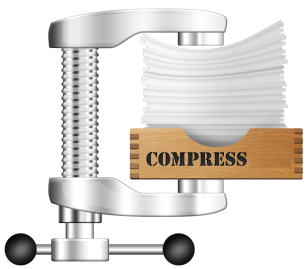
WEEK 4, PART 1:
SOURCE CODING THEOREM (BOOK CHAPTER 5)

Prof. Michael Gastpar

Slides by Prof. M. Gastpar and Prof. em. B. Rimoldi

EPFL

Spring Semester 2024



Outline

LAST LECTURE:

- $H(X|Y,Z) \leq H(X|Z) \leq H(X)$

"CONDITIONING REDUCES ENTROPY"

- $H(S_1, S_2, \ldots, S_n)$ **CHAIN RULE**

$$= H(S_1) + H(S_2|S_1) + H(S_3|S_1,S_2) + H(S_4|S_1,S_2,S_3) + \ldots + H(S_n|S_1, S_2, \ldots, S_{n-1})$$

Our Main Source Coding Result, so far

Theorem (Textbook Thm 3.3)

The **per-letter** average codeword-length of a D-ary Shannon-Fano code for the random variable $(S_1, S_2, \ldots, S_n)$ fulfills

$$\frac{H_D(S_1, S_2, \ldots, S_n)}{n} \leq \frac{L((S_1, S_2, \ldots, S_n), \Gamma_{SF})}{n} < \frac{H_D(S_1, S_2, \ldots, S_n)}{n} + \frac{1}{n}.$$

As $n \to \infty$, the left and the right bound coincide, thus leading to an exact expression.

Earlier, we studied the IID source where:

- $\frac{H_D(S_1, S_2, \ldots, S_n)}{n} = H_D(S)$,
- and thus, $H_D(S)$ characterizes the minimum number of bits (more precisely, of D-ary code symbols) per original source symbol we could ever hope to spend to compress the source without loss.

Our Main Source Coding Result, more generally

Theorem (Textbook Thm 3.3)

The **per-letter** average codeword-length of a D-ary Shannon-Fano code for the random variable $(S_1, S_2, \ldots, S_n)$ fulfills

$$\frac{H_D(S_1, S_2, \ldots, S_n)}{n} \leq \frac{L((S_1, S_2, \ldots, S_n), \Gamma_{SF})}{n} < \frac{H_D(S_1, S_2, \ldots, S_n)}{n} + \frac{1}{n}.$$

As $n \to \infty$, the left and the right bound coincide, thus leading to an exact expression.

Now, more generally,

- if we suppose the $R^*(\mathcal{S}) = \lim_{n\to\infty} \frac{H_D(S_1, S_2, \ldots, S_n)}{n}$ exists,
- then $R^*(\mathcal{S})$ characterizes the minimum number of bits (more precisely, of D-ary code symbols) per original source symbol we could ever hope to spend to compress the source without loss.

What Next?

Therefore, for today, the main question is:

- For which sources does $R^*(\mathcal{S}) = \lim_{n\to\infty} \frac{H_D(S_1,S_2,\ldots,S_n)}{n}$ exist?

To this end, we will introduce a number of concrete source models and show how to find $R^*(\mathcal{S}) = \lim_{n\to\infty} \frac{H_D(S_1,S_2,\ldots,S_n)}{n}$.

RANDOM PROCESSES

A.K.A. SOURCE MODELS.

$S_1, S_2, S_3, S_4, S_5, S_6, \ldots$

↳ HOW TO THINK ABOUT SUCH SEQUENCES (?)

↳ IID SOURCE.

Definition (Coin-Flip Source)

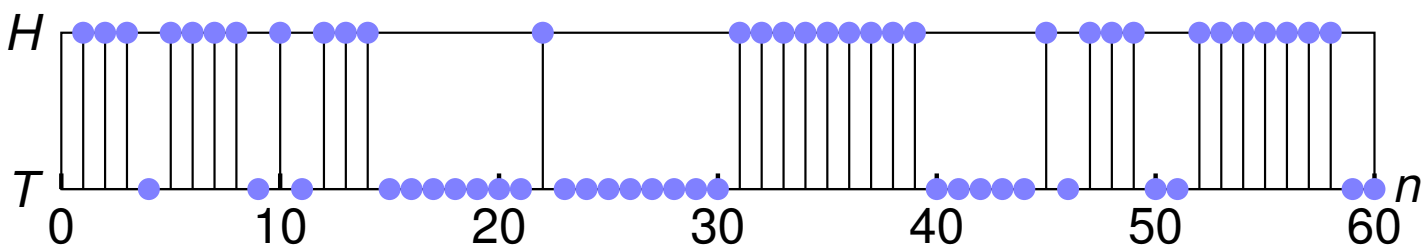
The source models a sequence $S_1, S_2, \ldots, S_n$ of n coin flips.

So $S_i \in \mathcal{A} = \{H, T\}$, where H stands for heads, T for tails, $i = 1, 2, \ldots, n$.

$p_{S_i}(H) = p_{S_i}(T) = \frac{1}{2}$ for all i, and coin flips are independent.

