Introduction to Information Theory

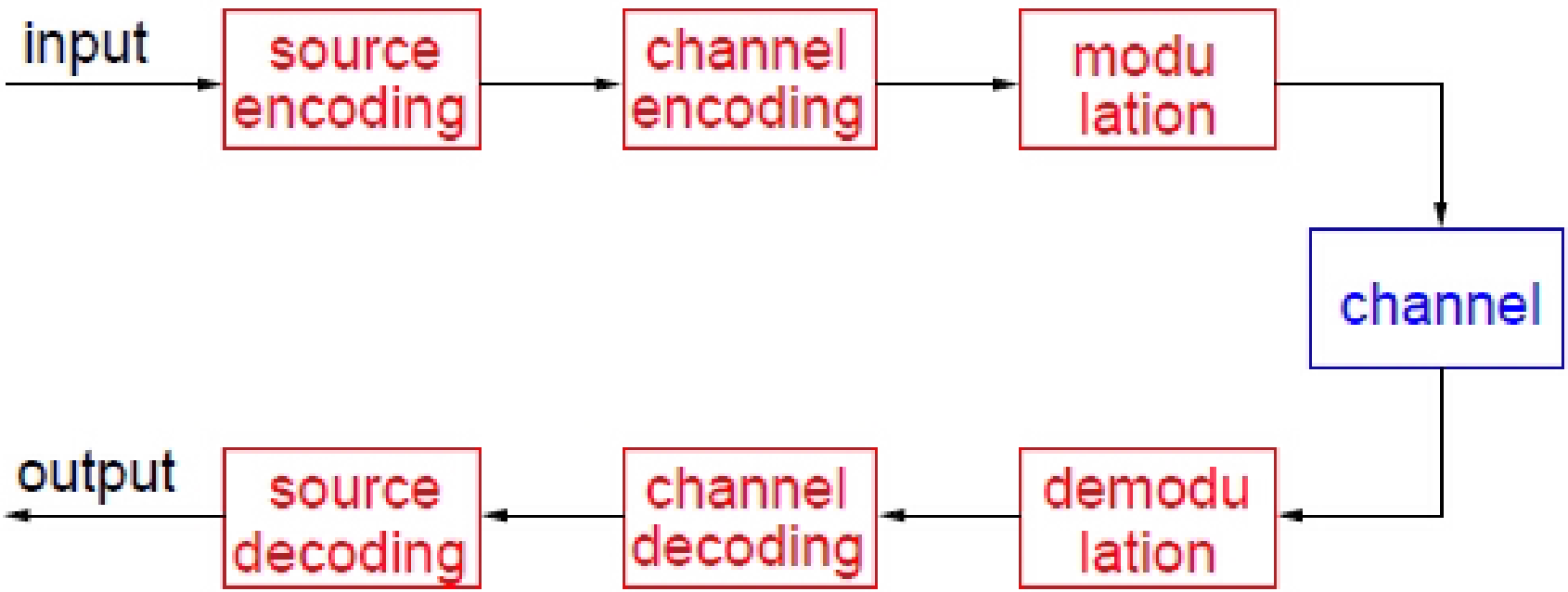
Impressive slide presentations

Radu Trîmbițaș

UBB

October 2012

Transmission of information



- A digital transmission scheme generally involves (see figure)
- Input/output are considered digital (analogue sampled/quantised)

Introduction

- Information theory answers two fundamental questions in communication theory:

Introduction

- Information theory answers two fundamental questions in communication theory:
  - What is the ultimate data compression? Answer: the entropy H

Introduction

- Information theory answers two fundamental questions in communication theory:
  - What is the ultimate data compression? Answer: the entropy H
  - What is the ultimate transmission rate of communication? Answer: the channel capacity C

Introduction

- Information theory answers two fundamental questions in communication theory:
  - What is the ultimate data compression? Answer: the entropy H
  - What is the ultimate transmission rate of communication? Answer: the channel capacity C
- For this reason information theory is considered to be a subset of communication theory

Introduction

- Information theory answers two fundamental questions in communication theory:
  - What is the ultimate data compression? Answer: the entropy H
  - What is the ultimate transmission rate of communication? Answer: the channel capacity C
- For this reason information theory is considered to be a subset of communication theory
- It is much more, for it has fundamental contributions in

