INFORMATION THEORY

Master 1 - Informatique - Univ. Rennes 1 / ENS Rennes

Aline Roumy

January 2020

Outline

1. Non mathematical introduction
2. Mathematical introduction: definitions
3. Typical vectors and the Asymptotic Equipartition Property (AEP)
4. Lossless Source Coding
5. Variable length Source coding - Zero error Compression

About me

Aline Roumy
Researcher at Inria, Rennes
Expertise: compression for video streaming
image/signal processing, information theory, machine learning

Web: http://people.rennes.inria.fr/Aline.Roumy/
email: aline.roumy@inria.fr

Course schedule (tentative)

Information theory (IT):

a self-sufficient course with a lot of connections to probability

- Lecture 1: introduction, reminder on probability
- Lecture 2-3: Data compression (theoretical limits)
- Lecture 4: Construction of codes that can compress data
- Lecture 5: Beyond classical information theory (universality...)

Course organization:

- slides (file available online)
- summary (file available online+hardcopy)
- proofs (see blackboard): take notes!

On my webpage:
http://people.rennes.inria.fr/Aline.Roumy/roumy_teaching.html

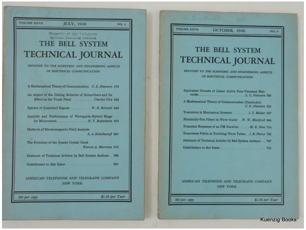
Course grading and documents

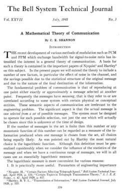
- Homework:
  - exercises (in class and at home)
  - correction in front of the class give bonus points.
- Middle Exam:
  - (in group) written exam,
  - home.
- Final Exam:
  - (individual) written exam
  - questions de cours, et exercices (in French)
  - 2h
- All documents on my webpage:
  http://people.rennes.inria.fr/Aline.Roumy/roumy_teaching.html

Course material

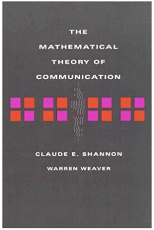
C.E. Shannon, "A mathematical theory of communication", *Bell Sys. Tech. Journal*, 27: 379–423, 623–656, 1948.

seminal paper



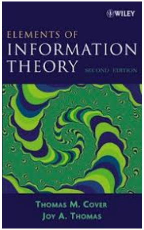

The Bell System Technical Journal

A Mathematical Theory of Communication



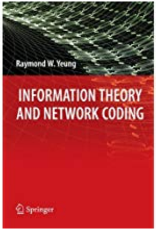
Course material

T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications. Wiley, New York, 2006.
THE reference



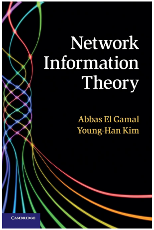
Course material

R. Yeung. *Information Theory and Network Coding*.
Springer 2008.
network coding



Course material

A. E. Gamal and Y-H. Kim. *Network Information Theory*.
Cambridge University Press 2011.
network information theory



Slides:
A. E. Gamal and Y-H. Kim.
Lecture Notes on Network Information Theory. arXiv:1001.3404v5, 2011. **web**

Outline

1. Non mathematical introduction
2. Mathematical introduction: definitions
3. Typical vectors and the Asymptotic Equipartition Property (AEP)
4. Lossless Source Coding
5. Variable length Source coding - Zero error Compression

Lecture 1

Non mathematical introduction

What does “communicating” means?

What it is about? A bit of history...

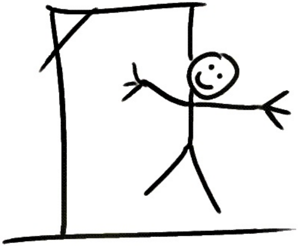
- Information theory (IT) =
  "The fundamental problem of **communication** is that of reproducing at one point, either exactly or approximately, a message selected at another point."
- IT established by Claude E. Shannon (1916-2001) in 1948.
  - Seminal paper: "A Mathematical Theory of Communication" in the Bell System Technical Journal, 1948.
  - revolutionary and groundbreaking paper

Teaser 1: compression

Hangman game

- **Objective:** play... and explain your strategy

_ _ _ _ _

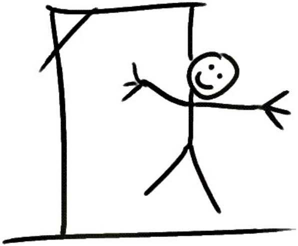


Teaser 1: compression

Hangman game

- **Objective:** play... and explain your strategy

_ _ _ _ _



- **2 winning ideas**
  - Letter frequency — **probability**
  - Correlation between successive letters — **dependence**

Teaser 1: compression

Analogy Hangman game-compression

- word
- Answer to a question (yes/no)

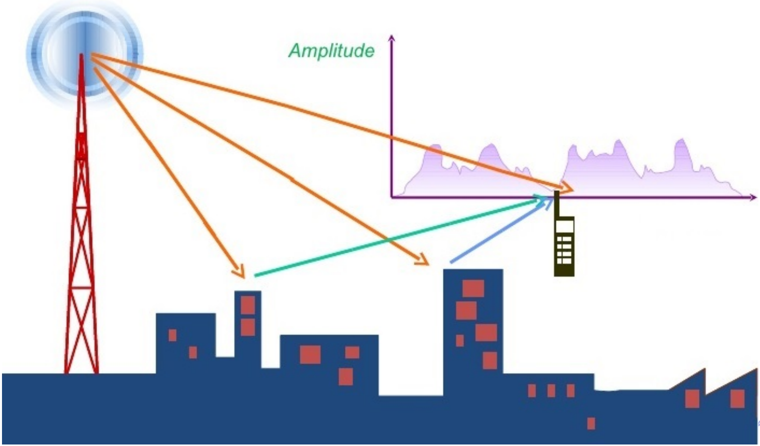
  removes uncertainty in word
- **Goal:** propose a minimum number of letter