Hence,

$$p_{S_1,S_2,\ldots,S_n}(s_1, s_2, \ldots, s_n) = \frac{1}{2^n} \quad \text{for all } (s_1, s_2, \ldots, s_n) \in \mathcal{A}^n$$

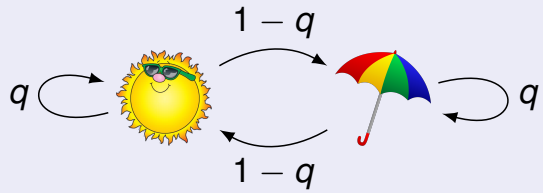


Definition (Sunny-Rainy Source)

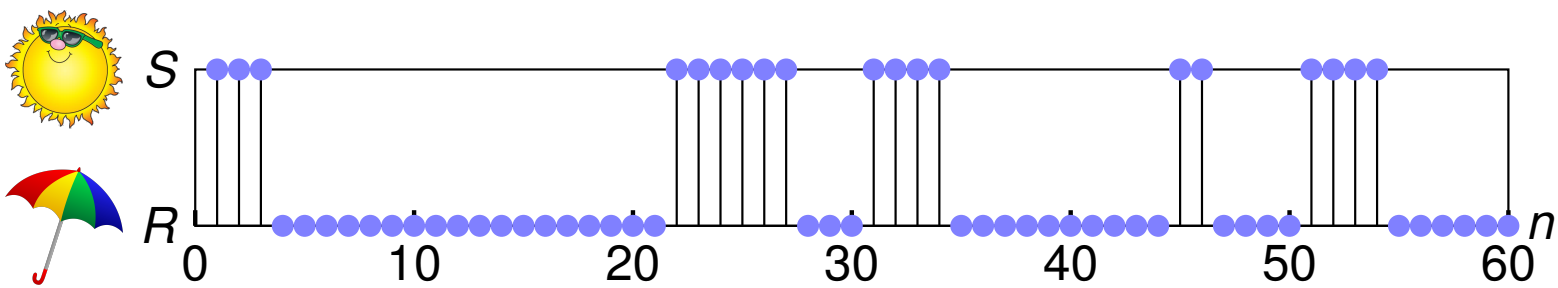
The source models a sequence $S_1, S_2, \ldots, S_n$ of weather conditions.

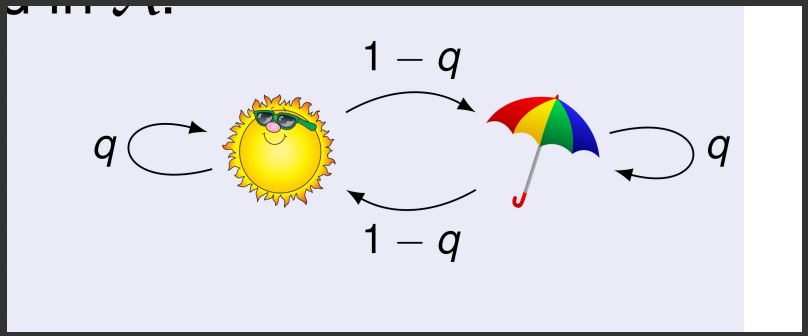
So $S_i \in \mathcal{A} = \{S, R\}$, where S stands for sunny, R for rainy, $i = 1, 2, \ldots, n$.

The weather on the first day is uniformly distributed in $\mathcal{A}$.



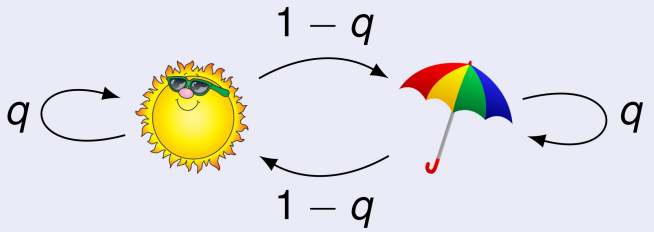
For all other days, with probability $q = \frac{6}{7}$ the weather is as for the day before.





S_1 : is selected according to $P_{S_1}(s) = \begin{cases} p, & \text{if } s = \text{sun} \\ 1-p, & \text{if } s = \text{rain} \end{cases}$

Ex: $S_1 \equiv$ sun

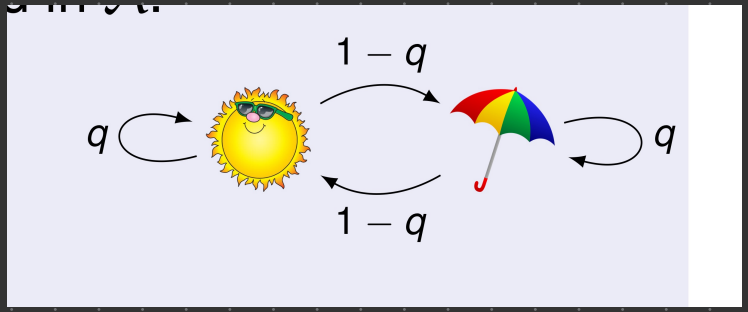


$$P(S_2 = \text{sun} \mid S_1 = \text{sun}) = q$$

$$P(S_2 = \text{rain} \mid S_1 = \text{sun}) = 1 - q$$

$$P(S_2 = \text{sun} \mid S_1 = \text{rain}) = 1 - q$$

$$P(S_2 = \text{rain} \mid S_1 = \text{rain}) = q$$



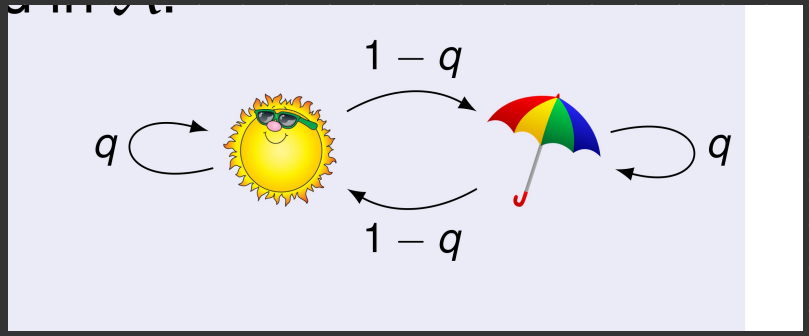
$$p(S_3 = \text{sun} \mid S_1 = \text{sun}, S_2 = \text{sun})$$

