I232 Information Theory
Chapter 5: Source Coding for Memoryless Sources

Brian Kurkoski

Japan Advanced Institute of Science and Technology

2023 April

Just for Fun: Eve's Birthday

Eve has just become friends with Michael and David, and they want to know when her birthday is. Eve gives them a list of 10 possible dates:

4 Mar, 5 Mar, 8 Mar

4 Jun, 7 Jun

1 Sep, 5 Sep

1 Dec, 2 Dec, 8 Dec

Eve tells the month to Michael, and the day to David.

Michael said, "I don't know Eve's birthday, but I know that David does not know either."

David said: "At first I did not know Eve's birthday. But now I know"

Michael said: "Now I know Eve's birthday, too"

What is Eve's birthday? How is this possible?

Homework: Method of Lagrange Multipliers

The method of Lagrange multipliers finds the local maximum and minimum of a function subject to equality constraints:

$$\text{optimize } f(\mathbf{x}) \text{ such that } g(\mathbf{x}) = 0$$

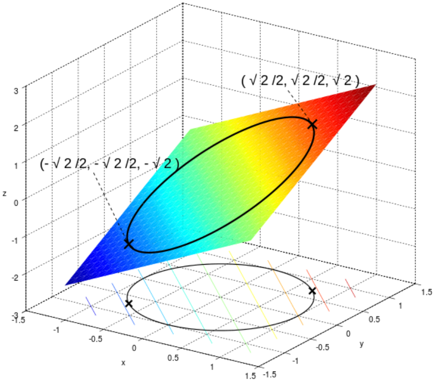
The general method is:

1. Form the Lagrangian $L = f(\mathbf{x}) + \lambda g(\mathbf{x})$
2. Take the partial derivatives $\frac{\partial}{\partial x_i} L(\mathbf{x}, \lambda) = 0$
3. Find λ
4. Use λ to find the optimal $\mathbf{x}$

Example: Method of Lagrange Multipliers

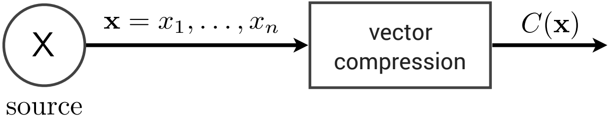
Solve the following problem using the method of Lagrange multipliers.

Optimize $f(x, y) = x + y$ subject to the constraint $x^2 + y^2 = 1$.



Outline

5.1 Vector Source Coding

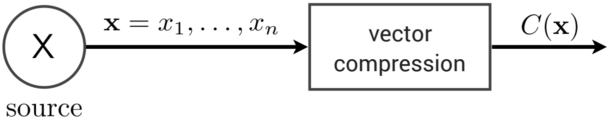


Definition

The *code rate* R for a vector source code C is:

$$R = \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{X}^n} p_{\mathbf{X}}(\mathbf{x}) \ell(\mathbf{x})$$

Vector Source Coding



Proposition

Vector Source Coding Theorem Let $\mathbf{X} = (\mathsf{X}_1, \ldots, \mathsf{X}_n)$ be n independent and identically distributed random variables X_i, with $\mathsf{X}_i \sim p_{\mathsf{X}}(x)$ having entropy $H(\mathsf{X})$. Let $\epsilon' > 0$. Then, there exists a vector source code with rate R that satisfies:

$$R \leq H(\mathsf{X}) + \epsilon',$$

for n sufficiently large.

This chapter shows:

- How to prove a typical theorem in information theory.
- We study the AEP for the simplest case, single variable compression.
- More advanced proofs, such as the channel coding theorem, follow the same basic principles.

5.2 Sample Entropy and Typical Sets

5.2.1 Sample Entropy

Definition

For any fixed sequence $\mathbf{x}$, jointly distributed according to $p_{\mathsf{X}}(\mathbf{x})$, the *sample entropy* is defined as:

$$-\frac{1}{n}\log p_{\mathsf{X}}(\mathbf{x}).$$

When the X_i are i.i.d. with distribution $p_{\mathsf{X}}(x_i)$, the sample entropy of $\mathbf{x} = x_1 x_2 \ldots x_n$ is:

$$-\frac{1}{n}\sum_{i=1}^{n}\log p_{\mathsf{X}}(x_i) \tag{1}$$

Note that (1) is similar to the sample mean.

