}$ be two events such that

$$\mathbb{P}(\mathcal{E}) \geq 1 - \alpha \quad \text{and} \quad \mathbb{P}(\mathcal{F}) \geq 1 - \beta$$

We have

$$\begin{aligned} \mathbb{P}(\overline{\mathcal{E}} \cup \overline{\mathcal{F}}) &\leq \mathbb{P}(\overline{\mathcal{E}}) + \mathbb{P}(\overline{\mathcal{F}}) \\ &\leq \alpha + \beta \end{aligned}$$

Therefore,

$$\mathbb{P}(\mathcal{E} \cap \mathcal{F}) = 1 - \mathbb{P}(\overline{\mathcal{E}} \cup \overline{\mathcal{F}}) \geq 1 - \alpha - \beta$$

1. Apply the above reasoning
2. For n being sufficiently large,

$$\begin{aligned} 1 - \varepsilon - \delta &\leq \mathbb{P}\left(\mathrm{x} \in S_\delta^{(n)} \cap A_\varepsilon^{(n)}\right) \\ &= \sum_{\mathrm{x} \in S_\delta^{(n)} \cap A_\varepsilon^{(n)}} p(\mathrm{x}) \\ &\leq \sum_{\mathrm{x} \in S_\delta^{(n)} \cap A_\varepsilon^{(n)}} 2^{-n(\mathcal{H} - \varepsilon)} \\ &\leq \sharp S_\delta^{(n)} \, 2^{-n(\mathcal{H} - \varepsilon)} \end{aligned}$$

To conclude the proof use the $\log_2$ properties and $\frac{1}{n} \log_2 \mathrm{Cst} \xrightarrow[n \to +\infty]{} 0$

Asymptotic coding average length:

Given a stochastic process, $\{X_i\}$, the asymptotic expected length of a symbol code φ is defined as $\big($if the following limits exists$\big)$,

$$L_{\text{asympt}}(\varphi, \mathcal{X}) \stackrel{\text{def}}{=} \lim_{n \to +\infty} \frac{1}{n} \sum_{x_1, \dots, x_n} \ell(x_1, \dots, x_n)\, p(x_1, \dots, x_n)$$

where $\ell(x_1, \dots, x_n)$ bit-length of $\varphi(x_1, \dots, x_n)$

Shannon source coding theorem:

Given a source verifying the AEP and with entropy per symbol $H(\mathcal{X})$,

1. All unambiguous coding φ verifies $L_{\text{asympt}}(\varphi, \mathcal{X}) \geq \mathcal{H}$
2. It exists an unambiguous coding φ such that $L(\varphi, \mathcal{X}) \leq \mathcal{H} + \varepsilon$

Proof:

1. φ can be defined as an unambiguous code over $\mathcal{X}^n$, then by results of Lecture 2 $\left(\text{"Shannon's theorem"}\right)$,

$$\frac{1}{n}\sum_{x_1,\ldots,x_n} \ell(x_1,\ldots,x_n)\; p(x_1,\ldots,x_n) \geq \frac{1}{n}\; H(\mathrm{X}_1,\ldots,\mathrm{X}_n)$$

The right-hand term as limit $\mathcal{H}$ $\left(\text{when } n \text{ tends to } +\infty\right)$

2. The idea is to distinguish elements according to $\mathbf{x} \in A_\varepsilon^{(n)}$ or not.

(*i*) Define a one-to-one mapping,

$$\varphi_0 : \mathcal{X}^n \to \{0,1\}^{\lceil n \log_2 \sharp\mathcal{X}\rceil}$$

(*ii*) Define a one-to-one mapping,

$$\varphi_1 : A_\varepsilon^{(n)} \longrightarrow \{0,1\}^{\lceil \sharp A_\varepsilon^{(n)}\rceil}$$