Introduction

- Information theory answers two fundamental questions in communication theory:
  - What is the ultimate data compression? Answer: the entropy H
  - What is the ultimate transmission rate of communication? Answer: the channel capacity C
- For this reason information theory is considered to be a subset of communication theory
- It is much more, for it has fundamental contributions in
  - statistical physics (thermodynamics);

Introduction

- Information theory answers two fundamental questions in communication theory:
  - What is the ultimate data compression? Answer: the entropy H
  - What is the ultimate transmission rate of communication? Answer: the channel capacity C
- For this reason information theory is considered to be a subset of communication theory
- It is much more, for it has fundamental contributions in
  - statistical physics (thermodynamics);
  - computer science (Kolmogorov complexity or algorithmic complexity);

Introduction

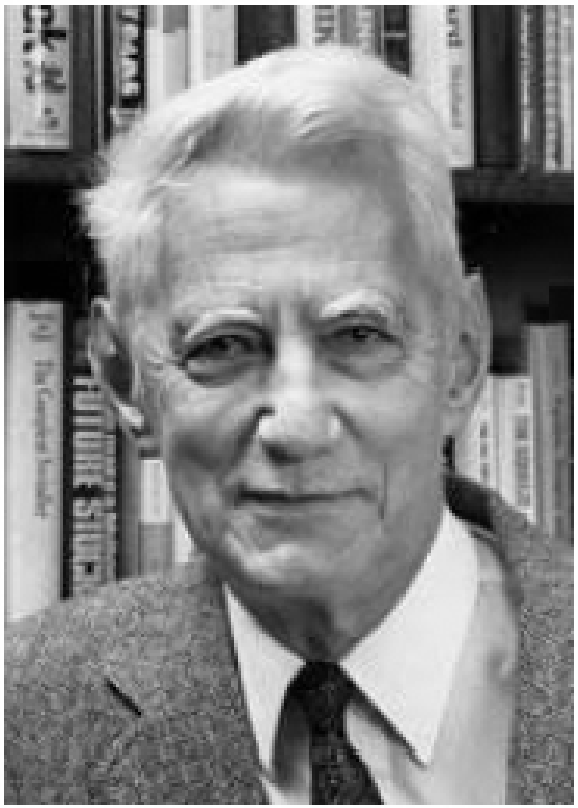
- Information theory answers two fundamental questions in communication theory:
  - What is the ultimate data compression? Answer: the entropy H
  - What is the ultimate transmission rate of communication? Answer: the channel capacity C
- For this reason information theory is considered to be a subset of communication theory
- It is much more, for it has fundamental contributions in
  - statistical physics (thermodynamics);
  - computer science (Kolmogorov complexity or algorithmic complexity);
  - statistical inference (Occam's Razor: "the simplest explanation is best");

Introduction

- Information theory answers two fundamental questions in communication theory:
  - What is the ultimate data compression? Answer: the entropy H
  - What is the ultimate transmission rate of communication? Answer: the channel capacity C
- For this reason information theory is considered to be a subset of communication theory
- It is much more, for it has fundamental contributions in
  - statistical physics (thermodynamics);
  - computer science (Kolmogorov complexity or algorithmic complexity);
  - statistical inference (Occam's Razor: "the simplest explanation is best");
  - probability and statistics (error exponents for optimal hypothesis testing and estimation.

Claude Shannon

- Claude Elwood Shannon, "A mathematical theory of communication," Bell System Technical Journal, 1948.
- Almost all important topics in information theory were initiated by Shannon