But an important difference is that to compute the sample entropy, knowledge of $p_{\mathsf{X}}(\mathbf{x})$ is needed.

Example: Sample Entropy

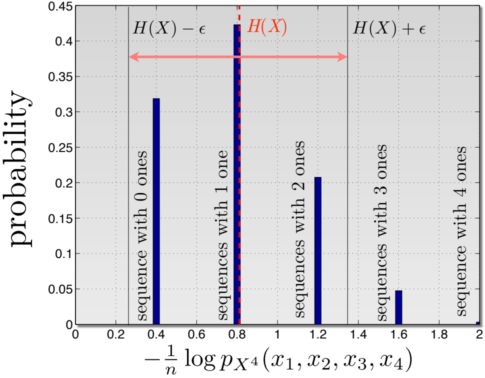
Consider $n = 4$ i.i.d. {0,1} binary random variables $\mathsf{X}_1\mathsf{X}_2\mathsf{X}_3\mathsf{X}_4$ with $p_{\mathsf{X}}(x) = [\frac{3}{4}, \frac{1}{4}]$ that is, $p_{\mathsf{X}}(1) = \frac{1}{4}$.

Find the sample entropy of all possible sequences.

★1

Example 5.4: Distribution of Sample Entropy

$\mathbf{x}$	$p_{\mathbf{X}}(\mathbf{x})$	sample entropy $-\frac{1}{n}\log p_{\mathbf{X}}(\mathbf{x})$
0 0 0 0	0.31641	0.41504
1 0 0 0	0.10547	0.81128
0 1 0 0	0.10547	0.81128
0 0 1 0	0.10547	0.81128
0 0 0 1	0.10547	0.81128
1 1 0 0	0.03516	1.20752
1 0 1 0	0.03516	1.20752
0 1 1 0	0.03516	1.20752
1 0 0 1	0.03516	1.20752
	$\vdots$	



5.2.2 Typical Sets and Typical Sequences

Definition

The *typical set* $\mathcal{T}_\epsilon^{(n)}$ is the set of sequences $\mathbf{x} \in \mathcal{X}^n$ with sample entropy ϵ-close to the true entropies:

$$\mathcal{T}_\epsilon^{(n)} = \left\{ \mathbf{x} \in \mathcal{X}^n : \left| -\frac{1}{n} \log p_{\mathbf{X}}(\mathbf{x}) - H(\mathsf{X}) \right| < \epsilon \right\}$$

The size of the typical set is $|\mathcal{T}_\epsilon^{(n)}|$.

A *typical sequence* is a member of the typical set, for a given value of n and ϵ.

★2

Example 5.5: Typical Set

Continue Example 5.4 of the binary random vector with $p_{\mathsf{X}}(x) = [\frac{3}{4}, \frac{1}{4}]$ and $n = 4$. Find the typical set, its size, and its probability when:

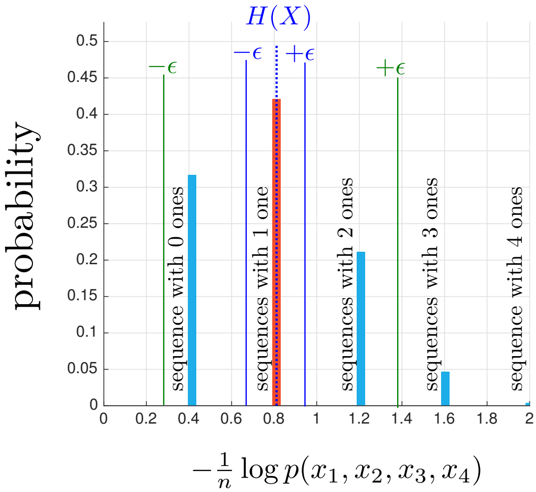
(a) $\epsilon = 0.01$ and

(b) $\epsilon = 0.45$.

★3

Example 5.5: Distribution of Sample Entropy

$\mathbf{x}$	$p_{\mathbf{X}}(\mathbf{x})$	sample entropy $-\frac{1}{n}\log p_{\mathbf{X}}(\mathbf{x})$
0 0 0 0	0.31641	0.41504
1 0 0 0	0.10547	0.81128
0 1 0 0	0.10547	0.81128
0 0 1 0	0.10547	0.81128
0 0 0 1	0.10547	0.81128
1 1 0 0	0.03516	1.20752
1 0 1 0	0.03516	1.20752
0 1 1 0	0.03516	1.20752
1 0 0 1	0.03516	1.20752
	$\vdots$	



Example: Typical Sequences Partition

The set $\mathcal{X}^n$ can be partitioned into two sets, the typical set $\mathcal{T}_\epsilon^{(n)}$ and its compliment $\overline{\mathcal{T}}_\epsilon^{(n)}$.

