Classical Information Theory

Adrian Hutter

ETH Zürich

March 2, 2009

Overview

Overview:

Overview

Overview:

- Fundamental Concepts

Overview

Overview:

- Fundamental Concepts
  - What is Communication?

Overview

Overview:

- Fundamental Concepts
  - What is Communication?
  - Logarithmic Measure of Information

Overview

Overview:

- Fundamental Concepts
  - What is Communication?
  - Logarithmic Measure of Information
  - Entropy

Overview

Overview:

- Fundamental Concepts
  - What is Communication?
  - Logarithmic Measure of Information
  - Entropy
  - Joint and Conditional Entropy, Mutual Information

Overview

Overview:

- Fundamental Concepts
  - What is Communication?
  - Logarithmic Measure of Information
  - Entropy
  - Joint and Conditional Entropy, Mutual Information
  - Asymptotics

Overview

Overview:

- Fundamental Concepts
  - What is Communication?
  - Logarithmic Measure of Information
  - Entropy
  - Joint and Conditional Entropy, Mutual Information
  - Asymptotics
- Data Compression

Overview

Overview:

- Fundamental Concepts
  - What is Communication?
  - Logarithmic Measure of Information
  - Entropy
  - Joint and Conditional Entropy, Mutual Information
  - Asymptotics
- Data Compression
  - Shannon's Source Coding Theorem

Overview

Overview:

- Fundamental Concepts
  - What is Communication?
  - Logarithmic Measure of Information
  - Entropy
  - Joint and Conditional Entropy, Mutual Information
  - Asymptotics
- Data Compression
  - Shannon's Source Coding Theorem
- Channel Coding

Overview

Overview:

- Fundamental Concepts
  - What is Communication?
  - Logarithmic Measure of Information
  - Entropy
  - Joint and Conditional Entropy, Mutual Information
  - Asymptotics
- Data Compression
  - Shannon's Source Coding Theorem
- Channel Coding
  - The Binary Symmetric Channel

Overview

Overview:

- Fundamental Concepts
  - What is Communication?
  - Logarithmic Measure of Information
  - Entropy
  - Joint and Conditional Entropy, Mutual Information
  - Asymptotics
- Data Compression
  - Shannon's Source Coding Theorem
- Channel Coding
  - The Binary Symmetric Channel
  - Channel Capacity

Overview

Overview:

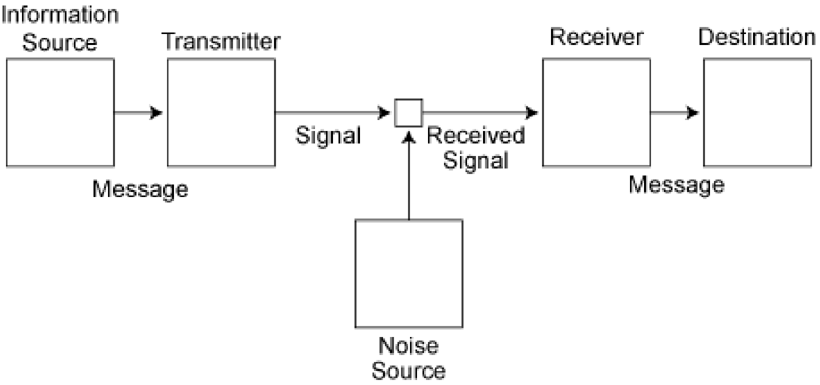
- Fundamental Concepts
  - What is Communication?
  - Logarithmic Measure of Information
  - Entropy
  - Joint and Conditional Entropy, Mutual Information
  - Asymptotics
- Data Compression
  - Shannon's Source Coding Theorem
- Channel Coding
  - The Binary Symmetric Channel
  - Channel Capacity
  - The Channel Coding theorem

Overview

Overview:

- Fundamental Concepts
  - What is Communication?
  - Logarithmic Measure of Information
  - Entropy
  - Joint and Conditional Entropy, Mutual Information
  - Asymptotics
- Data Compression
  - Shannon's Source Coding Theorem
- Channel Coding
  - The Binary Symmetric Channel
  - Channel Capacity
  - The Channel Coding theorem
- Summarization

Fundamental Concepts: A General Communication System



Fundamental Concepts: Definition of Communication

"The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point." (Shannon, 1948)

Fundamental Concepts: Definition of Communication

"The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point." (Shannon, 1948)

The significant aspect is that the actual message is one *selected from a set* of possible messages. If the number of messages in the set is finite then this number or any monotonic function of this number can be regarded as a measure of the information produced when one message is chosen from the set, all choices being equally like.