- data (image)
- 1 bit of the bistream that represents the data

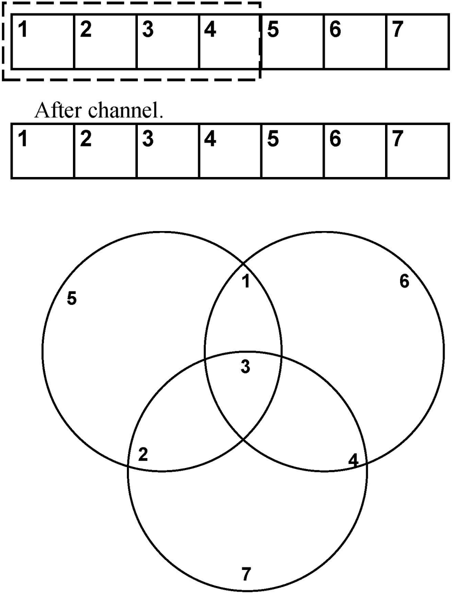
  removes uncertainty in data
- **Goal:** propose a minimum number of bits

Teaser 2: communication over a noisy channel

- Context:
  storing/communicating data on a channel **with errors**
  - scratches on a DVD
  - lost data packets: webpage sent over the internet.
  - lost or modified received signals: wireless links



Teaser 2: communication over a noisy channel



1. choose binary vector (x_1, x_2, x_3, x_4)
2. compute x_5, x_6, x_7 s.t. XOR in each circle is 0
3. add 1 or 2 errors
4. correct errors s.t. rule 2 is satisfied

Quiz 1:

Assume you know how many errors have been introduced.
Can one correct 1 error?
Can one correct 2 errors?

Teaser 2: communication over a noisy channel

Take Home Message (THM):

- To get zero error at the receiver, one can send a **FINITE** number of additional of bits.
- For a finite number of additional of bits, there is a limit on the number of errors that can be corrected.

Summary

- one can **compress** data by using two ideas:
  - Use **non-uniformity** of the probabilities

    this is the source coding theorem (first part of the course)

    very surprising...
  - Use **dependence** between the data

    in middle exam

Summary

- one can **compress** data by using two ideas:
  - Use **non-uniformity** of the probabilities

    this is the source coding theorem (first part of the course)

    very surprising...
  - Use **dependence** between the data

    in middle exam

- one can **send** data over a **noisy** channel and recover the data without any error

  provided the data is **encoded** (send additional data)

  this is the channel coding theorem (second part of the course)

Communicate what?

Definition

Source of information: something that produces **messages**.

Definition

Message: a stream of **symbols** taking their values in an **alphabet**.

Example

Source: camera
Message: picture
Symbol: pixel value: 3 coef. (RGB)
Alphabet=$\{0, \ldots, 255\}^3$

Example

Source: writer
Message: a text
Symbol: letter
Alphabet=$\{a, \ldots, z, !, ., ?, ...\}$

How to model the communication?

- Model for the source:

  communication
  a source of information
  a message of the source
  a symbol of the source
  alphabet of the source

- Model for the communication chain:

How to model the communication?

- Model for the source:

communication		**mathematical model**
a source of information	$\rightarrow$	a random process
a message of the source	$\rightarrow$	a realization of a random **vector**
a symbol of the source	$\rightarrow$	a realization of a random variable
alphabet of the source	$\rightarrow$	alphabet of the random variable

- Model for the communication chain:

Point-to-point Information theory

Shannon proposed and proved three fundamental theorems for point-to-point communication (1 sender / 1 receiver):

1. **Lossless source** coding theorem: For a given source, what is the minimum rate at which the source can be **compressed losslessly**?
   rate = nb bits / source symbol
2. **Lossy source** coding theorem: For a given source and a given distortion D, what is the minimum rate at which the source can be **compressed within distortion** D.
   rate = nb bits / source symbol
3. **Channel coding** theorem: What is the maximum rate at which data can be **transmitted** reliably?
   rate = nb bits / sent symbol over the channel

Application of Information Theory

Information theory is everywhere...

1. **Lossless source** coding theorem:
2. **Lossy source** coding theorem:
3. **Channel coding** theorem:

Quiz 2: On which theorem (1/2/3) rely these applications?

(1) zip compression
(2) jpeg and mpeg compression
(3) sending a jpeg file onto internet
(4) the 15 digit social security number
(5) movie stored on a DVD

Reminder (1)

Definition (Convergence in probability)

Let $(X_n)_{n\geq 1}$ be a sequence of r.v. and X a r.v. both defined over $\mathbb{R}$. $(X_n)_{n\geq 1}$ converges in probability to the r.v. X if

$$\forall \epsilon > 0, \lim_{n\to+\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0.$$

Notation:

$$X_n \xrightarrow{p} X$$

Quiz 3: Which of the following statements are true?

(1) X_n and X are random

(2) X_n is random and X is deterministic (constant)

Reminder (2)

Theorem (Weak Law of Large Numbers (WLLN))

Let $(X_n)_{n\geq 1}$ be a sequence of r.v. over $\mathbb{R}$.
If $(X_n)_{n\geq 1}$ is i.i.d., $\mathcal{L}^2$ (i.e. $\mathbb{E}[X_n^2] < \infty$) then

$$\frac{X_1 + \ldots + X_n}{n} \xrightarrow{p} \mathbb{E}[X_1]$$

Quiz 4: Which of the following statements are true?