For example with $n = 4$, $\epsilon = 0.01$, the partition is:

$$
\begin{aligned}
\mathcal{X}^n &= \color{red}{\mathcal{T}_\epsilon^{(n)}} \cup \color{cyan}{\overline{\mathcal{T}}_\epsilon^{(n)}} \\
&= \Big\{ \color{cyan}{0000}, \color{red}{0001, 0010, 0100, 1000}, \\
&\qquad \color{cyan}{0011, 0101, 1001, 0110, 1010, 1100, 0111, 1011, 1101, 1110, 1111} \Big\}
\end{aligned}
$$

where red indicate elements of the typical set, and blue indicates elements of its complement.

5.3 Asymptotic Equipartition Property (AEP)

"Information theory is the clever application of the law of large numbers."

The Markov inequality is used to prove the Chebyshev inequality is used to prove the law of large numbers, etc.

Asymptotic Equipartition Property (AEP)

Main idea, as $n \to \infty$:

$$\begin{aligned} \Pr(\mathbf{X} \in \mathcal{T}_\epsilon^{(n)}) &\to 1 && \text{Almost all sequences are typical} \\ |\mathcal{T}_\epsilon^{(n)}| &\to 2^{nH(\mathsf{X})} && \text{Size of typical set is small*} \end{aligned}$$

*for $H(\mathsf{X}) \ll \log |\mathcal{X}|$

Proposition 5.2: AEP

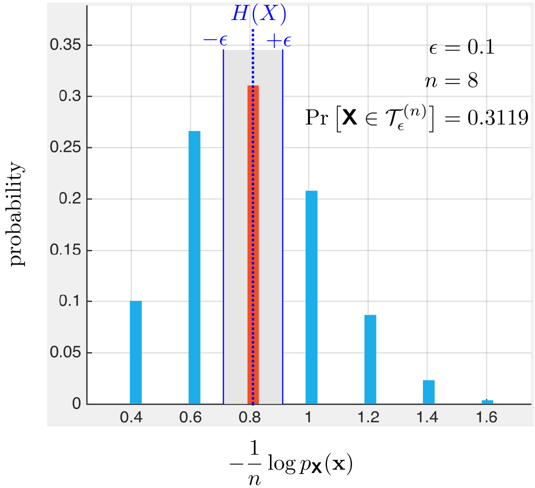
Proposition

Asymptotic Equipartition Property If $\mathsf{X}_1, \mathsf{X}_2, \ldots, \mathsf{X}_n$ are a sequence of n independent and identically distributed random variables with probability distribution $p_{\mathsf{X}}(x)$ then:

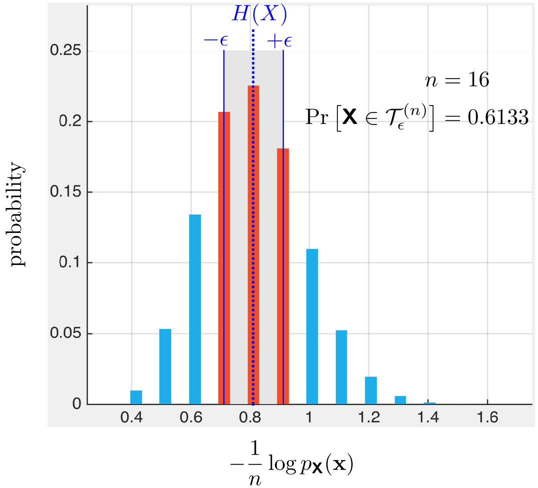
$$\lim_{n\to\infty} \Pr[\mathbf{X} \in \mathcal{T}_\epsilon^{(n)}] = 1.$$