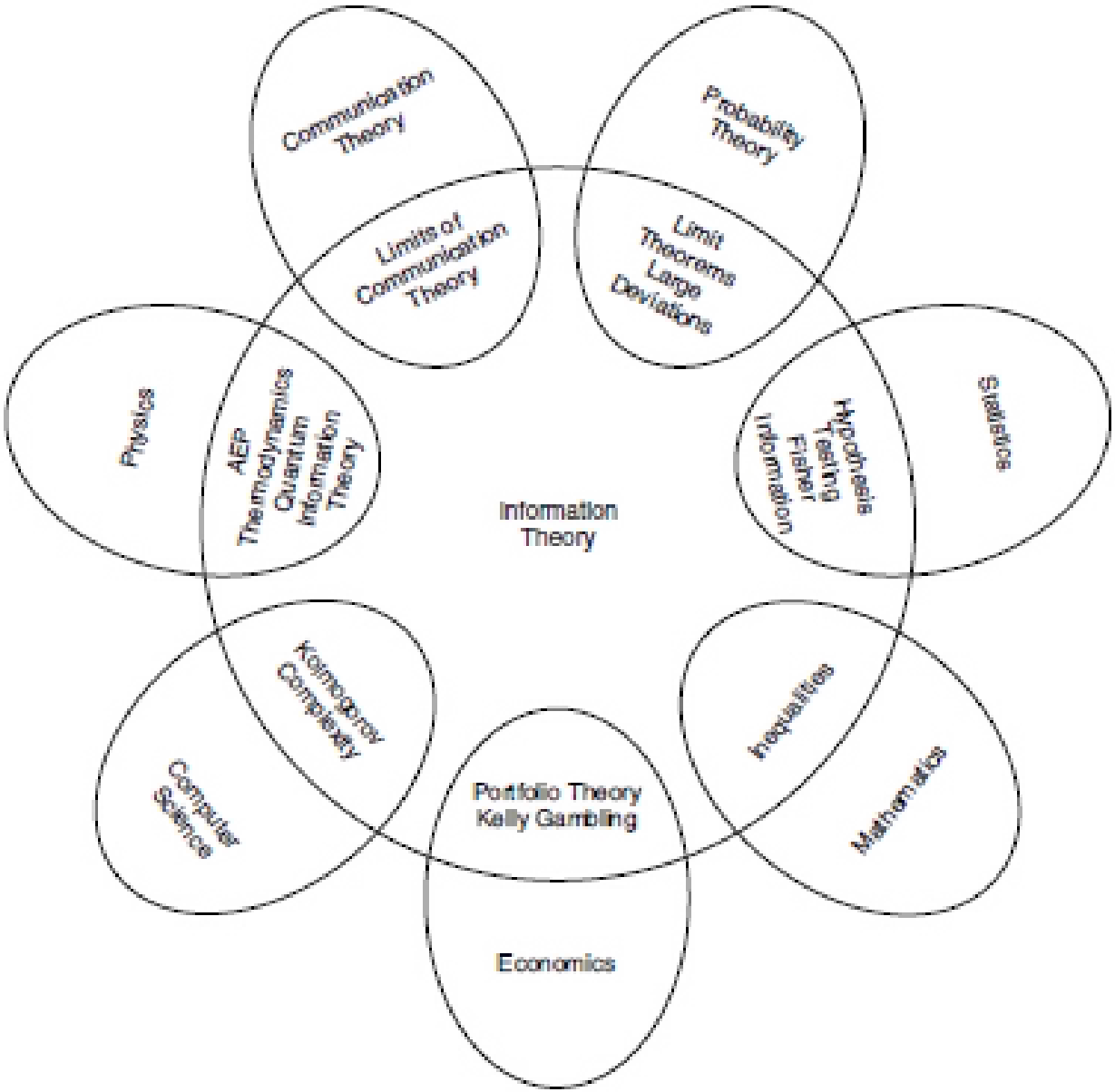


1916-2001

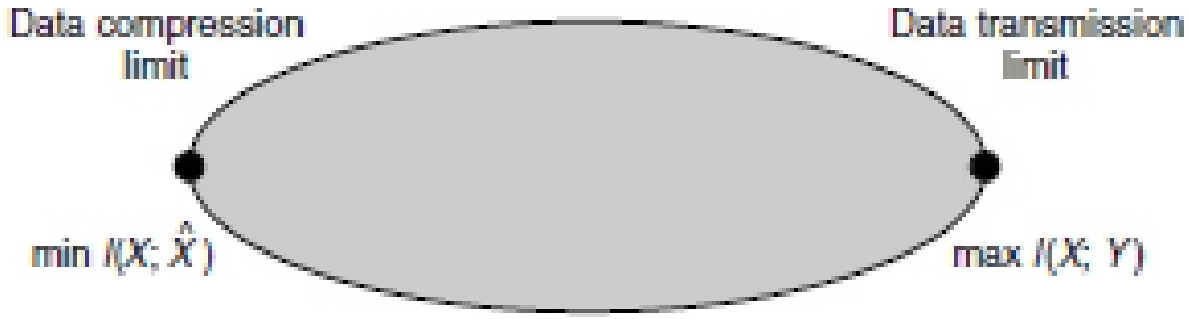
Origin of Information Theory

- Common wisdom in 1940s:
  - It is impossible to send information error-free at a positive rate
  - Error control by using retransmission: rate $\longrightarrow 0$ if error-free
- Still in use today
  - ARQ (automatic repeat request) in TCP/IP computer networking
- Shannon showed reliable communication is possible for all rates below channel capacity
- As long as source entropy is less than channel capacity, asymptotically error-free communication can be achieved
- And anything can be represented in bits
  - Rise of digital information technology

Relationship of information theory to other fields



IT and limits of communication theory



- Information theory today represents the extreme points of the set of all possible communication schemes
- The data compression minimum $I(X;\widehat{X})$ lies at one extreme of the set of communication ideas. All data compression schemes require description rates at least equal to this minimum.
- At the other extreme is the data transmission maximum $I(X;Y)$, known as the channel capacity.
- All modulation schemes and data compression schemes lie between these limits.

Relationship of information theory to other fields I

- **Computer Science (Kolmogorov Complexity)**. Kolmogorov, Chaitin, and Solomonoff put forth the idea that the complexity of a string of data can be defined by the length of the shortest binary computer program for computing the string. Thus, the complexity is the minimal description length. This definition of complexity turns out to be universal, that is, computer independent, and is of fundamental importance. Thus, Kolmogorov complexity lays the foundation for the theory of descriptive complexity. Gratifyingly, the Kolmogorov complexity K is approximately equal to the Shannon entropy H if the sequence is drawn at random from a distribution that has entropy H. So the tie-in between information theory and Kolmogorov complexity is perfect. Indeed, we consider Kolmogorov complexity to be more fundamental than Shannon entropy. It is the ultimate data compression and leads to a logically consistent procedure for inference.

Relationship of information theory to other fields II

- **Physics (Thermodynamics)**. Statistical mechanics is the birthplace of entropy and the second law of thermodynamics. Entropy always increases. Among other things, the second law allows one to dismiss any claims to perpetual motion machines.
