

Information Theory . . . Without Theory

Bixio Rimoldi

School of Computer and Communication Sciences
École Polytechnique Fédérale de Lausanne



Scientific Event to Commemorate 40 Years of ATED
Manno, October 7, 2011

The purpose of this talk ...

is to give a taste of information theory ...

with as little theory as possible.

A Bit of History

Shannon, The Genius That Created Information Theory

Like few other sciences, Information Theory has a clear birthday, namely when Claude Shannon's revolutionary paper was published.

- 1916 Shannon was born
- 1936 (age 20) bachelor degrees from University of Michigan (Math and EE)
- 1937 MS Thesis work from MIT (EE Department) on how to use Boolean algebra for the design of relay and switching circuits
- 1940 Ph.D. Thesis at MIT (Math Department) on genetics

*A quarter of a century later, H. H. Goldstine, in his book *The Computer from Pascal to Von Neumann*, called this work “one of the most important master’s thesis ever written...a landmark in that it helped to change digital circuit design from an art to a science.”*

- 1945 Wrote a classified Bell Labs memorandum “A Mathematical Theory of Cryptography”
- 1948 Birth of Information Theory with the publication “A Mathematical Theory of Communication”. It is widely recognized that this and subsequent Shannon’s papers have altered most profoundly all aspects of communication theory and practice.
- Shannon died in 2001

Shannon

(continued)

Throughout his careers Shannon made numerous inventions working across disciplines. (127 papers on source/channel coding, cryptography, computers, circuits, games theory, and genetics, mostly as a single author.)

Shannon has designed and built (mostly in his basement) a number of amusing devices including:

- obstacle-avoiding “turtles”
- electrical mouse that finds its way through a maze
- game playing machines of various types and sizes
- machine with 2 hands that juggles 3 balls
- the “mind reading” machine
- a computer program to play chess

Information Theory: Three Fundamental Quests

Source Coding

First Quest: The concept of information:

- which objects contain/produce information
- how do we measure the amount of information

These questions have led to the field of **source coding**.

Second Quest: How to protect information from natural deterioration:

- in storage (think of a DVD)
- in transmission (think of noise/interference/fading)

These questions have led to the field of **channel coding**.

Third Quest: How to protect information from potential fraud:

- can we ensure privacy
- can we ensure authenticity

These questions pertain to **cryptography**.

Recall the first-quest questions:

- which objects contain/produce information
- how do we measure the amount of information

We define the amount of information produced by a source as the smallest number of bits (in average) needed to store and reconstruct the source output.

Smallest # of Bits: The Context Matters

Example: The professor gives a reading assignment every week. What should the teacher communicates at the beginning of each week?

Sample options:

- Professor posts the file (lots of bytes)
- Professor posts a book title and chapter number (few bytes)
- Professor posts a chapter number (few bits)

In each case the professor tell the students which element of a set he has selected. The number of bits depends on the size of the set.

Smallest # of Bits: The Probability Matters Too

Example: How many bits are necessary to store a sequence of grades?

Assume the grades take value in the set $\{3, 4, 5\}$.

Assume they are given with probabilities $\frac{1}{4}$, $\frac{1}{2}$, $\frac{1}{4}$, respectively.

We store using the following *prefix-free* binary code:

3	→	01
4	→	1
5	→	00

The average number of bits is then

$$L = \frac{1}{4} \times 2 + \frac{1}{2} \times 1 + \frac{1}{4} \times 2 = 1.5$$

Hence the information content is (less or) equal 1.5 bits per grade.

First Quest: We Are Ready To Reformulate and Answer

- Q: How much information is produced by a source?
- A: By definition, the amount of information produced by a source equals the minimum number of bits necessary (in average) to store its output

- Q: How do we assess/measure it concretely?
- A: It is the source entropy $H = \sum_{i=1}^n p_i \log_2 \frac{1}{p_i}$ [bits]
(This has to be proved. See later.)

- Q: how do we model information sources?
- A: by random variables

Example

Computing the Entropy

For the above example the probabilities are $\frac{1}{4}$, $\frac{1}{2}$, $\frac{1}{4}$

Hence the source entropy is

$$H = \frac{1}{4} \log 4 + \frac{1}{2} \log 2 + \frac{1}{4} \log 4 = 1.5 \text{ [bits]}.$$

Kraft Inequality

For Prefix-Free Codeword Lengths

Lemma: For any binary prefix-free code, the codeword lengths l_1, l_2, \dots, l_n must satisfy Kraft's inequality:

$$\sum_{i=1}^n 2^{-l_i} \leq 1$$

Conversely, given a set of codeword lengths that satisfy this inequality, there exists a prefix-free code with these word lengths.

Proof: To be done on the board.