Define the unambiguous $\left(\text{and fixed-length}\right)$ code $\varphi_\varepsilon^{(n)}$,

$$\varphi_\varepsilon^{(n)}(\mathbf{x}) \stackrel{\text{def}}{=} \begin{cases} 0||\varphi_0(\mathbf{x}) \text{ if } \mathbf{x} \notin A_\varepsilon^{(n)} \\ 1||\varphi_1(\mathbf{x}) \text{ if } \mathbf{x} \in A_\varepsilon^{(n)} \end{cases}$$

Proof:

- The idea is to distinguish elements according to $\mathbf{x} \in A_\varepsilon^{(n)}$ or not

  (*i*) Define a one-to-one mapping,

$$\varphi_0 : \mathcal{X}^n \to \{0,1\}^{\lceil n \log_2 \sharp\mathcal{X} \rceil}$$

  (*ii*) Define a one-to-one mapping,

$$\varphi_1 : A_\varepsilon^{(n)} \longrightarrow \{0,1\}^{\lceil \sharp A_\varepsilon^{(n)} \rceil}$$

  Define the unambiguous $\big($and fixed-length$\big)$ code $\varphi_\varepsilon^{(n)}$,

$$\varphi_\varepsilon^{(n)}(\mathbf{x}) \overset{\text{def}}{=} \begin{cases} 0||\varphi_0(\mathbf{x}) \text{ if } \mathbf{x} \notin A_\varepsilon^{(n)} \\ 1||\varphi_1(\mathbf{x}) \text{ if } \mathbf{x} \in A_\varepsilon^{(n)} \end{cases}$$

By taking n large enough, $\mathbb{P}\left(\mathbf{x} \in A_\varepsilon^{(n)}\right) \geq 1 - \varepsilon$ and $\sharp A_\varepsilon^{(n)} \leq 2^{n(\mathcal{H}+\varepsilon)}$,

$$\begin{aligned} \sum_{\mathbf{x}} p(\mathbf{x})\, \ell(\mathbf{x}) &= \mathbb{P}\left(\mathbf{x} \in A_\varepsilon^{(n)}\right) \lceil \sharp A_\varepsilon^{(n)} \rceil + \left(1 - \mathbb{P}\left(\mathbf{x} \in A_\varepsilon^{(n)}\right)\right) \lceil n \log_2 \sharp\mathcal{X} \rceil \\ &\leq 1\, \lceil n(\mathcal{H} + \varepsilon) + \varepsilon \lceil n \log_2 \sharp\mathcal{X} \rceil \end{aligned}$$

To conclude: $n \to +\infty$

WHAT ELSE?

We defined the entropy per symbol as an entropy quantity to quantify optimal compression

We defined the AEP property as a necessary condition to reach optimal compression

What else?

WHAT ELSE?

We defined the entropy per symbol as an entropy quantity to quantify optimal compression

We defined the AEP property as a necessary condition to reach optimal compression

What else?

Which $\left(\text{interesting}\right)$ sources have an entropy per symbol be defined **and** verify the AEP?

MEMORYLESS SOURCES VERIFY AEP

Memoryless source:

A source $\{X_i\}_i$ is said to be memoryless if the X_i's are i.i.d.

Proposition:

Memoryless sources verify the AEP

Proof:

Let $\{X_i\}_i$ be a memoryless process defined as X

1. We have the following computation,

$$\frac{1}{n}H(X_1, \ldots, X_n) \overset{\text{(indep)}}{=} \frac{1}{n}\sum_{i=1}^{n} H(X_i) = H(X)$$

as they are identically distributed. Therefore: the entropy rate is defined and $H(\mathcal{X}) = H(X)$

2. By independence,

$$\log_2 \mathbb{P}(X_1, \ldots, X_n) = \sum_{i=1}^{n} \log_2 \mathbb{P}(X_i)$$

By linearity of the expectation,

$$\mathbb{E}\left(-\log_2 \mathbb{P}(X_1, \ldots, X_n)\right) = \mathbb{E}\left(-\sum_{i=1}^{n} \log_2 \mathbb{P}(X_i)\right) = nH(X)$$