- **Mathematics (Probability Theory and Statistics)**. The fundamental quantities of information theory—entropy, relative entropy, and mutual information—are defined as functionals of probability distributions. In turn, they characterize the behavior of long sequences of random variables and allow us to estimate the probabilities of rare events (large deviation theory) and to find the best error exponent in hypothesis tests.
- **Economics (Investment)**. Repeated investment in a stationary stock market results in an exponential growth of wealth. The growth rate of the wealth is a dual of the entropy rate of the stock market. The parallels between the theory of optimal investment in the stock market and information theory are striking.

Relationship of information theory to other fields III

- **Computation vs. Communication**. As we build larger computers out of smaller components, we encounter both a computation limit and a communication limit. Computation is communication limited and communication is computation limited. These become intertwined, and thus all of the developments in communication theory via information theory should have a direct impact on the theory of computation.
- **Philosophy of Science (Occam's Razor)**. William of Occam said "Causes shall not be multiplied beyond necessity," or to paraphrase it, "The simplest explanation is best." Solomonoff and Chaitin argued persuasively that one gets a universally good prediction procedure if one takes a weighted combination of all programs that explain the data and observes what they print next. Moreover, this inference will work in many problems not handled by statistics. For example, this procedure will eventually predict the subsequent digits of π.When this procedure is applied to coin flips that come up heads with probability

Relationship of information theory to other fields IV

0.7, this too will be inferred. When applied to the stock market, the procedure should essentially find all the "laws" of the stock market and extrapolate them optimally. In principle, such a procedure would have found Newton's laws of physics. Of course, such inference is highly impractical, because weeding out all computer programs that fail to generate existing data will take impossibly long. We would predict what happens tomorrow a hundred years from now.