(1) for any nonzero margin, with a sufficiently large sample there will be a very high probability that the average of the observations will be close to the expected value; that is, within the margin.
(2) LHS and RHS are random
(3) averaging kills randomness
(4) the statistical mean ((a.k.a. true mean) converges to the empirical mean (a.k.a. sample mean)

Outline

1. Non mathematical introduction
2. Mathematical introduction: definitions
3. Typical vectors and the Asymptotic Equipartition Property (AEP)
4. Lossless Source Coding
5. Variable length Source coding - Zero error Compression

Lecture 2

Mathematical introduction

Definitions: Entropy and Mutual Information

Some Notation

Specific to information theory are denoted in red

- Upper case letters $X, Y, \ldots$ refer to random process or random variable
- Calligraphic letters $\mathcal{X}, \mathcal{Y}, \ldots$ refer to alphabets
- $|\mathcal{A}|$ is the cardinality of the set $\mathcal{A}$
- $X^n = (X_1, X_2, \ldots, X_n)$ is an n-sequence of random variables or a random vector

  $X_i^j = (X_i, X_{i+1}, \ldots, X_j)$
- Lower case $x, y, \ldots$ and $x^n, y^n, \ldots$ mean scalars/vectors realization
- $X \sim p(x)$ means that the r.v. X has probability mass function (pmf) $\mathbb{P}(X = x) = p(x)$
- $X^n \sim p(x^n)$ means that the discrete random vector X^n has joint pmf $p(x^n)$
- $p(y^n | x^n)$ is the conditional pmf of Y^n given $X^n = x^n$.

Lecture 1: Entropy (1)

Definition (Entropy)

the **entropy** of a **discrete** random variable $X \sim p(x)$:

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$$

$H(X)$ in **bits/source sample** is the **average length of the shortest description** of the r.v. X. *(Shown later)*

Notation: $\log := \log_2$

Convention: $0 \log 0 := 0$

Properties

E1 $H(X)$ only depends on the pmf $p(x)$ and not x.

E2 $H(X) = -\mathbb{E}_X \log p(X)$

Entropy (2)

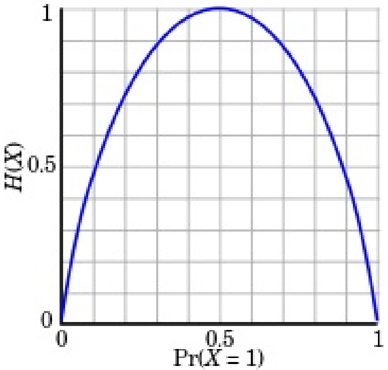
E3 $H(X) \geq 0$ with equality iff X is constant.

E4 $H(X) \leq \log |\mathcal{X}|$. The uniform distribution maximizes entropy.

Example

Binary entropy function: Let $0 \leq p \leq 1$

$$h_b(p) = -p \log p - (1-p) \log (1-p)$$



$H(X)$ for a binary rv.

$H(X)$ measures the amount of uncertainty on the rv X.

Entropy (3)

E4 **(con't)** Alternative proof with the positivity of the **Kullback-Leibler (KL) divergence**.

Definition (Kullback-Leibler (KL) divergence)

Let $p(x)$ and $q(x)$ be 2 pmfs defined on the same set $\mathcal{X}$.
The **KL divergence** between p and q is:

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

Convention: $c \log c/0 = \infty$ for $c > 0$.

Quiz 5: Which of the following statements are true?

(1) $D(p||q) = D(q||p)$.
(2) If $\text{Support}(q) \subset \text{Support}(p)$ then $D(p||q) = \infty$.

Entropy (4)

KL1 **Positivity of KL** [Cover Th. 2.6.3]: $D(p||q) \geq 0$
with equality iff $\forall x, p(x) = q(x)$.

This is a consequence of Jensen's inequality [Cover Th. 2.6.2]:
If f is a convex function and Y is a random variable with numerical values, then

$$\mathbb{E}[f(Y)] \geq f(\mathbb{E}[Y])$$

with equality when $f(.)$ is not strictly convex, or when $f(.)$ is strictly convex and Y follows a degenerate distribution (i.e. is a constant).

KL2 Let $X \sim p(x)$ and $q(x) = \frac{1}{|\mathcal{X}|}$, then $D(p||q) = -H(X) + \log|\mathcal{X}|$

Reminder (independence)

Definition (independence)

The random variables X and Y are independent, denoted by $X \perp\!\!\!\perp Y$, if

$$\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, \quad p(x, y) = p(x)p(y).$$

Definition (Mutual independence – mutuellement indépendant)

For $n \geq 3$, the random variables $X_1, X_2, \ldots, X_n$ are mutually independent if

$$\forall (x_1, \ldots, x_n) \in \mathcal{X}_1 \times \ldots \times \mathcal{X}_n, \quad p(x_1, \ldots, x_n) = p(x_1)p(x_2) \ldots p(x_n).$$

Definition (Pairwise independence – indépendance 2 à 2)

For $n \geq 3$, the random variables X_i, X_j are pairwise independent if $\forall (i, j)$ s.t. $1 \leq i < j \leq n$, X_i and X_j are independent.

Quiz 6

Quiz 6: Which of the following statements are/is true?

(1) mutual independence implies pairwise independence.

(2) pairwise independence implies mutual independence

Reminder (conditional independence)

Definition (conditional independence)

Let X, Y, Z be r.v.
X is independent of Z given Y, denoted by $X \perp\!\!\!\perp Z | Y$, if

$$\forall (x, y, z) \quad p(x, z|y) = \begin{cases} p(x|y)p(z|y) & \text{if } p(y) > 0 \\ 0 & \text{otherwise} \end{cases}$$

or equivalently

$$\forall (x, y, z) \quad p(x, y, z) = \begin{cases} \frac{p(x,y)p(y,z)}{p(y)} = p(x, y)p(z|y) & \text{if } p(y) > 0 \\ 0 & \text{otherwise} \end{cases}$$

or equivalently

$$\forall (x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}, \quad p(x, y, z)p(y) = p(x, y)p(y, z),$$

Definition (Markov chain)

Let $X_1, X_2, ..., X_n$, $n \geq 3$ be r.v.
$X_1 \to X_2 \to ... \to X_n$ forms a Markov chain if $\forall(x_1, ..., x_n)$

$$p(x_1, x_2, ..., x_n) = \begin{cases} p(x_1, x_2)p(x_3|x_2)...p(x_n|x_{n-1}) & \text{if } p(x_2),...,p(x_{n-1})>0 \\ 0 & \text{otherwise} \end{cases}$$

or equivalently $\forall(x_1, ..., x_n)$

$$p(x_1, x_2, ..., x_n)p(x_2)p(x_3)...p(x_{n-1}) = p(x_1, x_2)p(x_2, x_3)...p(x_{n-1}, x_n)$$

Quiz 7: Which of the following statements are true?