$$= p(S_3 = \text{sun} \mid S_2 = \text{sun}) = q$$

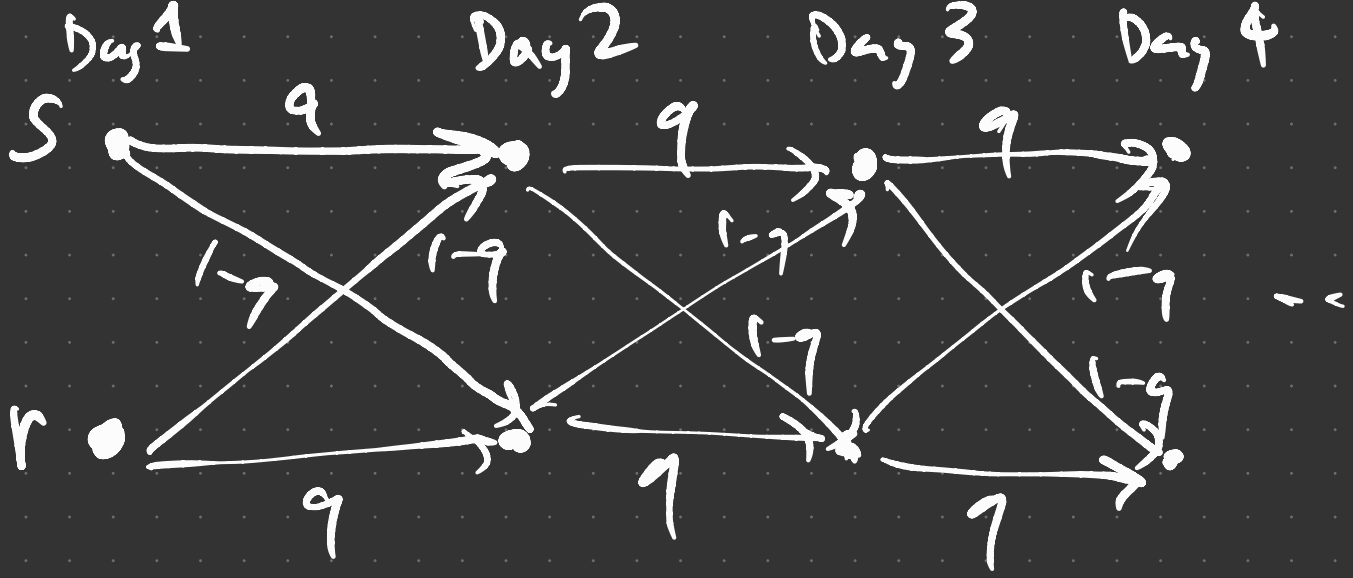
MORE GENERALLY:

$$p(S_n \mid S_1, S_2, \ldots S_{n-1}) =$$

$$p(S_n \mid S_{n-1})$$



EQUIVALENT PICTURE



Example

For the Sunny-Rainy source:

- $p_{S_1}(S) = \frac{1}{2}$
- $p_{S_1,S_2}(R,R) = p_{S_1}(R)p_{S_2|S_1}(R|R) = \frac{1}{2}q$
- $p_{S_1,S_2}(R,S) = p_{S_1}(R)p_{S_2|S_1}(S|R) = \frac{1}{2}(1-q)$
- $p_{S_1,S_2,S_3,S_4}(R,S,S,R) = \frac{1}{2}(1-q)q(1-q) = \frac{1}{2}q(1-q)^2$

In general, if c is the number of weather changes ($0 \leq c \leq n-1$), then

$$p_{S_1,S_2,\ldots,S_n}(s_1,s_2,\ldots,s_n) = \frac{1}{2}q^{n-1-c}(1-q)^c.$$

$$p(s_1, s_2, s_3, s_4)$$

BAYES

$$= p(s_1)\, p(s_2|s_1)\, p(s_3|s_1, s_2)\, p(s_4|s_1, s_2, s_3)$$

SUNNY - RAINY MODEL

$$= p(s_1)\, p(s_2|s_1)\, p(s_3|s_2)\, p(s_4|s_3)$$

$$p(x, y) = p(x)\, p(y|x)$$

$$p(\underbrace{s_1, s_2, s_3}_{z}, s_4) = p(z)\, p(s_4|z)$$

EXERCISE

Let $i = 2, 3, \ldots$

For the Sunny-Rainy source:

- Find $p_{S_i}(s_i)$
- Find $p_{S_i|S_{i-1}}(s_i|s_{i-1})$
- Are S_i and S_{i-1} independent?

$S_1 - S_2 - S_3 - \dots$

$p(S_i) \rightsquigarrow p(S_1 = R) = 1/2$

$$p(S_2) = \sum_{S_1} p(S_1, S_2)$$

$$p(S_2 = R) = p(S_1 = R, S_2 = R) + p(S_1 = S, S_2 = R)$$

$$= \frac{1}{2} q + \frac{1}{2}(1-q)$$

$$= \frac{1}{2}(q + 1 - q) = \frac{1}{2}.$$

$$p(S_3 = R) = \sum_{S_2} p(S_2, S_3 = R)$$

$$= \sum_{S_2} p(S_2)\, p(S_3 = R \mid S_2)$$

$$= p(S_2 = S)\, p(S_3 = R \mid S_2 = S) + p(S_2 = R)\, p(S_3 = R \mid S_2 = R)$$

$$= \frac{1}{2}(1 - q) + \frac{1}{2} \cdot q = 1/2$$

Solution (Sunny-Rainy source)

By definition, $p_{S_i|S_{i-1}}(j|k) = q$ if $j = k$ and $(1-q)$ otherwise.

Hence S_{i-1} and S_i are not independent.

To determine the statistic of the marginals, we use the law of total probability and induction to show that p_{S_i} is uniform.

It is true by definition for $i = 1$.