By the weak law of large number,

$$\left(\left|\frac{1}{n}\log_2 1/\mathbb{P}(X_1, \ldots, X_n) - H(X)\right| \leq \varepsilon\right) \xrightarrow[n \to +\infty]{} 1$$

Important remark:

don't think that elements of the typical set as being exactly equiprobable

By definition of the typical set, for $\mathbf{x} \in A_\varepsilon^{(n)}$, quantities $\log_2 \frac{1}{\mathbb{P}(\mathbf{x})}$ are within $2n\varepsilon$ of each other

But how did we choose $n\varepsilon$?

Important remark:

don't think that elements of the typical set as being exactly equiprobable

By definition of the typical set, for $\mathbf{x} \in A_\varepsilon^{(n)}$, quantities $\log_2 \frac{1}{\mathbb{P}(\mathbf{x})}$ are within $2n\varepsilon$ of each other

But how did we choose $n\varepsilon$?

By the weak law of large numbers, $\mathbb{P}\left(\mathbf{x} \in A_\varepsilon^{(n)}\right) \geq 1 - \frac{\sigma}{\varepsilon^2 n^2}$ $\left(\sigma \text{ variance of } \mathsf{X}\right)$

$\longrightarrow n = \alpha \frac{1}{\varepsilon^2}$ for some constant α

We deduce that the most probable string in the typical can be of order $2^{n\varepsilon} = 2^{\alpha/\varepsilon} = 2^{\sqrt{\alpha n}}$ times greater than the least probable string in the typical set

$2^{\sqrt{\alpha n}}$ is an exponential quantity!

Asymptotic coding average length (reminder):

Given a stochastic process, $\{X_i\}$, the asymptotic expected length of a symbol code φ is defined as $\Big($if the following limits exists$\Big)$,

$$L_{\text{asympt}}(\varphi, \mathcal{X}) \stackrel{\text{def}}{=} \lim_{n \to +\infty} \frac{1}{n} \sum_{x_1, \ldots, x_n} \ell(x_1, \ldots, x_n)\; p(x_1, \ldots, x_n)$$

where $\ell(x_1, \ldots, x_n)$ bit-length of $\varphi(x_1, \ldots, x_n)$

Theorem: Shannon source coding for memoryless sources

For any memory less source $(X_i)_i$, it exists an unambiguous coding φ such that

$$L_{\text{asympt}}(\varphi, \mathcal{X}) = H(X)$$

Furthermore, for any unambiguous coding φ, $L_{\text{asympt}}(\varphi, \mathcal{X})/H(X) \geq 1$

Typical series:

Let $\mathcal{X} = \{a_1, \ldots, a_k\}$. The ε-typical series is defined as,

$$\left\{ (x_1, \ldots, x_n) \in \mathcal{X}^n, \ \left| \sum_{i=1}^{k} \left(\frac{n_{a_i}(\mathbf{x})}{n} - p(a_i) \right) \log_2 p(a_i) \right| \leq \varepsilon \right\}$$

where $n_{a_i}(\mathbf{x})$ the number of times that a_i occurs in $\mathbf{x} = (x_1, \ldots, x_n)$

$\longrightarrow$ Both definitions are equivalent in the case of memoryless sources

One paper about this theory of AEP and compression for memoryless sources in the quantum case

- Chapter 11 up to 12.3 in *Quantum Computation and Quantum Information*, Michael A. Nielsen and Isaac L. Chuang

MARKOV: GENERAL SOURCES VERIFYING AEP

Stochastic matrix:

Given a finite set $\mathcal{X}$, a matrix $\mathbf{P} = (p(x,y))_{x,y\in\mathcal{X}}$ is said to be stochastic if

- $p(x,y) \geq 0$ for all $x, y \in \mathcal{X}$
- $\sum_{y\in\mathcal{X}} p(x,y) = 1$