(1) $X \perp\!\!\!\perp Z|Y$ is equivalent to $X \to Z \to Y$
(2) $X \perp\!\!\!\perp Z|Y$ is equivalent to $X \to Y \to Z$
(3) $X_1 \to X_2 \to ... \to X_n \Rightarrow X_n \to ... \to X_2 \to X_1$

Joint and conditional entropy

Definition (Conditional entropy)

For discrete random variables $(X, Y) \sim p(x, y)$,
the **Conditional entropy** **for a given y** is:

$$H(X \mid Y = y) \quad = \quad -\sum_{x \in \mathcal{X}} p(x|y) \log p(x|y)$$

the **Conditional entropy** is:

$$H(X \mid Y) = \sum_{y \in \mathcal{Y}} p(y) H(X \mid Y = y) = -\mathbb{E}_{XY} \log p(X \mid Y)$$

$$= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x|y) = -\sum_{y \in \mathcal{Y}} p(y) \sum_{x \in \mathcal{X}} p(x|y) \log p(x|y)$$

$H(X|Y)$ in **bits/source sample** is the **average length of the shortest description** of the r.v. X when Y is known.

Joint entropy

Definition (Joint entropy)

For discrete random variables $(X, Y) \sim p(x, y)$, the **Joint entropy** is:

$$H(X, Y) = -\mathbb{E}_{XY} \log p(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y)$$

$H(X, Y)$ in **bits/source sample** is the **average length of the shortest description** of ???.

Properties

JCE1 **trick** $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$

JCE2 $H(X,Y) \leq H(X) + H(Y)$ with equality iff X and Y are independent (denoted $X \perp\!\!\!\perp Y$).

JCE3 **Conditioning reduces entropy**
$H(X|Y) \leq H(X)$ with equality iff $X \perp\!\!\!\perp Y$

JCE4 **Chain rule** for entropy (formule des conditionnements successifs)
Let X^n be a discrete random vector

$$\begin{aligned} H(X^n) &= H(X_1) + H(X_2|X_1) + \ldots + H(X_n|X_{n-1},\ldots,X_1) \\ &= \sum_{i=1}^{n} H(X_i|X_{i-1},\ldots,X_1) \\ &= \sum_{i=1}^{n} H\left(X_i|X^{i-1}\right) \leq \sum_{i=1}^{n} H(X_i) \end{aligned}$$

with notation $H(X_1|X^0) = H(X_1)$.

JCE5 $H(X|Y) \geq 0$ with equality iff $X = f(Y)$ a.s.

JCE6 $H(X|X) = 0$ and $H(X, X) = H(X)$

JCE7 **Data processing inequality.** Let X be a discrete random variable and $g(X)$ be a function of X, then

$$H(g(X)) \leq H(X)$$

with equality iff $g(x)$ is injective on the support of $p(x)$.

JCE8 **Fano's inequality:** link between entropy and error prob. Let $(X, Y) \sim p(x, y)$ and $P_e = \mathbb{P}\{X \neq Y\}$, then

$$H(X|Y) \leq h_b(P_e) + P_e \log(|\mathcal{X}| - 1) \leq 1 + P_e \log(|\mathcal{X}| - 1)$$

JCE9 $H(X|Z) \geq H(X|Y, Z)$ with equality iff X and Y are independent given Z (denoted $X \perp\!\!\!\perp Y|Z$).

JCE10 $H(X, Y|Z) \leq H(X|Z) + H(Y|Z)$ with equality iff $X \perp\!\!\!\perp Y|Z$.

Venn diagram

	is represented by	
X (a r.v.)	$\rightarrow$	set (set of realizations)
$H(X)$	$\rightarrow$	area of the set
$H(X, Y)$	$\rightarrow$	area of the union of sets

Exercise

1. Draw a Venn Diagram for 2 r.v. X and Y.
   Show $H(X), H(Y), H(X, Y)$ and $H(Y|X)$.
2. Show the case $X \perp\!\!\!\perp Y$
3. Draw a Venn Diagram for 3 r.v. X, Y and Z and show the decomposition $H(X, Y, Z) = H(X) + H(Y|X) + H(Z|X, Y)$.
4. Show the case $X \perp\!\!\!\perp Y|Z$

Mutual Information

Definition (Mutual Information)

For discrete random variables $(X, Y) \sim p(x, y)$, the **Mutual Information** is:

$$\begin{aligned} I(X;Y) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\ &= H(X) - H(X|Y) = H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned}$$

Exercise Show $I(X;Y)$ on the Venn Diagram representing X and Y.

Mutual Information: properties

MI1 $I(X;Y)$ is a function of $p(x,y)$

MI2 $I(X;Y)$ is symmetric: $I(X;Y) = I(Y;X)$

MI3 $I(X;X) = H(X)$

MI4 $I(X;Y) = D(p(x,y) || p(x)p(y))$

MI5 $I(X;Y) \geq 0$
with equality iff $X \perp\!\!\!\perp Y$

MI6 $I(X;Y) \leq \min(H(X), H(Y))$
with equality iff $X = f(Y)$ **a.s. or** $Y = f(X)$ **a.s.**

Conditional Mutual Information

Definition (Conditional Mutual Information)

For discrete random variables $(X, Y, Z) \sim p(x, y, z)$, the **Conditional Mutual Information** is:

$$
\begin{aligned}
I(X; Y|Z) &= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \sum_{z\in\mathcal{Z}} p(x, y, z) \log \frac{p(x, y|z)}{p(x|z)p(y|z)} \\
&= H(X|Z) - H(X|Y, Z) \\
&= H(Y|Z) - H(Y|X, Z)
\end{aligned}
$$

Exercise Show $I(X; Y|Z)$ and $I(X; Z)$ on the Venn Diagram representing X, Y, Z.