Suppose that p_{S_i} is uniform for $i = 1, \ldots, n-1$. We show that it is uniform also for $i = n$:

$$p_{S_n}(j) = \sum_{k \in \{S,R\}} p_{S_n|S_{n-1}}(j|k) p_{S_{n-1}}(k) = \frac{1}{2} \sum_{k \in \{S,R\}} p_{S_n|S_{n-1}}(j|k)$$
$$= \frac{1}{2}(q + (1-q)) = \frac{1}{2}.$$

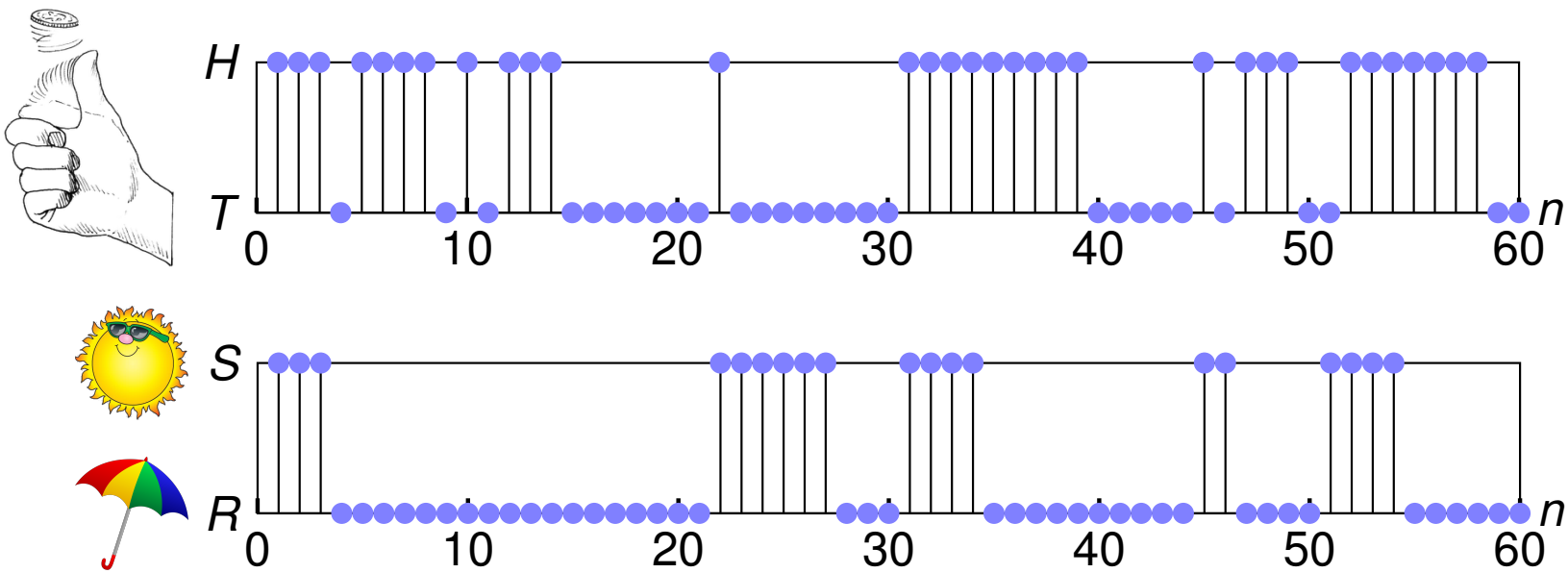
Hence the marginals are uniformly distributed (like for the Coin-Flip source).

EXERCISE

Let $i = 2, 3, \ldots$

For the Coin-Flip (CF) and Sunny-Rainy (SR) sources:

- Compute $H(S_i)$
- Compute $H(S_i|S_1, \ldots, S_{i-1})$



Solution ($H(S_i)$)

The entropy depends only on the distribution, and for a uniform distribution, it is the log of the alphabet's cardinality. Hence

$$H_{CF}(S_i) = H_{SR}(S_i) = \log 2 = 1$$

Solution ($H(S_i|S_1, \ldots, S_{i-1})$ for the Coin-Flip source)

S_i is independent of $S_1, \ldots, S_{i-1}$

Hence, $H(S_i|S_1, \ldots, S_{i-1}) = H(S_i)$.

Solution ($H(S_i|S_1, \ldots, S_{i-1})$ for the Sunny-Rainy source)

S_i depends only on S_{i-1}. Hence

$$H_{SR}(S_i|S_1 = s_1, \ldots, S_{i-1} = s_{i-1}) = H_{SR}(S_i|S_{i-1} = s_{i-1}).$$

When $S_{i-1} = k \in \{S, R\}$, the probabilities for S_i are q and $(1-q)$. Hence

$$H_{SR}(S_i|S_{i-1} = s_{i-1}) = -q \log q - (1-q) \log(1-q).$$

Taking the average on both sides yields

$$H_{SR}(S_i|S_{i-1}) = -q \log q - (1-q) \log(1-q).$$

For $q = \frac{6}{7}$, we have

$$H_{SR}(S_i|S_{i-1}) = -q \log q - (1-q) \log(1-q) = 0.592.$$

BACK TO THE THEORY

The main question is:

- For which sources does $R^*(\mathcal{S}) = \lim_{n\to\infty} \frac{H_D(S_1,S_2,\ldots,S_n)}{n}$ exist?
- We now introduce an alternative criterion.

TERMINOLOGY

The current (French) version of the textbook makes a difference between

- "source" S
- "source composée" $S_1, S_2, \ldots, S_n$
- "source étendue" $\mathcal{S} = S_1, S_2, \ldots$

This distinction has its merits, but we will not work with it in our class.

We will instead think of a source as described by its statistical property, and from it, it is implicit whether it produces one, n, or ∞ symbols.

For convenience, we do reserve the symbol $\mathcal{S}$ (calligraphic version of S) for sources that produce infinite sequences.

Definition

The source $\mathcal{S} = (S_1, S_2, \dots)$ is said to be regular if

$$H(\mathcal{S}) \overset{def}{=} \lim_{n \to +\infty} H(S_n) \quad \text{and}$$

$$H^*(\mathcal{S}) \overset{def}{=} \lim_{n \to +\infty} H(S_n | S_1, S_2, \dots, S_{n-1})$$

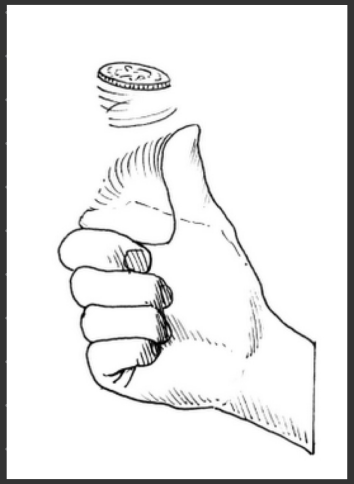
exist and are finite.