Fundamental fact

If $\mathbf{x} = (q(x))_{x\in\mathcal{X}}$ is a distribution[a] and $\mathbf{P}$ is a stochastic matrix. Then, $\mathbf{x}^\top \mathbf{P}$ is a distribution

[a] for all $x \in \mathcal{X}$, $q(x) \geq 0$ and $\sum_{x\in\mathcal{X}} q(x) = 1$

- Let $(r(y))_{y\in\mathcal{X}}$ be the distribution defined as $\mathbf{x}^\top \mathbf{P}$. We have,

$$r(y) = \sum_{x\in\mathcal{X}} q(x)p(x,y)$$

- r defines the distribution: pick x according to q and then pick y with probability $p(x,y)$

Markov chains give a rule to walk from one point to the other independently of the path we followed in the past

Markov chain:

Let $\mathcal{X}$ be a finite set, $(q(x))_{x\in\mathcal{X}}$ be a distribution and $\mathbf{P} = (p(x,y))_{x,y\in\mathcal{X}}$ be a stochastic matrix.

A $\Big($homogenous$\Big)$ Markov chain with state space $\mathcal{X}$, initial distribution q and transition matrix $\mathbf{P}$ is a sequence of random variables $\mathsf{X}_0, \ldots, \mathsf{X}_t, \ldots$ such that

$$\mathbb{P}\left(\mathsf{X}_0 = x_0\right) = q(x_0) \quad \text{and} \quad \mathbb{P}\left(\mathsf{X}_{t+1} = x_{t+1} \mid \mathsf{X}_t = x_t, \ldots, \mathsf{X}_0 = x_0\right) = p(x_t, x_{t+1})$$

for all $t \in \mathbb{N}$ and $x_0, \ldots, x_{t+1} \in \mathcal{X}$ such that $\mathbb{P}\left(\mathsf{X}_0 = x_0, \ldots, \mathsf{X}_t = x_t\right) > 0$

Remark:

The homogenous term refers to the fact that for each t the transition matrix is the same

Proposition:

Given a Markov chain $(X_t)_t$ with initial distribution $(q(x))_{x\in\mathcal{X}}$, transition matrix $\mathbf{P} = (p(x,y))_{x,y\in\mathcal{X}}$,

$$\mathbb{P}(X_t = x_t) = q^{(t)}(x) \quad \text{where} \quad \left(q^{(t)}(x)\right)_{x\in\mathcal{X}} \stackrel{\text{def}}{=} \left(q(x)\right)_{x\in\mathcal{X}}^{\top} \mathbf{P}^{t}$$

and,

$$\mathbb{P}(X_{t+1} = x_{t+1} \mid X_t = x_t) = p(x_t, x_{t+1})$$

$\Big(p(x,y)$: rule for moving from x to y, we read from left to right$\Big)$

Proof:

Exercise

Notation:

Given $\mathbf{P} = \left(p(x,y)\right)_{x,y\in\mathcal{X}}$, we denote $\mathbf{P}^{t} = \left(p^{(t)}(x,y)\right)_{x,y\in\mathcal{X}}$

Starting from the distribution $\mathbf{x} = (q(x))_{x \in \mathcal{X}}$ and after t walks we are distributed as $\mathbf{x}^\top \mathbf{P}^t$

Stationary distribution:

Let $\mathbf{P}$ be a stochastic matrix. A stationary distribution for $\mathbf{P}$ is a distribution π such that

$$\pi^\top = \pi^\top \mathbf{P}$$

$\longrightarrow$ Starting from the stationary distribution and applying the walk keeps invariant the distribution!

$$\left(\text{given } \mathbf{P} = \left(p(x,y)\right)_{x,y\in\mathcal{X}}, \text{ we denote } \mathbf{P}^t = \left(p^{(t)}(x,y)\right)_{x,y\in\mathcal{X}}\right)$$

Ergodicity:

A stochastic matrix $\mathbf{P}$ is said ergodic if there exists $t_0 \in \mathbb{N}$ such that

$$\forall x, y \in \mathcal{X}, \quad p^{(t_0)}(x,y) > 0$$

Theorem:

A stochastic matrix $\mathbf{P}$ is ergodic if and only if there exists a strict probability distribution[a] π on $\mathcal{X}$ such that

$$\forall x, y \in \mathcal{X}, \quad p^{(t)}(x,y) \xrightarrow[t\to+\infty]{} \pi(y)$$

Furthermore, when $\mathbf{P}$ is ergodic, the above distribution π is the unique stationary distribution

We deduce that for any Markov chain $\{X_t\}_t$ with ergodic matrix $\mathbf{P}$,

$$\forall y \in \mathcal{X}, \quad \mathbb{P}(X_t = y) \xrightarrow[t\to+\infty]{} \pi(y)$$

where π is the unique stationary distribution of $\mathbf{P}$

[a] $\pi(x) > 0$ for any $x \in \mathcal{X}$

Proof:

Suppose that **P** is ergodic and $\varepsilon \stackrel{\text{def}}{=} \min_{x,y\in\mathcal{X}} p^{(t_0)}(x,y) \in (0,1)$,

$$M^{(t)}(y) \stackrel{\text{def}}{=} \max_{x\in\mathcal{X}} p^{(t)}(x,y) \quad m^{(t)}(y) \stackrel{\text{def}}{=} \min_{x\in\mathcal{X}} p^{(t)}(x,y)$$

We have,

$$m^{(t)}(x,y) \le \sum_z p(x,z)m^{(t)}(x,y) \le \sum_z p(x,z)p^{(t)}(z,y) = p^{(t+1)}(x,y) \le \sum_z p(x,z)M^{(t)}(y) = M^{(t)}(y)$$

We deduce that $t \mapsto M^{(t)}(x,y)$ and $t \mapsto m^{(t)}(x,y)$ are decreasing and increasing. Therefore they convergence as belonging to $(0,1)$. Call $\pi_1(y)$ and $\pi_2(y)$ their limits. For any $r \ge 0$ we have:

$$\begin{aligned}
p^{(t_0+r)}(x,y) &= \sum_z p^{(t_0)}(x,z)p^{(r)}(z,y) \\
&= \sum_z \left(p^{(t_0)}(x,z) - \varepsilon p^{(r)}(y,z)\right) p^{(r)}(z,y) + \varepsilon \cdot \sum_z p^{r}(y,z)p^{(r)}(z,y) \\
&\ge m^{(r)}(y) \sum_z \left(p^{(t_0)}(x,z) - \varepsilon p^{(r)}(y,z)\right) + \varepsilon \cdot p^{(2r)}(y,y) \\
&= (1-\varepsilon)\cdot m^{(r)}(y) + \varepsilon \cdot p^{(2r)}(y,y) \ge (1-\varepsilon)m^{(r)}(y) + \varepsilon \cdot m^{(2r)}(y,y)
\end{aligned}$$

where the inequality follows from the fact that $\left(\text{as } \varepsilon \ge p^{(t_0)}(x,z)\right)$,

$$p^{(t_0)}(x,z) - \varepsilon p^{(r)}(y,z) \ge p^{(t_0)}(x,z)\left(1 - p^{(r)}(y,z)\right) \ge 0$$

Proof:

Similarly $M^{(n_0+r)}(y) \leq (1-\varepsilon)M^{(r)}(y) + \varepsilon \cdot M^{(2r)}(y,y)$. We deduce that for any k,

$$M^{(kn_0+r)}(y) - m^{(kn_0+r)}(y) \leq (1-\varepsilon)^k \left(M^{(r)}(y) - m^{(r)}(y)\right) \xrightarrow[k\to+\infty]{} 0$$

