# I232 Information Theory
# Chapter 6: Source Coding for Markov Sources

Brian Kurkoski

Japan Advanced Institute of Science and Technology

2023 April

# Outline

# 6.1 Markov Chains

# Markov Chains

Markov chains were introduced in Section 3.3. This section generalizes Markov chains, which are a special case of a stochastic process.

## 6.1.1 Stochastic Processes

A *stochastic process* $X_1, X_2, \ldots, X_n$ is an indexed sequence of random variables (stochastic means random). The random variable $X_t$ is called the *state* at time $t$. We will refer to $t$ as time, but could also refer to distance, etc. The stochastic process has a joint probability distribution:

$$\Pr[X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n] = p_{\mathbf{X}}(x_1, x_2, \ldots, x_n)$$

Let $X_t$ take values from the set of size $m$, $\mathcal{X} = \{1, 2, \ldots, m\}$.

- $\mathcal{X}$ is called the *state space*
- X and $x \in \mathcal{X}$ are called the *state*
- $s_t(x) = \Pr[X_t = x]$ is the probability of state $x$ at time $t$.
- The *state vector* is $\mathbf{s}$:

$$\mathbf{s}_t = [s_t(1), s_t(2), \ldots, s_t(m)].$$

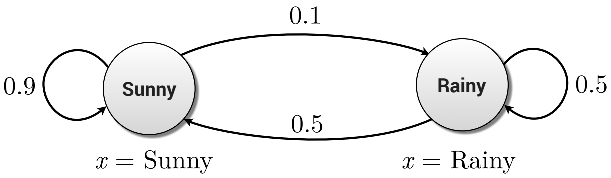
- The state vector at time $t = 1$ is the *initial state vector*:

## 6.1.2 Markov Chain

A concrete example of a Markov source is a simple model of the weather with $\mathcal{X} = \{\text{Sunny}, \text{Rainy}\}$. Given today is Sunny, the probability that tomorrow is Rainy is $\Pr[\mathsf{X}_{n+1} = \text{Rainy} | \mathsf{X}_n = \text{Sunny}] = 0.1$. All the possible transitions are given by a transition matrix:

$$\mathbf{P} = \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix}.$$

The state-transition diagram is shown below.



## 6.1.3 Steady-State Distribution (Stationary Distribution)

Recall the state vector $\mathbf{s}_{t+1}$ at time $t+1$ can be conveniently expressed in matrix form:

$$\mathbf{s}_{t+1} = \mathbf{s}_t \cdot \mathbf{P}.$$

### Definition

A *steady-state distribution* or stationary distribution $\mathbf{z}$ is a state vector such that:

$$\mathbf{z} = \mathbf{z}\mathbf{P}.$$

The state vector of the steady-state distribution does not change after taking one or more steps in the Markov chain.

★1

# Example: Steady-State Distribution

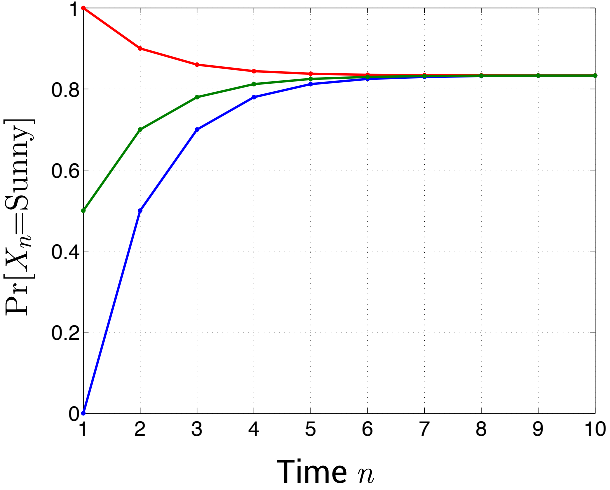



Figure 1: Evolution of state vector when the initial state distribution is $[0, 1]$ (blue), or $[0.5, 0.5]$ (green) or $[1, 0]$ (red). All three evolve towards the same steady-state distribution $[5/6, 1/6]$.