For a regular source $\mathcal{S}$, $H(\mathcal{S})$ is called the entropy of a symbol, and $H^*(\mathcal{S})$ the entropy rate.

Exercise: We have $H^*(\mathcal{S}) \leq H(\mathcal{S})$, with equality if the symbols are independent.

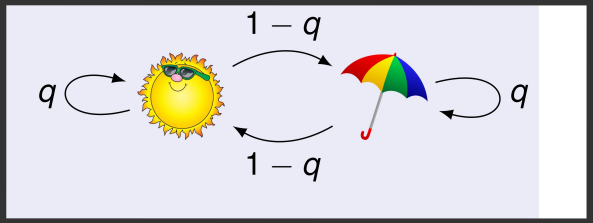
Ex: COIN FLIP

$H(S) = 1$ bit



$H^*(S) = 1$ bit

Ex: SUNNY-RAINY



$$H(\mathcal{S}) = \lim_{n\to\infty} H(S_n) = 1 \text{ bit}$$

$$H^{\$}(\mathcal{S}) = \lim_{n\to\infty} \underbrace{H(S_n \mid S_1, S_2, \ldots, S_{n-1})}_{H(S_n \mid S_{n-1}) = 0.592}$$

English - French Translation

symbol	English	French
$H(\mathcal{S})$	entropy of a symbol	entropie d'un symbole
$H^*(\mathcal{S})$	entropy rate	entropie par symbole

EXERCISE

Which of the following sources is regular?

1. The Coin-Flip source
2. The Sunny-Rainy source
3. Both

SOLUTION

Both

Main theorem

Theorem

For any regular source,

$$\lim_{n\to\infty} \frac{H_D(S_1, S_2, \ldots, S_n)}{n} = H_D^*(\mathcal{S}).$$

To prove this theorem, we need the following result that you have likely encountered earlier:

THEOREM (CESARO MEANS)

Let $a_1, a_2, \ldots$ be a real-valued sequence and let $c_1, c_2, \ldots$ be the sequence of running averages defined by

$$c_n = \frac{a_1 + a_2 + \cdots + a_n}{n}.$$

If $\lim_{n\to\infty} a_n$ exists, then $\lim_{n\to\infty} c_n$ also exists and

$$\lim_{n\to\infty} c_n = \lim_{n\to\infty} a_n.$$

PROOF OF THE MAIN THEOREM

By the chain rule of entropy,

$$\frac{H_D(S_1, S_2, \ldots, S_n)}{n} = \frac{H_D(S_1) + H_D(S_2|S_1) + \cdots + H_D(S_n|S_1, \ldots, S_{n-1})}{n}$$

and by the Cesàro means theorem, both sides converge to the limit of

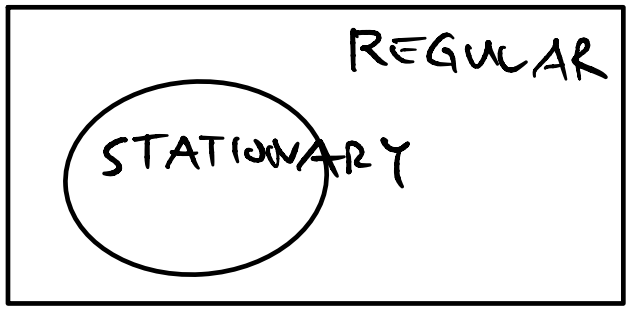
$$H_D(S_n|S_1, \ldots, S_{n-1})$$

which is

$$H_D^*(\mathcal{S}),$$

which thus completes the proof.

The Class of Stationary Sources



- The class of regular sources is a bit too abstract to most people's taste. It does not really give a good intuition for the types of sources for which the limit exists.

- One important subclass of regular sources are so-called **Stationary Sources.**

- This class is more intuitive and instructive. Therefore, we now discuss this class in more detail.

Definition (Stationary source)

A source $S_1, S_2, \ldots$ is stationary if, for all positive integers n and k, the blocks $(S_1, S_2, \ldots, S_n)$ and $(S_{k+1}, S_{k+2}, \ldots, S_{k+n})$ have the same statistic.

This implies

- $p_{S_1} = p_{S_m}$ for all m $\quad (n=1,\ k=m-1)$
- $p_{S_1,S_2} = p_{S_m,S_{m+1}}$ for all m
- $p_{S_1,S_2,S_3} = p_{S_m,S_{m+1},S_{m+2}}$ for all m
- $p_{S_m,S_t} = p_{S_{m+\ell},S_{t+\ell}}$ for all m, t, ℓ (Can you prove this?)
- etc.
- (For any subset $\mathcal{I}$ of indices, $p_{S_\mathcal{I}} = p_{S_{k+\mathcal{I}}}$.)

A source is stationary if its distribution is unaffected by an index shift (time shift).

- Coin-Flip and Sunny-Rainy are stationary.

↑ with $p(s_1)$ = uniform.

THEOREM

Stationary sources are regular, implying that

$$\lim_{n\to\infty} \frac{H_D(S_1, S_2, \ldots, S_n)}{n} = H_D^*(\mathcal{S}).$$

Moreover, for stationary sources,

$$\frac{H_D(S_1, S_2, \ldots, S_n)}{n}$$

is non-increasing in n.