CMI1 $I(X; Y|Z) \geq 0$ with equality iff $X \perp\!\!\!\perp Y|Z$

Exercise Compare $I(X; Y, Z)$ with $I(X; Y|Z) + I(X; Z)$ on the Venn Diagram representing X, Y, Z.

CMI2 Chain rule

$$I\left(X^n;Y\right)=\sum_{i=1}^{n} I\left(X_i;Y\left|X^{i-1}\right.\right)$$

CMI3 If $X\to Y\to Z$ form a Markov chain, then $I(X;Z|Y)=0$

CMI4 Corollary: If $X\to Y\to Z$, then $I(X;Y)\geq I(X;Y|Z)$

CMI5 Corollary: **Data processing inequality:**
If $X\to Y\to Z$ form a Markov chain, then $I(X;Y)\geq I(X;Z)$

Exercise Draw the Venn Diagram of the Markov chain $X\to Y\to Z$

CMI6 There is **no order relation** between $I(X;Y)$ and $I(X;Y|Z)$
Faux amis: Recall $H(X|Z)\leq H(X)$

Hint: show an example s.t. $I(X;Y)>I(X;Y|Z)$ and an example s.t. $I(X;Y)<I(X;Y|Z)$

Exercise Show the area that represents $I(X;Y)-I(X;Y|Z)$ on the Venn Diagram...

Outline

1. Non mathematical introduction
2. Mathematical introduction: definitions
3. Typical vectors and the Asymptotic Equipartition Property (AEP)
4. Lossless Source Coding
5. Variable length Source coding - Zero error Compression

Lecture 3

Typical vectors and
Asymptotic **Equipartition** Property (AEP)

Re-reminder

Definition (Convergence in probability)

Let $(X_n)_{n\geq 1}$ be a sequence of r.v. and X a r.v. both defined over $\mathbb{R}^d$.
$(X_n)_{n\geq 1}$ converges in probability to the r.v. X if

$$\forall \epsilon > 0, \lim_{n\to+\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0.$$

Notation:

$$X_n \xrightarrow{p} X$$

Theorem (Weak Law of Large Numbers (WLLN))

Let $(X_n)_{n\geq 1}$ be a vector of r.v. over $\mathbb{R}$.
If $(X_n)_{n\geq 1}$ is i.i.d., $\mathcal{L}^2$ (i.e. $\mathbb{E}[X_n^2] < \infty$) then

$$\frac{X_1 + \ldots + X_n}{n} \xrightarrow{p} \mathbb{E}[X_1]$$

Theorem (Asymptotic Equipartition Property (AEP))

Let $X_1, X_2, \ldots$ *be i.i.d.* $\sim p(x)$ **finite** *random process (source), let us denote* $p(x^n) = \prod_{i=1}^{n} p(x_i)$, *then*

$$-\frac{1}{n} \log p(X^n) \to H(X) \qquad \textit{in probability}$$

Definition (Typical set)

Let $\epsilon > 0, n > 0$ and $X \sim p(x)$, the set $A_\epsilon^{(n)}(X)$ of ϵ-**typical** vectors x^n, where $p\left(x^n\right) = \prod_{i=1}^n p\left(x_i\right)$ is defined as

$$A_\epsilon^{(n)}(X) = \left\{ x^n : \left| -\frac{1}{n} \log p\left(x^n\right) - H(X) \right| \leq \epsilon \right\}$$

Properties

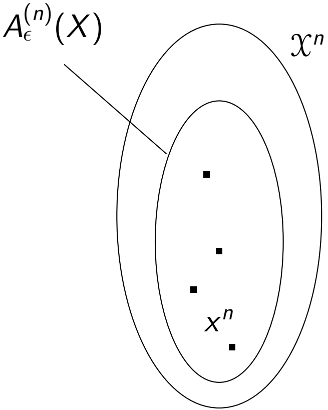
AEP1 $\forall(\epsilon, n)$, all these statements are equivalent:

$$\begin{aligned} x^n \in A_\epsilon^{(n)} \quad &\Leftrightarrow \quad 2^{-n(H(X)+\epsilon)} \leq p\left(x^n\right) \leq 2^{-n(H(X)-\epsilon)} \\ &\Leftrightarrow \quad p\left(x^n\right) \doteq 2^{-n(H(X)\pm\epsilon)} \end{aligned}$$

Notation: $a_n \doteq 2^{n(b\pm\epsilon)} \Leftrightarrow \left| \frac{1}{n} \log a_n - b \right| \leq \epsilon$ for n sufficiently large.

"uniform distribution on the typical set"

Interpretation of typicality



Example of typical vectors

$\mathbb{P}[X = x] = p(x)$
$x^n = (x_1, \ldots x_i, \ldots x_n)$
$n_x = |\{i : x_i = x\}|$

Let x^n satisfies $\frac{n_x}{n} = p(x)$ then

$$p(x^n) = \prod_i p(x_i) = \prod_{x \in \mathcal{X}} p(x)^{n_x}$$
$$= 2^{\sum_x np(x) \log p(x)} = 2^{-nH(X)}$$

x^n represents well the distribution
So, x^n is ϵ-typical, $\forall \epsilon$.

Quiz

- Let $X \sim \mathcal{B}(0.2)$, $\epsilon = 0.1$ and n=10. Which of the following x^n vector is ϵ-typical? $a = (0100000100)$ $b = (1100000000)$ $c = (1111111111)$
- Let $X \sim \mathcal{B}(0.5)$, $\epsilon = 0.1$ and n=10. Which x^n vectors are ϵ-typical?