Therefore $\pi(y) \stackrel{\text{def}}{=} \pi_1(y) = \pi_2(y)$ and from above with the fact that $M^{(t)}$ and $m^{(t)}$ are decreasing and increasing, for $t = kn_0 + r$ where $0 \leq r \leq n_0$,

$$\left|p^{(t)}(x,y) - \pi(y)\right| \leq M^{(t)}(y) - m^{(t)}(y) \leq (1-\varepsilon)^{n/n_0}$$

and therefore $p^{(t)}(x,y) \xrightarrow[t\to+\infty]{} \pi(y)$. Furthermore,

$$p^{(t+1)}(x,y) \sum_z p^{(t)}(x,z)p(z,y)$$

we get with $t \to +\infty$,

$$\pi(y) = \sum_z \pi(z)p(z,y)$$

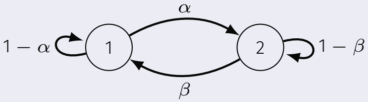
which shows that π is a stationary distribution $\left(\text{it is a distribution as } \sum_z p^{(t)}(x,z) = 1 \text{ and } p^{(t)}(x,z) \geq 0\right)$. It is strict as $m^{(t)}(y) \geq \varepsilon > 0$.

Conversely, suppose that $p^{(t)}(x,y) \xrightarrow[t\to+\infty]{} \pi(y) > 0$. We deduce easily that **P** is ergodic $\left(\text{a finite number of } y\right)$. To prove uniqueness let π' be another stationary distribution,

$$\pi'(y) = \sum_x \pi'(x)p^{(t)}(x,y) \xrightarrow[t\to+\infty]{} \sum_x \pi'(x)\pi(y) = \pi(y)$$

Two-state Markov chain with a probability transition matrix

$$\mathbf{M} = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$$



The stationary distribution is: $\begin{pmatrix} \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \end{pmatrix}$

Proposition:

The entropy rate of any ergodic Markov chain with transition matrix $\mathbf{P}$ exists and is equal to:

$$H(\mathcal{X}) = \lim_{L \to +\infty} H(\mathsf{X}_L \mid \mathsf{X}_{L-1}) = -\sum_{x_1, x_2} \pi(x_1)\ p(x_1, x_2)\ \log_2 p(x_1, x_2)$$

where π is the unique stationary distribution

Furthermore, if the initial condition of the Markov chain is its stationary distribution, then

$$H(\mathcal{X}) = H(\mathsf{X}_2 \mid \mathsf{X}_1)$$

Proof:

The entropy rate may not be defined as the process is not stationary. Let us first show that $\lim_{L\to+\infty} H(X_L \mid X_1, \dots, X_{L-1})$ exists. First, by definition of the Markov chain $\big($it has order 1$\big)$,

$$H(X_L \mid X_1, \dots, X_{L-1}) = H(X_L \mid X_{L-1})$$

We have now the following computation,

$$\begin{aligned} H(X_L \mid X_{L-1}) &= -\sum_{x,y} \mathbb{P}(X_L = y \mid X_{L-1} = x)\mathbb{P}(X_{L-1} = x) \log_2 \mathbb{P}(X_L = y \mid X_{L-1} = x) \\ &= -\sum_{x,y} p(x,y) \underbrace{\mathbb{P}(X_{L-1} = x)}_{\xrightarrow[L\to+\infty]{} \pi(x)} \log_2 p(x,y) \end{aligned}$$

Therefore $\lim_{L\to+\infty} H(X_L \mid X_1, \dots, X_{L-1})$ exists and as we did using Cesaro Theorem $\big($see Slide 19$\big)$, $H(\mathcal{X})$ exists and

$$H(\mathcal{X}) = \lim_{L\to+\infty} H(X_L \mid X_{L-1}) = -\textstyle\sum_{x,y} \pi(x)p(x,y)\ \log_2 p(x,y)$$

Furthermore, if the initial condition is the stationary distribution π, then

$$\pi(x)p(x,y) = \mathbb{P}(X_1 = x)\ \mathbb{P}(X_2 = y \mid X_1 = x) = \mathbb{P}(X_2 = y, X_1 = x)$$