PROOF: "REGULAR" MEANS:

1) $H(\mathcal{S}) = \lim_{n\to\infty} H(S_n)$ exists ✓

2) $H^*(\mathcal{S}) = \lim_{n\to\infty} H(S_n | S_1, \dots, S_{n-1})$ exists

$H(S_1)$, $H(S_2|S_1)$, $H(S_3|S_1,S_2)$, $H(S_4|S_1S_2S_3)$, ...

$= H(S_2)$ $= H(S_3|S_2)$ $= H(S_4|S_2S_3)$

$\vdots$ $= H(S_4|S_3)$ $= \dots$

$\vdots$

$= H(S_n) \geq H(S_n|S_{n-1}) \geq H(S_n|S_{n-1}, S_{n-2}) \geq \dots$

Proof: $H_D(\mathcal{S}) = \lim_{n\to\infty} H_D(S_n)$ is well since $H_D(S_n)$ is constant.

Moreover, we know that $H_D(S_2|S_1) \leq H_D(S_2)$ but since the source is stationary, we also have that $H_D(S_2) = H_D(S_1)$, thus,

$$H_D(S_2|S_1) \leq H_D(S_1).$$

Next, we know that $H_D(S_3|S_1, S_2) \leq H_D(S_3|S_2)$ but since the source is stationary, we also have that $H_D(S_3|S_2) = H_D(S_2|S_1)$, thus,

$$H_D(S_3|S_1, S_2) \leq H_D(S_2|S_1).$$

Continuing in this manner, we find that

$$H_D(S_1), H_D(S_2|S_1), \ldots, H_D(S_n|S_1, \ldots, S_{n-1})$$

is a non-increasing sequence. Moreover, it is bounded from below by zero.

Hence $H_D^*(\mathcal{S}) = \lim_{n\to\infty} H_D(S_n|S_1, \ldots, S_{n-1})$ is well defined.

Hence a stationary source is regular.

It remains to be shown that $\frac{H_D(S_1,\ldots,S_n)}{n}$ in non-increasing, i.e., that

$$\frac{H_D(S_1,\ldots,S_n)}{n} \geq \frac{H_D(S_1,\ldots,S_{n+1})}{n+1}.$$

We prove that

$$(n+1)H_D(S_1,\ldots,S_n) \geq nH_D(S_1,\ldots,S_{n+1}),$$

$= H(S_1 \cdots S_n) + H(S_{n+1}|S_1 \cdots S_n)$

or equivalently, that

$$H_D(S_1,\ldots,S_n) \geq nH_D(S_{n+1}|S_1,\ldots,S_n).$$

Namely:

$$H(S_1, S_2, \ldots S_n)$$

$$= H(S_1) + H(S_2|S_1) + H(S_3|S_1, S_2) + \ldots + H(S_n|S_1 \ldots S_{n-1})$$

$$= H(S_n) + H(S_n|S_{n-1}) + H(S_n|S_{n-2}S_{n-1}) \ldots H(S_n|S_1 \ldots S_{n-1})$$

$$\geq H(S_n|S_1 \ldots S_{n-1}) + H(S_n|S_1 \ldots S_{n-1}) + \ldots$$

$$= n\, H(S_n|S_1 \ldots S_{n-1})$$

$$= n\, H(S_{n+1}|S_2, \ldots, S_n) \geq n\, H(S_{n+1}|S_{n-1})$$

$$
\begin{aligned}
H_D(S_1, \ldots, S_n) &= H_D(S_1) + H_D(S_2|S_1) + \cdots + H_D(S_n|S_1, \ldots, S_{n-1}) \\
&\overset{(\star)}{=} H_D(S_{n+1}) + H_D(S_{n+1}|S_n) + \cdots + H_D(S_{n+1}|S_2, \ldots, S_n) \\
&\overset{(\star\star)}{\geq} H_D(S_{n+1}|S_1, \ldots, S_n) + \cdots + H_D(S_{n+1}|S_1, \ldots, S_n) \\
&= nH_D(S_{n+1}|S_1, \ldots, S_n),
\end{aligned}
$$

where

- $(\star)$ follows from the source stationarity;
- $(\star\star)$ holds because "conditioning reduces entropy".

□

Exercise

Determine $H(S_1, S_2, \ldots, S_n)$ for the Coin-Flip source.

Solution

The source produces independent and identically distributed symbols. Hence

$$H(S_1, S_2, \ldots, S_n) \overset{\text{(indep.)}}{=} H(S_1) + H(S_2) + \cdots + H(S_n) \overset{\text{(identically distributed)}}{=} nH(S_1)$$

Moreover, the distribution is uniform, therefore $H(S_1) = 1$ bit. Putting things together,

$$H(S_1, S_2, \ldots, S_n) = n \text{ bits}$$

EXERCISE

Determine $H(S_1, S_2, \ldots, S_n)$ for the Sunny-Rainy source with $q = \frac{6}{7}$.

SOLUTION

$$H(S_1, S_2, \ldots, S_n) = H(S_1) + H(S_2|S_1) + \cdots + H(S_n|S_1, \ldots, S_{n-1})$$

For $i = 2, 3, \ldots, n$, the statistic of S_i depends only on S_{i-1}. Hence

$$H(S_i|S_1, S_2, \ldots, S_{i-1}) = H(S_i|S_{i-1})$$

$$H(S_1, S_2, \ldots, S_n) = H(S_1) + H(S_2|S_1) + \cdots + H(S_n|S_{n-1})$$

We have already determined that $H(S_1) = 1$ bit and $H(S_i|S_{i-1}) = 0.592$ bits. Therefore

$$H(S_1, S_2, \ldots, S_n) = 1 + 0.592(n-1) \text{ bits}$$

We summarize a main result of source coding.

THEOREM

Let $S_1, S_2, \ldots$ be the infinite sequence produced by a regular source $\mathcal{S}$ (which, in particular, includes stationary sources).

1. By encoding blocks of symbols into D-ary codewords, the average codeword-length per symbol of a uniquely decodable code can be made as close as desired to $H_D^*(\mathcal{S})$.
2. No uniquely decodable D-ary code can achieve a smaller average codeword-length.