Properties

AEP2 $\forall \epsilon > 0, \lim_{n \to +\infty} \mathbb{P}\left(\left\{X^n \in A_\epsilon^{(n)}(X)\right\}\right) = 1$

"for a given ϵ, asymptotically a.s. typical"

Theorem (CT Th. 3.1.2)

Given $\epsilon > 0$. Assume that $\forall n, X^n \sim \prod_{i=1}^n p(x_i)$.

Then, for n **sufficiently large**, we have

1. $\mathbb{P}\left(A_\epsilon^{(n)}(X)\right) = \mathbb{P}\left(\left\{X^n \in A_\epsilon^{(n)}(X)\right\}\right) > 1 - \epsilon$
2. $\left|A_\epsilon^{(n)}(X)\right| \leq 2^{n(H(X)+\epsilon)}$
3. $\left|A_\epsilon^{(n)}(X)\right| > (1-\epsilon) 2^{n(H(X)-\epsilon)}$

2 and 3 can be summarized in $\left|A_\epsilon^{(n)}\right| \doteq 2^{n(H(X) \pm 2\epsilon)}$.

Outline

1. Non mathematical introduction
2. Mathematical introduction: definitions
3. Typical vectors and the Asymptotic Equipartition Property (AEP)
4. Lossless Source Coding
5. Variable length Source coding - Zero error Compression

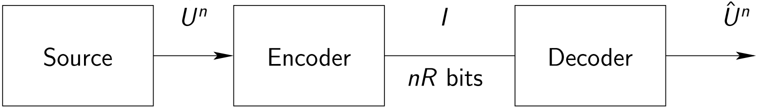
Lecture 4

Lossless Source Coding

↓

data compression

Compression system model:



We assume a **finite** alphabet i.i.d. source $U_1, U_2, \ldots \sim p(u)$.

Definition (Fixed-length Source code (FLC))

Let $R \in \mathbb{R}^+, n \in \mathbb{N}^*$. A $\left(2^{nR}, n\right)$ fixed-length source code consists of:

1. An **encoding function** that assigns to each $u^n \in \mathcal{U}^n$ an index $i \in \{1, 2, \ldots, 2^{nR}\}$, i.e., a codeword of length nR bits:

$$\begin{aligned} \mathcal{U}^n &\rightarrow \mathcal{I} = \{1, 2, ..., 2^{nR}\} \\ u^n &\mapsto i(u^n) \end{aligned}$$

2. A **decoding function** that assigns an estimate $\hat{u}^n(i)$ to each received index i

$$\begin{aligned} \mathcal{I} &\rightarrow \mathcal{U}^n \\ i &\mapsto \hat{u}^n(i) \end{aligned}$$

Definition (Probability of decoding error)

Let $n \in \mathbb{N}^*$. The **probability of decoding error** is $P_e^{(n)} = \mathbb{P}\{\hat{U}^n \neq U^n\}$

R is called the **compression rate**: number of bits per source sample.

Definition (Achievable rate)

Let $R \in \mathbb{R}^+$. A rate R is **achievable** if there exists a sequence of $(2^{nR}, n)$ codes with $P_e^{(n)} \to 0$ as $n \to \infty$

Source Coding Theorem

The source coding problem is to find the infimum of all achievable rates.

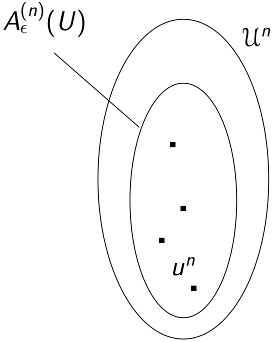
Theorem (Source coding theorem (Shannon'48))

Let $U \sim p(u)$ be a **finite** alphabet i.i.d. source. Let $R \in \mathbb{R}^+$.
[Achievability]. If $R > H(U)$,
then there exists a sequence of $\left(2^{nR}, n\right)$ codes s.t. $P_e^{(n)} \to 0$.

[Converse]. For any sequence of $\left(2^{nR}, n\right)$ codes s.t. $P_e^{(n)} \to 0$,
$R \geq H(U)$

Classical (and equivalent) statement of **[Converse]**:
If there exists a sequence of $\left(2^{nR}, n\right)$ codes s.t. $P_e^{(n)} \to 0$,
then $R \geq H(U)$

Proof of achievability [CT Th. 3.2.1]



Let $U \sim p(u)$ a finite alphabet i.i.d. process.
Let $R \in \mathbb{R}$, $\epsilon > 0$.

- Assume that $R > H(U) + \epsilon$.
  Then $\left|A_\epsilon^{(n)}\right| \leq 2^{n(H(U)+\epsilon)} < 2^{nR}$.
  Assume that nR is an integer.
- Encoding: Assign a distinct index $i\left(u^n\right)$ to each $u^n \in A_\epsilon^{(n)}$
  Assign the same index (not assigned to any typical vector) to all $u^n \notin A_\epsilon^{(n)}$
- The probability of error
  $P_e^{(n)} = 1 - \mathbb{P}\left(A_\epsilon^{(n)}\right) \to 0$ as $n \to \infty$

Proof of converse [Yeung Sec. 5.2, ElGamal Page 3-34]

- Given a sequence of $\left(2^{nR}, n\right)$ codes with $P_e^{(n)} \to 0$, let I be the random variable corresponding to the index of the $\left(2^{nR}, n\right)$ encoder.
  By Fano's inequality

  $$H\left(U^n \,|I\right) \leq H\left(U^n \,\middle|\, \hat{U}^n\right) \leq nP_e^{(n)} \log |\mathcal{U}| + 1 \triangleq n\epsilon_n$$

  where $\epsilon_n \to 0$ as $n \to \infty$, since $|\mathcal{U}|$ is finite.
- Now consider

  $$\begin{aligned} nR &\geq H(I) \\ &= I\left(U^n; I\right) \\ &= nH(U) - H\left(U^n|I\right) \geq nH(U) - n\epsilon_n \end{aligned}$$