Fundamental Concepts: Logarithmic Measure of Information

Reasons for a logarithmic measure of information:

Fundamental Concepts: Logarithmic Measure of Information

Reasons for a logarithmic measure of information:

- Practicability (e.g. doubling the time of transmission squares the number of possible messages)

Fundamental Concepts: Logarithmic Measure of Information

Reasons for a logarithmic measure of information:

- Practicability (e.g. doubling the time of transmission squares the number of possible messages)
- Intuitive Feeling (e.g. two identical channels should have twice the capacity of one for transmitting information)

Fundamental Concepts: Logarithmic Measure of Information

Reasons for a logarithmic measure of information:

- Practicability (e.g. doubling the time of transmission squares the number of possible messages)
- Intuitive Feeling (e.g. two identical channels should have twice the capacity of one for transmitting information)
- Mathematical Suitability

Fundamental Concepts: Shannon Information

If the probability distribution of the possible outcomes is non-uniform, we define the *Shannon information content* of an outcome x to be
$h(x) = log \frac{1}{p(x)}$

Fundamental Concepts: Shannon Information

If the probability distribution of the possible outcomes is non-uniform, we define the *Shannon information content* of an outcome x to be

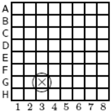
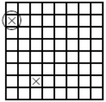
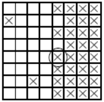
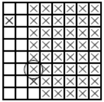
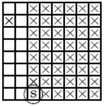
$h(x) = log \frac{1}{p(x)}$

$(log = log_2 \rightarrow$ binary digits or *bits*)

Fundamental Concepts: Shannon Information

If the probability distribution of the possible outcomes is non-uniform, we define the *Shannon information content* of an outcome x to be
$h(x) = log \frac{1}{p(x)}$
($log = log_2 \rightarrow$ binary digits or *bits*)
Improbable outcomes do convey more information than probable outcomes.

Fundamental Concepts: Battleships

					
move #	1	2	32	48	49
question	G3	B1	E5	F3	H3
outcome	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{y}$
$P(x)$	$\frac{63}{64}$	$\frac{62}{63}$	$\frac{32}{33}$	$\frac{16}{17}$	$\frac{1}{16}$
$h(x)$	0.0227	0.0230	0.0443	0.0874	4.0
Total info.	0.0227	0.0458	1.0	2.0	6.0

Fundamental Concepts: Shannon Entropy

Notation: Let X be a random variable with possible outcomes $x \in \Omega$ and respective probabilities $p(x)$.

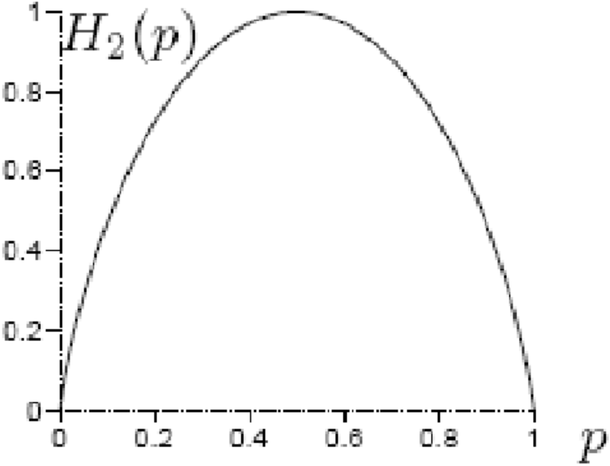
Fundamental Concepts: Shannon Entropy

Notation: Let X be a random variable with possible outcomes $x \in \Omega$ and respective probabilities $p(x)$.
The entropy of X is defined to be the average Shannon information content of an outcome:
$H(X) = -\sum_{x \in \Omega} p(x) logp(x)$

Fundamental Concepts: Shannon Entropy

Notation: Let X be a random variable with possible outcomes $x \in \Omega$ and respective probabilities $p(x)$.
The entropy of X is defined to be the average Shannon information content of an outcome:
$H(X) = -\sum_{x\in\Omega} p(x) log p(x)$
Theorem $H(X) \leq log \mid \Omega \mid$
with equality if X has a uniform distribution over Ω.