The above result justifies considering $H_D^*(\mathcal{S})$ as a **measure of information**.

In particular, $H_2^*(\mathcal{S})$ is a measure for the number of bits per source symbol produced by the source $\mathcal{S}$.

For an iid source $\mathcal{S}$, $H_D^*(\mathcal{S}) = H_D(\mathcal{S})$.

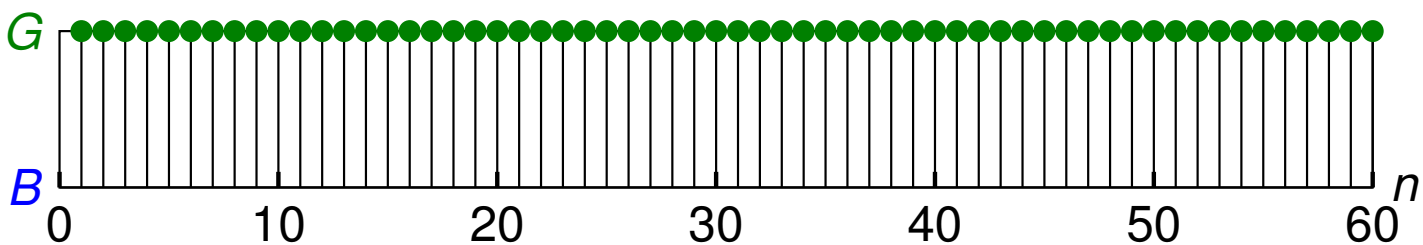
Time permitting, we will cover additional source examples, as presented in the following slides.

In any case, some of these (and yet further examples) will be covered in the homework.

Definition (Green-Blue Source)

The source models a sequence $S_1, S_2, \ldots, S_n$ of a person's votes from the alphabet $\mathcal{A} = \{G, B\}$.

- The first vote is chosen uniformly in $\mathcal{A}$.
- The next votes are always identical to the initial vote, i.e., $S_i = S_1$, $i = 2, \ldots, n$.

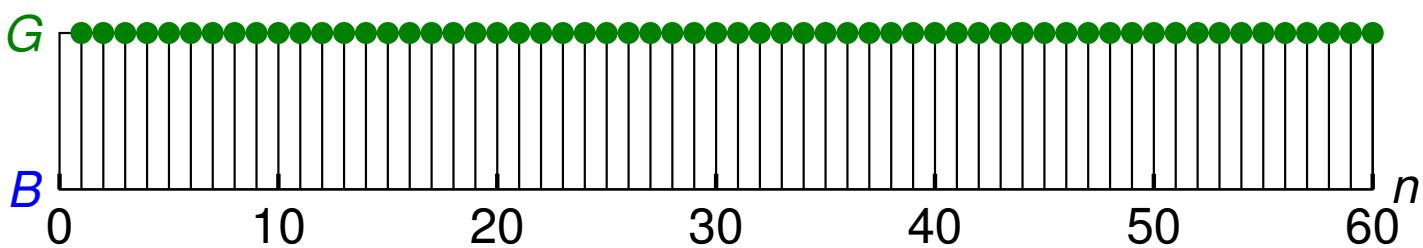


EXERCISE

Let $n = 1, 2, \ldots$

For the Green-Blue source $\mathcal{S} = S_1, S_2, \ldots$:

1. Find $p_{S_n}(s_n)$
2. Find $H_{GB}(S_n)$
3. Find $H_{GB}(S_n|S_{n-1})$
4. Is the source regular ? If yes, determine its symbol entropy $H_{GB}(\mathcal{S})$ and its entropy rate $H^*_{GB}(\mathcal{S})$.



SOLUTION

S_1 is uniformly distributed in $\mathcal{A} = \{G, B\}$.

$S_n = S_1$. Hence S_n is uniformly distributed in $\mathcal{A} = \{G, B\}$. Hence $H_{GB}(S_n) = 1$.

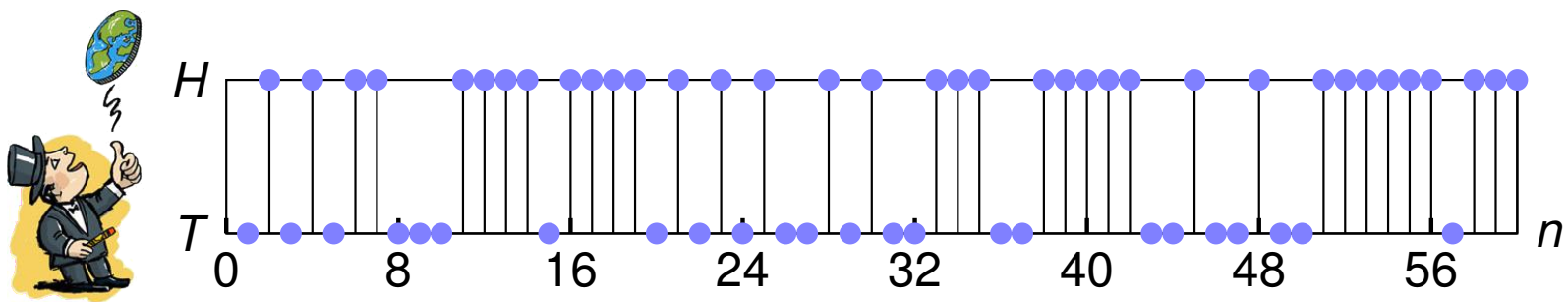
For $n = 2, 3, \ldots$, the value of S_n is a deterministic function of S_{n-1}. Hence, $H_{GB}(S_n|S_{n-1}) = 0$.

The source is regular, with $H_{GB}(\mathcal{S}) = 1$ and $H^*_{GB}(\mathcal{S}) = 0$.