Theorem: To store a random variable X by means of a prefix-free binary code we need in average

$$H(X) := \sum_{i=1}^n p_i \log_2 \frac{1}{p_i} \quad [\text{bits}].$$

Proof of necessity: Minimize $L = \sum_{i=1}^n p_i l_i$ subject to $\sum_{i=1}^n 2^{-l_i} \leq 1$.

Using the method of Lagrange multipliers we find that we should choose $l_i = \log_2 \frac{1}{p_i}$.

This proves only that $H(X)$ is a lower bound since we have neglected that l_i needs to be integer.

Example

Storing Grades

Recall the probabilities: $\frac{1}{4}$, $\frac{1}{2}$, $\frac{1}{4}$.

Recall the prefix-free binary code

3 \rightarrow 01

4 \rightarrow 1

5 \rightarrow 00

Recall the average number of bits: $L = \frac{1}{4} \times 2 + \frac{1}{2} \times 1 + \frac{1}{4} \times 2 = 1.5$

Compare to the entropy: $H = \frac{1}{4} \log 4 + \frac{1}{2} \log 2 + \frac{1}{4} \log 4 = 1.5$

What is going on? We chose $l_i = \log_2 \frac{1}{p_i}$.

The Impact of Source Coding

Source coding has led to various optimal compression algorithms, including

- Huffman source coding (shortest expected length), Lempel/Ziv universal coding, arithmetic coding, etc.

and has tight connections with seemingly unrelated areas such as

- the problem of simulating random variables
- gambling
- the game of 20 questions

A Useful Fact

For a fixed alphabet size n , the uniform distribution has the largest entropy, namely

$$H_n = \sum_{i=1}^n p_i \log_2 \frac{1}{p_i} = \log n \quad [\text{bits}].$$

The proof is straightforward (e.g. using Lagrange multipliers).

Example of Seemingly Unrelated Application

13 +1 Billiard Balls Game

Let us play the following game



- there 13 billiard balls numbered $1, \dots, 13$ and a white ball
- you **suspect** that **one of the numbered balls may be heavier/lighter** but don't know which.

- you are allowed to use a balance 3 times
- and then tell, for each ball, whether it is heavier/lighter/normal.

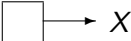



Back To Quest 2: Protection From Natural Deterioration

- We store and transmit information by means of signals
- In many situations we can not prevent signals from becoming corrupted (natural events)
- Can we ensure information integrity and at what price?

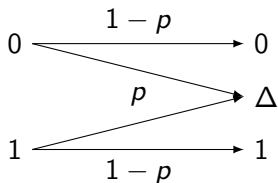
The answer is given by channel coding.

Analogy Between Source and Channel Coding

	Source	Channel
Diagram		
Mathematical Model	Random Variable X described by P_X	Collection of RVs described by $P_{V U}$
Fundamental Measure	Source Entropy $H(X)$	Channel Capacity $\max_{P_U} H(V) - H(V U)$
Operational Meaning	# of bits needed to store X (in average)	# of bits that can be sent reliably (in average)

Channel Capacity Example

The Binary Erasure Channel is defined as follows:



It's channel capacity is $C = 1 - p$ [bits per channel use].

We know this since Shannon's 1948 paper but only recently we know how to transmit reliably at C [bits/use].

The Evolution of Channel Coding

Relentless efforts to approach capacity on various channel classes have produced a series of success stories:

- Bell Labs motto for the late 50's: "coding is dead"
- In the 60's: success on wide-band channels ("deep space")
- In the 70's: success on coding for storage on optical disks
- In the 80's: success on narrowband channels (phone line)
- In the 90's: success on cellular networks (GSM)
- In the 00's: the golden age of Wi-Fi

Along the path several code techniques have been developed (Hamming codes (40's), RM codes (50's), RS codes (60's), Turbo codes (90's), LDPC codes (60s and 90's)).

Information theory has synergy with many fields

- mathematics (inequalities)
- computer science (Kolmogorov complexity)
- physics (thermodynamics)
- life science (genomics, evolution)
- probability theory and statistic (hypothesis testing, large deviation theory)
- economics (portfolio theory, gambling)

The Future of Information Theory

Guess 1: Network Information Theory

Network information theory for wireless communication is still widely uncharted territory.

The big picture unfolds slowly. Work will continue there.

The Future of Information Theory

Guess 2: Approaching Capacity on Fiber-Optic Communication

Past progress in fiber-optic communication has been hardware driven.

At 100 Tbits/sec, the bottle neck is no longer the electronic but the channel impairments. This is where information theory and coding can help.

Information theory will help finding the fundamental limits and coding will find ways to approach those limits.

The work has started.

- The most widely used introductory textbook in information theory is T. Cover and J. Thomas, *Elements of Information Theory* 2d ed., Wiley, 2006.
- Shannon's biography and papers are published in *Claude Elwood Shannon Collected Papers* Edited by N.J.A. Sloane and Aaron D. Wyner, IEEE Press, 1993.

Thank You!