  Thus as $n \to \infty, R \geq H(U)$
- The above source coding theorem also holds for any discrete stationary and ergodic source

Outline

Lecture 5

Variable length Source coding

Zero error Data Compression

A code

Definition (Variable length Source code (VLC))

Let X be a r.v. with finite alphabet $\mathcal{X}$. A **variable-length source code** C for a random variable X is a mapping

$$C : \mathcal{X} \rightarrow \mathcal{A}^*$$

where $\mathcal{X}$ is a set of M **symbols**,
$\mathcal{A}$ is a set of D **letters**, and
$\mathcal{A}^*$ the set of finite length sequences (or strings) of letters from $\mathcal{A}$.
$C(x)$ denotes the **codeword** corresponding to the symbol x.

In the following, we will say Source code for VLC.
Examples 1, 2

The length of a code

Let $L : \mathcal{A}^* \to \mathbb{N}$ denote the **length mapping** of a codeword (sequence of letters).

$L(C(x))$ is the number of letters of $C(x)$, and
$L(C(x)) \log |\mathcal{A}|$ the number of bits.

Definition

The **expected length** $L(C)$ of a source code C for a random variable X with pmf $p(x)$ is given by:

$$L(C) = \mathbb{E}[L(C(X))] = \sum_{x \in \mathcal{X}} L(C(x))p(x)$$

Goal Find a source code C for X with **smallest $L(C)$**.

Encoding a sequence of source symbols

Definition

A **source message**= a sequence of symbols
A **coded sequence**= a sequence of codewords

Definition

The **extension** of a code C is the mapping from finite length sequences of $\mathcal{X}$ (of any length) to finite length strings of $\mathcal{A}$, defined by:

$$\begin{array}{rccll} C: & \mathcal{X}^* & \rightarrow & \mathcal{A}^* & \\ & (x_1, \ldots, x_n) & \mapsto & C(x_1, \ldots, x_n) & = C(x_1)C(x_2)\ldots C(x_n) \end{array}$$

where $C(x_1)C(x_2)\ldots C(x_n)$ indicates the concatenation of the corresponding codewords.

Characteristics of good codes

Definition

A (source) code C is said to be **non-singular** iff C is injective:

$$\forall (x_i, x_j) \in \mathcal{X}^2, x_i \neq x_j \Rightarrow C(x_i) \neq C(x_j)$$

Definition

A code is called **uniquely decodable** iff its extension is non-singular.

Definition

A code is called a **prefix code** (or an **instantaneous code**) if no codeword is a prefix of any other codeword.

prefix code	$\Rightarrow$	**uniquely decodable**
uniquely decodable	$\not\Rightarrow$	**prefix** code

Examples

Kraft inequality

Theorem (prefix code $\Leftrightarrow$ KI [CT Th 5.2.1])

Let C be an **prefix code** *for the source X with $|\mathcal{X}| = M$ over an alphabet $\mathcal{A} = \{a_1, ..., a_D\}$ of size D. Let $l_1, l_2, ..., l_M$ the lengths of the codewords associated to the realizations of X. These codeword lengths must satisfy the* **Kraft inequality**

$$\sum_{i=1}^{M} D^{-l_i} \leq 1 \qquad \text{(KI)}$$

Conversely, *let $l_1, l_2, ..., l_M$ be M lengths that satisfy this inequality (KI), there exists an* **prefix code** *with M symbols, constructed with D letters, and with these word lengths.*

- from the lengths, one can always construct a prefix code
- finding prefix code is equivalent to finding the codeword lengths

uniquely decodable

Theorem (uniquely decodable code $\Leftrightarrow$ KI [CT Th 5.5.1])

The codeword lengths of any **uniquely decodable** *code must satisfy the Kraft inequality.*
Conversely, *given a set of codeword lengths that satisfy this inequality, it is possible to construct a* **uniquely decodable** *code with these codeword lengths.*

uniquely decodable

Theorem (uniquely decodable code ⇔ KI [CT Th 5.5.1])

The codeword lengths of any **uniquely decodable** *code must satisfy the Kraft inequality.*
Conversely, *given a set of codeword lengths that satisfy this inequality, it is possible to construct a* **uniquely decodable** *code with these codeword lengths.*

Good news!!

$$\begin{array}{rcl} \text{prefix code} & \Leftrightarrow & \text{KI} \\ \text{uniquely decodable code (UDC)} & \Leftrightarrow & \text{KI} \end{array}$$

$\Rightarrow$ same set of achievable codeword lengths for UDC and prefix

$\Rightarrow$ restrict the search of good codes to the set of prefix codes.

Optimal source codes

Let X be a r.v. taking M values in $\mathcal{X} = \{\alpha_1, \alpha_2, ..., \alpha_M\}$, with probabilities $p_1, p_2, ..., p_M$.

Each symbol α_i is associated with a codeword W_i i.e. a sequence of l_i letters, where each letter takes value in an alphabet of size D.

Goal

Find a uniquely decodable code with minimum expected length.

$$\Longleftrightarrow$$

Find a prefix code with minimum expected length.

$$\Longleftrightarrow$$

Find a set of lengths satisfying KI with minimum expected length.

$$\{l_1^*, l_2^*, ..., l_M^*\} = \arg \min_{\{l_1, l_2, ..., l_M\}} \sum_{i=1}^{M} p_i l_i \qquad \text{(Pb1)}$$

$$\text{s.t. } \forall i, l_i \geq 0 \text{ and } \sum_{i=1}^{M} D^{-l_i} \leq 1$$

Battle plan to solve (Pb1)

1. find a lower bound for $L(C)$,
2. find an upper bound,
3. construct an optimal prefix code.