★4

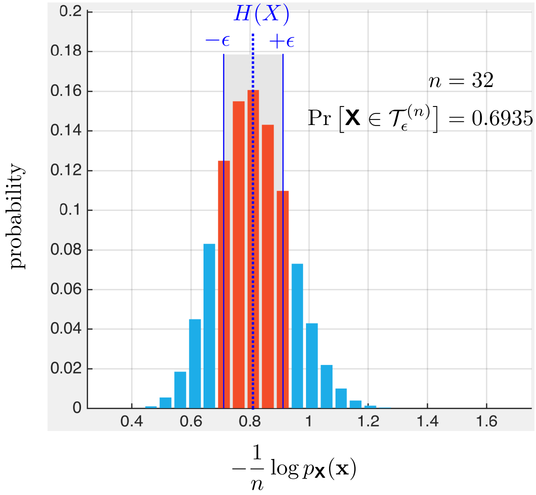
Example: Sample Entropy for $n = 8$, $p_X(x) = [\frac{3}{4}, \frac{1}{4}]$



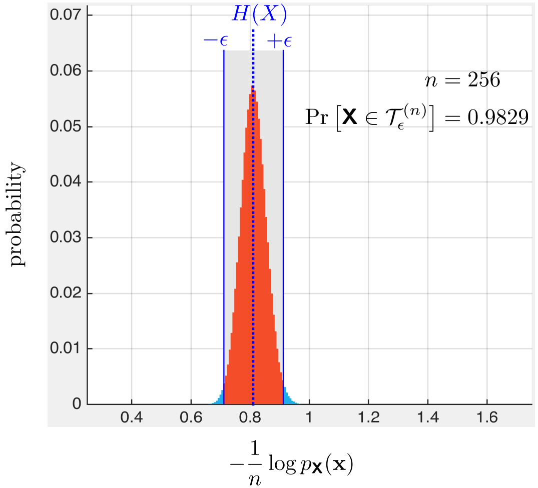
Example: Sample Entropy for $n = 16$, $p_X(x) = [\frac{3}{4}, \frac{1}{4}]$



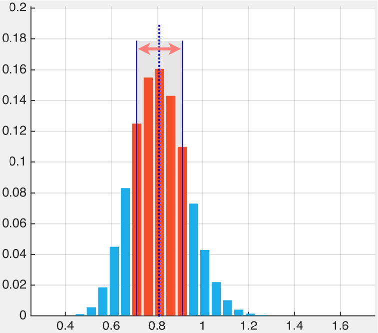
Example: Sample Entropy for $n = 32$, $p_X(x) = [\frac{3}{4}, \frac{1}{4}]$

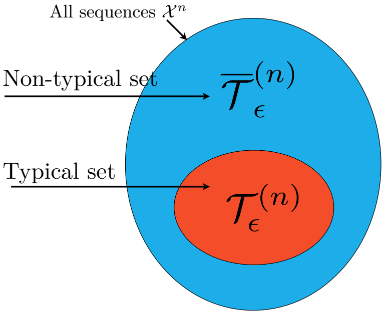


Example: Sample Entropy for $n = 256$, $p_X(x) = [\frac{3}{4}, \frac{1}{4}]$



Typical Set: Small Number of Sequences Represent Most of the Probability





Property 1 of AEP: Proposition 5.3

Proposition

Most sequences are typical For a random vector $\mathbf{X}$, most sequences are typical, that is:

$$\Pr\left[\mathbf{X} \in \mathcal{T}_\epsilon^{(n)}\right] > 1 - \epsilon,$$

for n sufficiently large. For the complementary event, $\Pr[\mathbf{X} \in \overline{\mathcal{T}}_\epsilon^{(n)}] \leq \epsilon$.

★5

Property 2 of AEP: Proposition 5.4

Proposition

The size of the typical set is bounded as:

$$(1-\epsilon)2^{n(H(\mathsf{X})-\epsilon)} \leq |\mathcal{T}_\epsilon^{(n)}| \leq 2^{n(H(\mathsf{X})+\epsilon)}.$$

The lower bound holds for n sufficiently large.

★6

5.4 Vector Source Coding

Recall the main result of this chapter.

Proposition

Vector Source Coding Theorem Let $\mathbf{X} = (\mathsf{X}_1, \ldots, \mathsf{X}_n)$ be n independent and identically distributed random variables X_i, with $\mathsf{X}_i \sim p_{\mathsf{X}}(x)$ having entropy $H(\mathsf{X})$. Let $\epsilon' > 0$. Then, there exists a vector source code with rate R that satisfies:

$$R \leq H(\mathsf{X}) + \epsilon',$$

for n sufficiently large.