Therefore $\big($see Proposition given in Slide 45$\big)$,

$$H(\mathcal{X}) = -\sum_{x,y} \mathbb{P}(X_2 = y, X_1 = x) \log_2 \mathbb{P}(X_2 = y \mid X_1 = x) = H(X_2 \mid X_1)$$

Proposition:

Any stationary and ergodic Markov chain with transition matrix **P** verifies the AEP and,

$$\frac{1}{n}\ \log_2 \mathbb{P}(X_1, \ldots, X_n) \xrightarrow[n \to +\infty]{P} H(\mathcal{X}) = -H(X_2 \mid X_1)$$

where π denotes the unique stationary distribution

Exercise:

Show that an ergodic Markov chain is stationary if and only its initial distribution is the unique stationary distribution

Proof:

$$\mathbb{P}(x_1,\dots,x_n) = \mathbb{P}(X_1)\mathbb{P}(X_2 \mid X_1)\mathbb{P}(X_3 \mid X_2, X_1)\cdots\mathbb{P}(X_n \mid X_{n-1},\dots,X_1)$$
$$= \mathbb{P}(X_1)\mathbb{P}(X_2 \mid X_1)\mathbb{P}(X_3 \mid X_2)\cdots\mathbb{P}(X_n \mid X_{n-1})$$

because Markov. Therefore,

$$\frac{1}{n}\log_2 \mathbb{P}(X_1,\dots,X_n) = \underbrace{\frac{1}{n}\log_2 \mathbb{P}(X_1)}_{\xrightarrow[n\to+\infty]{}0} + \frac{1}{n}\Big(\log_2 \mathbb{P}(X_2 \mid X_1) + \cdots + \log_2 \mathbb{P}(X_n \mid X_{n-1})\Big)$$

Weak law of large number for weak dependency (proof exercise session):

Let $Y_1,\dots,Y_n$ be identically random variables such that

$$\frac{1}{n^2}\sum_{i,j=1}^{n}\mathrm{Cov}\,(Y_i,Y_j) \xrightarrow[n\to+\infty]{} 0 \quad \text{where } \mathrm{Cov}(Y_i,Y_j) \overset{\text{def}}{=} \mathbb{E}(Y_iY_j) - \mathbb{E}(Y_i)\mathbb{E}(Y_j)$$

Then,

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow[n\to+\infty]{P} \mathbb{E}(Y_1)$$

Proof:

$$\frac{1}{n}\log_2 \mathbb{P}(X_1,\ldots,X_n) = \varepsilon(n) + \frac{1}{n}\sum_j Y_j$$

- $\varepsilon(n) = \frac{1}{n}\log_2 \mathbb{P}(X_1) \xrightarrow[n\to+\infty]{P} 0$
- $Y_j \overset{\text{def}}{=} -\log_2 \mathbb{P}\left(X_{j+1} \mid X_j\right)$ are identically random variable as the Markov chain is stationary

We have,

$$\mathbb{E}(Y_j) = H(X_{j+1} \mid X_j) = H(X_2 \mid X_1) = H(\mathcal{X})$$

Now, $\Big($only $3n$ Covariances are non-zero and they are independent of $n\Big)$

$$\frac{1}{n^2}\sum_{i,j=1}^{n^2} \mathrm{Cov}\left(Y_i, Y_j\right) \xrightarrow[n\to+\infty]{} 0$$

To conclude we apply the weak law of large number for weak dependency

Asymptotic coding average length:

Given a stochastic process, $\{X_i\}$, the asymptotic expected length of a symbol code φ is defined as $\big($if the following limits exists$\big)$,

$$L_{\text{asympt}}(\varphi, \mathcal{X}) \stackrel{\text{def}}{=} \lim_{n \to +\infty} \frac{1}{n} \sum_{x_1, \ldots, x_n} \ell(x_1, \ldots, x_n)\, p(x_1, \ldots, x_n)$$

where $\ell(x_1, \ldots, x_n)$ bit-length of $\varphi(x_1, \ldots, x_n)$

Theorem: Shannon source coding for Markov chains

For any ergodic and stationary Markov chain $(X_i)_i$, it exists an unambiguous coding φ such that

$$L_{\text{asympt}}(\varphi, \mathcal{X}) = H(X_2 \mid X_1)$$

Furthermore, for any unambiguous coding φ, $L_{\text{asympt}}(\varphi, \mathcal{X})/H(X_2 \mid X_1) \geq 1$

MORE GENERALITY: ERGODIC PROCESSES

There is more general processes for Shannon's theorem to be verified: **ergodic processes**

Intuitively:

Ergodic process: we can determine the distribution by observing a sufficiently long sequence

$\longrightarrow$ There is an average behaviour that we can determine, *i.e.*, the law of large number is "verified"

- Approximation of order 0 (all symbols, don't forget the "space", are i.i.d.):

XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGXYD QPAAMKBZAACIBZLHJQD

$$H_0 = \log_2 27 \approx 4.76$$

- Approximation of order 1 (the letters are chosen according to their frequency in English)

OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL

$$H_1 \approx 4.03$$

- Approximation of order 2 : same distribution of the pairs as in English

ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE

- Approximation of order 3 : same frequency of the triplets as in English

IN NO IST LAT WHEY CRATICT FROURE BERS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE

THE ENGLISH LANGUAGE(I)

- Approximation of order 0 $\left(\text{all symbols, don't forget the "space", are i.i.d.}\right)$:

XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGXYD QPAAMKBZAACIBZLHJQD

$$H_0 = \log_2 27 \approx 4.76$$

- Approximation of order 1 $\left(\text{the letters are chosen according to their frequency in English}\right)$

OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL

$$H_1 \approx 4.03$$

- Approximation of order 2 : same distribution of the pairs as in English

ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE

- Approximation of order 3 : same frequency of the triplets as in English

IN NO IST LAT WHEY CRATICT FROURE BERS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE

Any idea to generate correctly some English test?

- Markov model of order 3 of English (the frequency of quadruplets of letters matches English text. Each letter depends on the previous three letters)

  THE GENERATED JOB PRIVIDUAL BETTER TRAND THE DIPLAYED CODE, ABOVERY UPONDULTS WELL THE CODERST IN THESTICAL IT DO HOCK BOTH MERG. (INSTATES CONS ERATION. NEVER ANY OF PUBLE AND TO THEORY. EVENTIAL CALLEGAND TO ELAST BENERATED IN WITH PIES AS WITH THE)

- Markov model of order 1 on the words (the word transition probabilities match English text)

  THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED

Generate some English language by using the Markov chain model. Give an estimation of the entropy rate of the English

We know that:

$$\lim_{L \to +\infty} \frac{1}{L} H(X_1, \ldots, X_L) = \lim_{L \to +\infty} H(X_L \mid X_1, \ldots, X_{L-1}) = H(\mathcal{X})$$

$\longrightarrow$ Using Huffman encoding with packing of L $\Big($large$\Big)$ letters, *i.e.*, using $\mathcal{X}^L$ as source alphabet instead of $\mathcal{X}$, enables to optimally compress for instance the English

Issue:

Memory complexity in Huffman encoding is $O(\sharp\mathcal{Y})$ where $\mathcal{Y}$ is the source alphabet...

Overcoming this issue: Lecture 4

ABOUT HUFFMAN CODING(II)

Another issue with Huffman coding: we need to know the probabilities to compress
(in order to build the tree)

$\longrightarrow$ There are optimal compression even when nothing is known about the source!

- Lempel-Ziv compression algorithm in *Elements of Information Theory*, Chapter 13, Thomas M. Cover and Joy A. Thomas

EXERCISE SESSION