# 6.2 Entropy Rate

# 6.2.1 Entropy Rate of Stochastic Processes

## Definition

The *entropy rate* of a stochastic process $\mathsf{X}_1, \mathsf{X}_2, \ldots$ is defined by:

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(\mathsf{X}_1, \mathsf{X}_2, \ldots, \mathsf{X}_n)$$

when the limit exists.

## Definition

The *conditional entropy rate* is:

$$H'(\mathcal{X}) = \lim_{n\to\infty} H(\mathsf{X}_n | \mathsf{X}_{n-1}, \ldots, \mathsf{X}_2, \mathsf{X}_1)$$

- $H(\mathcal{X})$ is the entropy per symbol of $n$ random variables.
- $H'(\mathcal{X})$ is the conditional entropy of the last random variable, given the past.

# 6.2.2 Entropy Rate of Stationary Stochastic Processes

## Proposition

For a stationary stochastic process, $H(\mathsf{X}_n|\mathsf{X}_{n-1},\ldots,\mathsf{X}_1)$ is non-increasing in $n$ and has a limit $H'(\mathcal{X})$.

## Proposition

For a stationary stochastic process:

$$H(\mathcal{X}) \;\; = \;\; H'(\mathcal{X}).$$

# Source Codes for Stationary Processes

Now consider a *source code* for a stationary stochastic process $\mathsf{X}_1, \mathsf{X}_2, \ldots, \mathsf{X}_n$.

Form a vector source code with code rate $R_n$:

$$R_n = \frac{1}{n} \sum_{i=1}^{n} p_{\mathbf{X}}(\mathbf{x}) \ell(\mathbf{x})$$

Let $R_n^*$ be the rate of an optimal code, the code having the lowest possible rate.

## Proposition

If $\mathsf{X}_1, \mathsf{X}_2, \ldots . \mathsf{X}_n$ is a stationary stochastic process,

$$\lim_{n \to \infty} R_n^* = H(\mathcal{X})$$

where $H(\mathcal{X})$ is the entropy rate of the process.

## 6.2.3 Entropy Rate of Stationary Markov Chains

Any stationary source can be compressed at the entropy rate of the process.

Stationary Markov process: The initial state is equal to the steady-state distribution. Thus, $H(\mathsf{X}_{n+1}|\mathsf{X}_n) = H(\mathsf{X}_2|\mathsf{X}_1)$.

### Proposition

For a stationary Markov chain, the entropy rate is given by:

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

# Entropy Rate of Stationary Markov Chain

### Proposition

Let $\mathsf{X}_1, \mathsf{X}_2, \ldots,$ be a stationary Markov chain with steady-state distribution $\mathbf{z}$ and transition matrix $\mathbf{P}$. Let $\mathsf{X}_1 \sim \mathbf{z}$. Then the entropy rate is:

$$H(\mathcal{X}) = -\sum_{i=1}^{m}\sum_{j=1}^{m} z_i p_{i,j} \log p_{i,j}$$

## 6.3 On Compression of Text

Consider the compression of English text.

Consider a simplified 27-symbol alphabet:

A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y, Z, _

That is, we ignore capitalization (a $\rightarrow$ A) and punctuation (. , ? !, etc.)

Clearly, $|\mathcal{X}| = 27$, so compression at a rate $R = \log 27 \approx 4.75$ bits is possible.

Big question: what is the *lowest* possible rate of compression for English.

First step: use the iid source distribution. What is the most common letter in English? What is the least common letter?

# Compression Using Assuming IID

"A" and "E" are common letters include. "J" and "Q" and "Z" are uncommon letters.

| Letter | Probability | Letter | Probability | Letter | Probability |
|---|---|---|---|---|---|
| A | 8.29% | J | 0.21% | S | 6.33% |
| B | 1.43 | K | 0.48 | T | 9.27 |
| C | 3.68 | L | 3.68 | U | 2.53 |
| D | 4.29 | M | 3.23 | V | 1.03 |
| E | 12.8 | N | 7.16 | W | 1.62 |
| F | 2.20 | O | 7.28 | X | 0.20 |
| G | 1.71 | P | 2.93 | Y | 1.57 |
| H | 4.54 | Q | 0.11 | Z | 0.09 |
| I | 7.16 | R | 6.90 | | |