We now prove this result.

5.4.1 Vector Compression Scheme

Recall that the set of all sequences $\mathcal{X}^n$ can be partitioned

$$\mathcal{X}^n = \mathcal{T}_\epsilon^{(n)} \cup \overline{\mathcal{T}}_\epsilon^{(n)}.$$

- For $\mathbf{x} \in \mathcal{T}_\epsilon^{(n)}$, assign each element of $\mathcal{T}_\epsilon^{(n)}$ an index from the set:

  $$\{1, 2, 3, \ldots, |\mathcal{T}_\epsilon^{(n)}|\}$$

  Form a codeword:

  $$C(\mathbf{x}) = \text{“0” followed by the binary index}$$

- If $\mathbf{x} \in \overline{\mathcal{T}}_\epsilon^{(n)}$, no compression is performed. The codeword is:

  $$C(\mathbf{x}) = \text{“1” followed by } \mathbf{x} \text{ itself.} \quad \bigstar 7$$

Example: Typical Sequences Partition

11 01 10 00 arbitrary binary index

$$\mathcal{X}^4 = \Big\{0000, 0001, 0010, 0100, 1000,$$

$$0011, 0101, 1001, 0110, 1010, 1100, 0111, 1011, 1101, 1110, 1111\Big\}$$

$|\mathcal{T}_\epsilon^{(n)}| = 4, \epsilon = 0.01.$

$\mathbf{x}$	$p_{\mathbf{X}}(\mathbf{x})$	in typical set?	$C(\mathbf{x})$	$\ell(\mathbf{x})$
0 0 0 0	0.3164	no	1 0 0 0 0	5
1 0 0 0	0.1055	yes	0 0 0	3
0 1 0 0	0.1055	yes	0 1 0	3
1 1 0 0	0.0352	no	1 1 1 0 0	5
0 0 1 0	0.1055	yes	0 0 1	3
1 0 1 0	0.0352	no	1 1 0 1 0	5
0 1 1 0	0.0352	no	1 0 1 1 0	5
1 1 1 0	0.0117	no	1 1 1 1 0	5
0 0 0 1	0.1055	yes	0 1 1	3
1 0 0 1	0.0352	no	1 1 0 0 1	5
0 1 0 1	0.0352	no	1 0 1 0 1	5
1 1 0 1	0.0117	no	1 1 1 0 1	5
0 0 1 1	0.0352	no	1 0 0 1 1	5
1 0 1 1	0.0117	no	1 1 0 1 1	5
0 1 1 1	0.0117	no	1 0 1 1 1	5
1 1 1 1	0.0039	no	1 1 1 1 1	5

5.4.2 Proof of Vector Source Coding Theorem

- Bound the length of typical set codewords using Property 2
- Express the length of "not typical set codewords"
- Write the expected code length R
- Upper bound the expected length R using Property 1
- Show upper bound is $R \leq H(X) + \epsilon'$
- ϵ' can be as small as desired

★8

5.4.3 "Super-Alphabet" Perspective

There is another way to prove Proposition 1.

Suppose we have an independent and identically distributed vector source $\mathsf{X}_1, \mathsf{X}_2, \ldots, \mathsf{X}_n$ over $\mathcal{X}^n$.

Think of this vector source as a single-variable random variable

$$\mathsf{Z} = \mathsf{X}_1, \ldots, \mathsf{X}_n$$

which takes values from a **super-alphabet**

$$\mathcal{Z} = \mathcal{X}^n.$$

The relationship between the expected length L and the code rate R is:

$$R = \frac{1}{n} L.$$

★9

Source Coding for Vector Sources — What You Should Have Learned

- Compression of vectors is better than single-source compression:
  - Single-source compression achieves rate $H(\mathsf{X}) + 1$ in worst case
  - Vector compression achieves $H(\mathsf{X}) + \epsilon$ in worst case
- ϵ can be made as small as needed by letting n get big. If $n \to \infty$ then $\epsilon \to 0$.
- Central idea in information theory: prove things by letting n get big

Class Info

- Next lecture: Monday, May 1. Source Coding for Markov Sources. There will be a pop quiz.
- Homework 3 and 4 on LMS. Deadline: Monday, May 1 at 18:00. Be sure to use most recent version of Lecture Notes.