Course Objectives

- In this course we will (focus on communication theory):
  - Define what we mean by information.
  - Show how we can compress the information in a source to its theoretically minimum value and show the tradeoff between data compression and distortion.
  - Prove the channel coding theorem and derive the information capacity of different channels.
  - Generalize from point-to-point to network information theory.

Course Contents

1. Review of basic notions of probability theory
2. Entropy, informational divergence, joint entropy, conditional entropy, mutual information, conditional information
3. Fano's inequality and data processing lemma
4. Information sources, entropy rate, Markov chains
5. Data compression, optimal codes, Huffman codes, Shannon-Fano-Elias codes, Shannon codes
6. Channel capacity, the Channel Coding Theorem
7. Maximum entropy distributions, universal source coding

Basic rule of probability theory I

- S - *sample space*, elements of S *elementary events*, $A \subseteq S$ random *event*
- A collection $\mathcal{K}$ of events from S is a σ-*field* (σ-*algebra*) if
  1. $\mathcal{K} \neq \emptyset$
  2. $A \in \mathcal{K} \Longrightarrow \overline{A} \in \mathcal{K}$
  3. if $A_n \in \mathcal{K}$, $\forall n \in \mathbb{N}$, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{K}$
- $(S, \mathcal{K})$ *measurable space*
- $(S, \mathcal{K}, P)$ *probability space*, $A \subseteq S$ random event, $A \in \mathcal{K}$
- Axioms of probability $P : \mathcal{K} \longrightarrow \mathbb{R}$

  P1) $P(S) = 1$
  P2) $P(A) \geq 0, \forall A \in \mathcal{K}$

Basic rule of probability theory II

P3) If $A_i \cap A_j = \emptyset$, $i \neq j$, $i, j \in I \subseteq \mathbb{N}$, $A_i \in \mathcal{K}$

$$P\left(\bigcup_{i \in I} A_i\right) = \sum_{i \in I} P(A_i)$$

- Conditional probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \qquad \text{if } P(B) > 0.$$

- A and B are *independent events* if

$$P(A \cap B) = P(A)P(B)$$

- Properties
  - $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$

Basic rule of probability theory III

- *multiplication rule*

$$P(A_1 \cap \ldots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)\ldots P(A_n|A_1 \cap \cdots \cap A_{n-1})$$

- *total probability formula*: if $\{A_i\}_{i\in I}$ is a partition of S (i.e. $\bigcup_{i\in I} A_i = S$ and $A_i \cap A_j = \varnothing$, $\forall i \neq j$) then

$$P(A) = \sum_{i\in I} P(A_i)P(A|A_i)$$

- *Bayes' formula*: if $\{A_i\}_{i\in I}$ is a partition of S then

$$P(A_k|A) = \frac{P(A|A_k)P(A_k)}{\sum_{i\in I} P(A_i)P(A|A_i)}$$

- $P(A) = P(A|B)P(B) + P(A|\overline{B})P(\overline{B})$

Discrete Random Variables I

- $\mathcal{X}$ - the *alphabet* of X (the set of all values of X). X *discrete random variable*

$$X \sim \begin{pmatrix} x \\ p(x) \end{pmatrix}_{x \in \mathcal{X}}$$

- p is the *probability mass function* of X

$$p(x) = P(X = x)$$

- $p(x) \geq 0$, and $\sum_{x \in \mathcal{X}} p(x) = 1$
-

$$Y \sim \begin{pmatrix} y \\ p(y) \end{pmatrix}_{y \in \mathcal{Y}}$$

Discrete Random Variables II

- The *joint probability distribution* of X and Y

$$p(x,y) = P\left(X = x \cap Y = y\right) = p(y,x)$$
$$p(x,y) \geq 0, \qquad \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y) = 1$$

- *marginal probability distribution* of X and Y from joint distribution

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

Discrete Random Variables III

- *conditional probabilities*

$$p(x|y) = P\left(X = x|Y = y\right) = \frac{p(x,y)}{p(y)}$$

$$p(y|x) = P\left(Y = y|X = x\right) = \frac{p(x,y)}{p(x)} = \frac{p(x|y)p(y)}{p(x)}$$

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