$$H(X) = -\sum_{i=1}^{27} p_i \log p_i \approx 4.17 \text{ bits } < 4.75 \text{ bits}$$

| $i$ | $a_i$ | $p_i$ |
|---|---|---|
| 1 | a | 0.0575 |
| 2 | b | 0.0128 |
| 3 | c | 0.0263 |
| 4 | d | 0.0285 |
| 5 | e | 0.0913 |
| 6 | f | 0.0173 |
| 7 | g | 0.0133 |
| 8 | h | 0.0313 |
| 9 | i | 0.0599 |
| 10 | j | 0.0006 |
| 11 | k | 0.0084 |
| 12 | l | 0.0335 |
| 13 | m | 0.0235 |
| 14 | n | 0.0596 |
| 15 | o | 0.0689 |
| 16 | p | 0.0192 |
| 17 | q | 0.0008 |
| 18 | r | 0.0508 |
| 19 | s | 0.0567 |
| 20 | t | 0.0706 |
| 21 | u | 0.0334 |
| 22 | v | 0.0069 |
| 23 | w | 0.0119 |
| 24 | x | 0.0073 |
| 25 | y | 0.0164 |
| 26 | z | 0.0007 |
| 27 | – | 0.1928 |

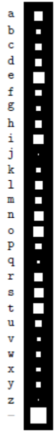



Figure 2.1. Probability distribution over the 27 outcomes for a randomly selected letter in an English language document

# Compression Using Markovity

Conditioning reduces entropy:

$$H(\mathsf{X}_{10}) > H(\mathsf{X}_{10}|\mathsf{X}_9) > H(\mathsf{X}_{10}|\mathsf{X}_8\mathsf{X}_9) > H(\mathsf{X}_{10}|\mathsf{X}_7\mathsf{X}_8\mathsf{X}_9) > \cdots$$

We know that certain letter pairs occur often in English:

THIS THAT THE THEN THOUGH

LOOK GOOD BOOK

What letter follows Q in the following sentences:

AT THE CASTLE WE MET THE Q__

THIS IS THE LIBRARY, PLEASE BE Q__

# First-order Markovity of English

Lex X denote one letter of English, with pX(x).

Now consider a sequence of letters:

$X_1\ X_2\ X_3\ X_4\ X_5\ X_6\ X_7\ X_8\ X_9\ X_{10}\ X_{11}\ X_{12}\ X_{13}\ \ldots.$

For example

T I M E _ F L I E S _ L I K E....

We can evaluate conditional probabilities

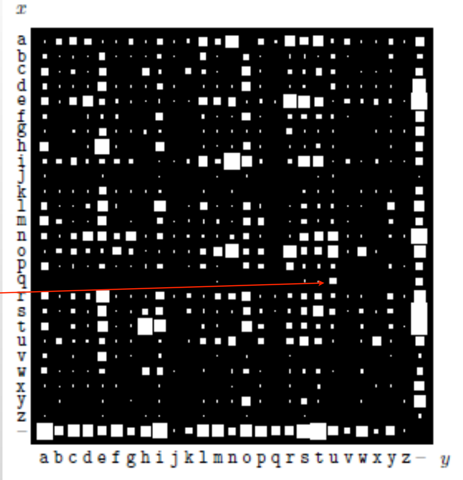
$\Pr[X_n = a \mid X_{n\text{-}1} = b]$

For example "q" usually comes before "u"

$\Pr[X_n = \text{"}\mathbf{u}\text{"} \mid X_{n\text{-}1} = \text{"}\mathbf{q}\text{"}]$