Definition (Weekly-Coin-Flip Source)

The source models a sequence $S_1, S_2, \ldots$ of coin flips in $\mathcal{A} = \{H, T\}$ such that

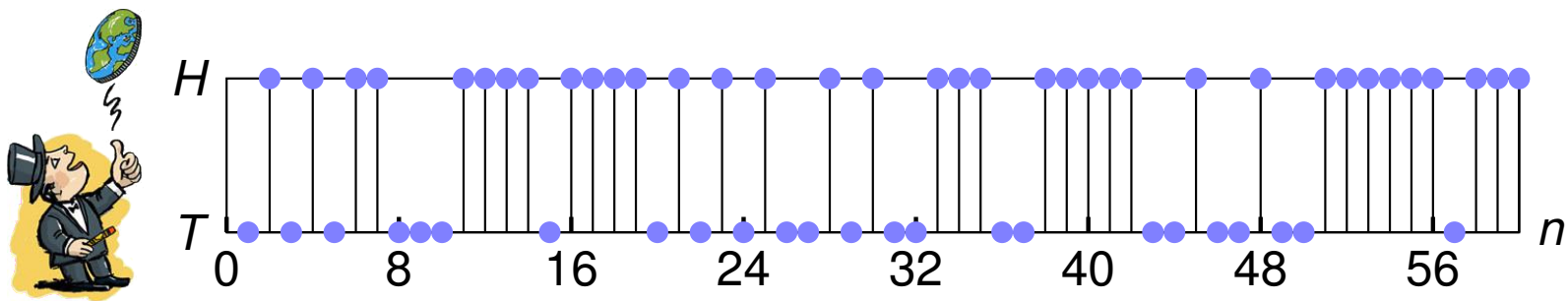
$$p_{S_{i+7k}}(T) = \frac{1}{i}, \quad i = \{1, 2, \ldots, 7\},\ k \in \mathbb{N}.$$



EXERCISE

Let $n = 1, 2, \ldots$ and $\mathcal{S} = S_1, S_2, \ldots$ be a Weekly-Coin-Flip source.

- Is it a regular source?
- If yes, determine its symbol entropy $H_{WCF}(\mathcal{S})$ and its entropy rate $H^*_{WCF}(\mathcal{S})$.



SOLUTION

$$H(S_1) = H(S_8) = \cdots = H(S_1 + 7k) = 0 \text{ bits}$$
$$H(S_2) = H(S_9) = \cdots = H(S_2 + 7k) = 1 \text{ bits}$$
$$\cdots$$

$\lim_{n\to\infty} H_{WCF}(S_n)$ does not exist (because $0 \neq 1$), hence the source is not regular.

Additional Questions of interest include:

- What if the source alphabet is not finite?
- What if we do not know the source distribution $p_X(x)$? (Universal source coding)

What if the source alphabet is infinite?

- In all of our previous discussion on actual codes, we have assumed that the source alphabet is discrete and finite.
- What if it is discrete but infinite?
- ... is this just an academic endeavour?
- In this class, we only touch the top of this iceberg...

BINARY PREFIX-FREE CODE FOR POSITIVE INTEGERS

The set of positive integers is infinite and no probability is assigned to its elements. Hence we cannot use Huffman's construction to encode integers.

First Attempt to Encode Positive Integers: "Standard Method"

n	$c(n)$
1	1
2	10
3	11
4	100
5	101
$\vdots$	$\vdots$

The code is not prefix-free.

The length of $c(n)$ is $l(n) = \lfloor \log_2 n \rfloor + 1$.

Note: The first digit is always 1.

Second Attempt: “Elias Code 1"

We prefix code $c(n)$ with $l(n) - 1$ zeros.

n	$c_1(n)$
1	1
2	010
3	011
4	00100
5	00101
$\vdots$	$\vdots$

The code is prefix-free. (Codewords of different length cannot have the same number of leading zeros.)

The length of $c_1(n)$ is

$$l_1(n) = l(n) - 1 + l(n) = 2\lfloor \log_2 n \rfloor + 1.$$

Note: we are essentially doubling the length to make the code prefix-free.

Third Attempt: “Elias Code 2"

Instead of $l(n) - 1$ zeros followed by a 1, we prefix with $c_1\big(l(n)\big)$, which is also prefix-free (hence can be identified). Like the zeros, it tells the length of the codeword.

Notation: $\tilde{c}(n)$ is $c(n)$ without the leading 1.

n	$c(n)$	$l(n)$	$c_1(n)$	$c_1\big(l(n)\big)\tilde{c}(n)$
1	1	1	1	$c_1(1) = 1$
2	10	2	010	$c_1(2)\,0 = 010\,0$
3	11	2	011	$c_1(2)\,1 = 010\,1$
4	100	3	00100	$c_1(3)\,00 = 011\,00$
5	101	3	00101	$c_1(3)\,01 = 011\,01$
⋮	⋮			

The code is prefix-free.

The codeword length is
$l_2(n) = l_1\big(l(n)\big) + l(n) - 1 = 2\lfloor \log_2(\lfloor \log_2 n \rfloor + 1) \rfloor + 1 + \lfloor \log_2 n \rfloor$.

What if the source distribution is not known?

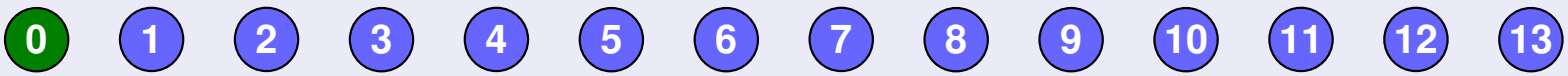
- Universal source coding.

- Practically important algorithms: “Lempel-Ziv” (LZ77, LZ78). Time permitting, we briefly discuss how they work. An analysis is beyond the scope of AICC-2.

CHALLENGE FOR NEXT LECTURE

EXERCISE

There are 14 billiard balls numbered as shown:



Among balls 1 - 13, at most one **could** be heavier/lighter than the others.

What is the minimum number of weightings to simultaneously determine:

- if one ball is different . . .
- if there is such a ball, which one, . . .
- and whether the different ball is heavier/lighter.