Lower bound of prefix code

Theorem (Lower bound on the expected length of any prefix code [CT Th. 5.3.1])

The **expected length** $L(C)$ *of* **any prefix** *D-ary code for the r.v. X taking M values in $\mathcal{X} = \{\alpha_1, \alpha_2, \ldots, \alpha_M\}$, with probabilities $p_1, p_2, \ldots, p_M$,*
is greater than or equal to the **entropy** $H(X)/\log(D)$ *i.e.,*

$$L(C) = \sum_{i=1}^{M} p_i l_i \geq \frac{H(X)}{\log D}$$

with equality iff $p_i = D^{-l_i}$, for $i = 1, \ldots, M$, and $\sum_{i=1}^{M} D^{-l_i} = 1$

Lower and upper bound of Shannon code

Definition

A **Shannon** code (defined on an alphabet with D symbols) for each source symbol $\alpha_i \in \mathcal{X} = \{\alpha_i\}_{i=1}^{M}$ of probability $p_i > 0$, assigns codewords of length $L(C(\alpha_i)) = l_i = \lceil -\log_D(p_i) \rceil$.

Theorem (Expected length of a Shannon code [CT Sec. 5.4])

Let X be a r.v. with entropy $H(X)$. The **Shannon code** *for the source X can be turned* **into a prefix code** *and its* **expected length $L(C)$ satisfies**

$$\frac{H(X)}{\log D} \leq L(C) < \frac{H(X)}{\log D} + 1 \quad (1)$$

Lower and upper bound of Shannon code

Definition

A **Shannon** code (defined on an alphabet with D symbols) for each source symbol $\alpha_i \in \mathcal{X} = \{\alpha_i\}_{i=1}^{M}$ of probability $p_i > 0$, assigns codewords of length $L(C(\alpha_i)) = l_i = \lceil -\log_D(p_i) \rceil$.

Theorem (Expected length of a Shannon code [CT Sec. 5.4])

Let X be a r.v. with entropy $H(X)$. The **Shannon code** *for the source X can be turned* **into a prefix code** *and its* **expected length $L(C)$ satisfies**

$$\frac{H(X)}{\log D} \leq L(C) < \frac{H(X)}{\log D} + 1 \qquad (1)$$

Corollary

Let X be a r.v. with entropy $H(X)$. There **exists a prefix code** *with expected length $L(C)$ that satisfies (1).*

Lower and upper bound of optimal code

Definition

A code is **optimal** if it achieves the lowest expected length **among all prefix codes**.

Theorem (Lower and upper bound on the expected length of an optimal code [CT Th 5.4.1])

Let X be a r.v. with entropy $H(X)$. Any **optimal code** C^* *for* X *with codeword lengths* $l_1^*, \dots, l_M^*$ *and* **expected length** $L(C^*) = \sum p_i l_i^*$ **satisfies**

$$\frac{H(X)}{\log D} \leq L(C^*) < \frac{H(X)}{\log D} + 1$$

Quiz Improve the upper bound.

Improved upper bound

Theorem (Lower and upper bound on the expected length of an optimal code for a sequence of symbols[CT Th 5.4.2])

Let X be a r.v. with entropy $H(X)$. Any **optimal code** *C^* for a sequence of s i.i.d. symbols $(X_1, ..., X_s)$ with expected length $L(C^*)$* **per source symbol** *X satisfies*

$$\frac{H(X)}{\log D} \leq L(C^*) < \frac{H(X)}{\log D} + \frac{1}{s}$$

This is the zero-error source coding Theorem.

Same average achievable rate for **vanishing** and **error-free** compression.
This is not true in general for distributed coding of multiple sources.

Construction of optimal codes

Lemma (Necessary conditions on optimal prefix codes[CT Le5.8.1])

Given a **binary** *prefix code C with word lengths $l_1, ..., l_M$ associated with a set of symbols with probabilities $p_1, ..., p_M$.*
Without loss of generality, assume that
(i) $p_1 \geq p_2 \geq ... \geq p_M$,
(ii) a group of symbols with the same probability is arranged in order of increasing codeword length (i.e. if $p_i = p_{i+1} = ... = p_{i+r}$ then $l_i \leq l_{i+1} ... \leq l_{i+r}$).
If **C is optimal** *within the class of* **prefix** *codes,* **C must satisfy**:

1. **higher** *probabilities symbols have* **shorter** *codewords ($p_i > p_k \Rightarrow l_i < l_k$),*
2. *the two least probable symbols have* **equal** *length ($l_M = l_{M-1}$),*
3. *among the codewords of* **length l_M**, *there must be at least two words that* **agree in all digits except the last**.

Huffman code

Let X be a r.v. taking M values in $\mathcal{X} = \{\alpha_1, \alpha_2, ..., \alpha_M\}$, with probabilities $p_1, p_2, ..., p_M$ s.t. $p_1 \geq p_2 \geq ... \geq p_M$.
Each letter α_i is associated with a codeword W_i i.e. a sequence of l_i letters, where each letter takes value in an alphabet of size $D = 2$.

1. **Combine** the last 2 symbols α_{M-1}, α_M into an **equivalent symbol** $\alpha_{M,M-1}$ w.p. $p_M + p_{M-1}$,
2. Suppose we can construct an optimal code C_2 $(W_1, ..., W_{M,M-1})$ for the new set of symbols $\{\alpha_1, \alpha_2, ..., \alpha_{M,M-1}\}$.
   Then, construct the code C_1 for the original set as:

$$
\begin{aligned}
C_1: \quad \alpha_i &\mapsto W_i, \ _{\forall i \in [1, M-2]}, \text{ same codewords as in } C_2 \\
\alpha_{M-1} &\mapsto W_{M,M-1}\, 0 \\
\alpha_M &\mapsto W_{M,M-1}\, 1
\end{aligned}
$$

Theorem (Huffman code is optimal [CT Th. 5.8.1])