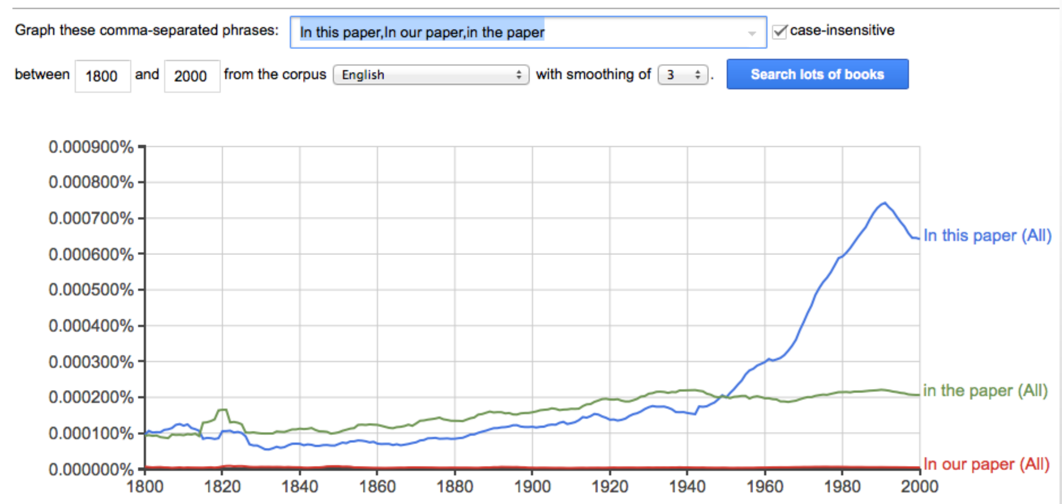
Square size proportional to probability

$x_{n-1}$




$x_n$

# Google N-Gram Viewer




What about higher order models?

For words not letters, more than 3-gram is hard

# Shannon's Human Experiment

**Prediction and Entropy of Printed English**

By C. E. SHANNON

(*Manuscript Received Sept. 15, 1950*)

A new method of estimating the entropy and redundancy of a language is described. This method exploits the knowledge of the language statistics possessed by those who speak the language, and depends on experimental results in prediction of the next letter when the preceding text is known. Results of experiments in prediction are given, and some properties of an ideal predictor are developed.

Show humans some text, and ask them to guess the next letter.

**Entropy is around 2.5 bits/letter**

the original text; the second line contains a dash for each letter correctly guessed. In the case of incorrect guesses the correct letter is copied in the second line.

```
(1) THE ROOM WAS NOT VERY LIGHT A SMALL OBLONG                    (8)
(2) ----ROO------NOT-V-----I------SM----OBL----
(1) READING LAMP ON THE DESK SHED GLOW ON
(2) REA----------O------D----SHED-GLO--O--
(1) POLISHED WOOD BUT LESS ON THE SHABBY RED CARPET
(2) P-L-S-----O---BU--L-S--O------SH-----RE--C------
```

# Let's Do Our Own Experiment

## In-Class Experiment: Compression of Text

Example. What is your best guess of the letter ?

- INFORMATION_THEO?
- JAIST_IS_A_UNIVER?
- PROBABILITY_?

# Let’s Do Our Own Experiment

In-Class Experiment: Compression of Text

From *Harry Potter and the Prisoner of Azkaban*

*Go to LMS on line. Make your best guess of the next 30 letters. Any letter or a space. Must type something*

NON_MAGIC_PEOPLE_WERE_PARTICULARLY_AFRAID_OF_MAGIC_IN_
MEDIEVAL_TIMES_?

# Source Coding for Markov Sources — What You Should Have Learned

- A stochastic process is an indexed sequence of random varibles
- A Markov chain is a stochastic process: future depends on the present, but not the past
- Markov sources can be compressed at lower rates because the Markov chain expresses redundancy in the source

# Class Info

- Tutorial Hours: Monday, May 1 at 13:30. Ask questions about Homework.
- Homework 3 and 4 on LMS. Deadline: Monday, May 1 at 18:00. Be sure to use most recent version of Lecture Notes.
- No Class on May 3. Enjoy Golden Week
- Next lecture: Monday, May 8 at 9:00. Channel Coding and Channel Capacity. There will be a pop quiz
- Homework 5 and 6 on LMS (soon). Deadline Monday, May 8 at 18:00.
- Midterm exam on May 15 at 13:30.

# Midterm Exam

The exam is closed book. You may use:

- One page of notes, A4-sized paper, double-sided OK.
- Blank scratch paper

You may not use anything else: No printed materials, including books, lecture notes, and slides. No notes (except as above). No internet-connected devices. No calculators. You may need to perform a $2 \times 2$ matrix inverse.

Exam Content

- Covers Chapters 1–6
- Study Homework 1–6. Solutions to Homework 1–6 are provided.
- No programming questions.

Practice problems will be provided.