Fundamental Concepts: Entropy of an Event with Two Possible Outcomes



Fundamental Concepts: Joint Entropy

The *joint entropy* of a pair of discrete random variables (X,Y) is defined as
$H(X,Y) = -\sum_{x \in \Omega} \sum_{y \in \Phi} p(x,y) log p(x,y)$

Fundamental Concepts: Joint Entropy

The *joint entropy* of a pair of discrete random variables (X,Y) is defined as
$H(X,Y) = -\sum_{x\in\Omega}\sum_{y\in\Phi} p(x,y) log p(x,y)$
Note: $H(X,Y) \leq H(X) + H(Y)$ with equality for independent events, i.e. $p(x,y) = p(x)p(y)$

Fundamental Concepts: Conditional Entropy

The *conditional entropy* of a pair of discrete random variables (X,Y) is defined as
$H(Y \mid X) = \sum_{x \in \Omega} p(x) H(Y \mid X = x)$

Fundamental Concepts: Conditional Entropy

The *conditional entropy* of a pair of discrete random variables (X,Y) is defined as
$H(Y \mid X) = \sum_{x \in \Omega} p(x) H(Y \mid X = x)$
It measures how uncertain we are of Y on the average when we know X. (In general $H(X \mid Y) \neq H(Y \mid X)$)

Fundamental Concepts: Conditional Entropy

The *conditional entropy* of a pair of discrete random variables (X,Y) is defined as
$H(Y \mid X) = \sum_{x \in \Omega} p(x) H(Y \mid X = x)$
It measures how uncertain we are of Y on the average when we know X. (In general $H(X \mid Y) \neq H(Y \mid X)$)
Chain Rule $H(X, Y) = H(X) + H(Y \mid X)$

Fundamental Concepts: Mutual Information

The *mutual information* of a pair of discrete random variables (X,Y) is defined as
$I(X;Y) = \sum_{x \in \Omega} \sum_{y \in \Phi} p(x,y) log \frac{p(x,y)}{p(x)p(y)}$

Fundamental Concepts: Mutual Information

Properties of mutual information:

Fundamental Concepts: Mutual Information

Properties of mutual information:

- Symmetry: $I(X;Y) = I(Y;X)$

Fundamental Concepts: Mutual Information

Properties of mutual information:

- Symmetry: $I(X;Y) = I(Y;X)$
- Reduction of the uncertainity of X due to the knowledge of Y: $I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$

Fundamental Concepts: Mutual Information

Properties of mutual information:

- Symmetry: $I(X;Y) = I(Y;X)$
- Reduction of the uncertainity of X due to the knowledge of Y: $I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$
- Entropy as *self-information*: $I(X;X) = H(X)$

Fundamental Concepts: Mutual Information

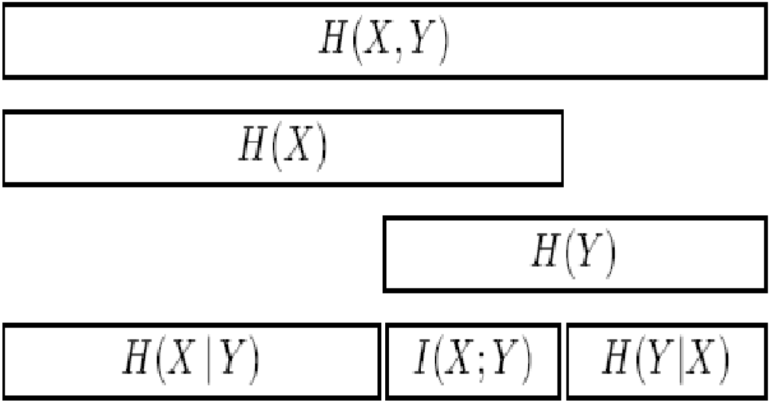
Properties of mutual information:

- Symmetry: $I(X;Y) = I(Y;X)$
- Reduction of the uncertainity of X due to the knowledge of Y:
  $I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$
- Entropy as *self-information*: $I(X;X) = H(X)$
- Non-negativity: $I(X;Y) \geq 0$
  (equality for independent variables)

Fundamental Concepts: Conditioning Reduces Entropy

Corollary $\mathrm{H(X} \mid Y) \leq H(X)$
("conditioning reduces entropy")

Fundamental Concepts: Roundup



Fundamental Concepts: Asymptotics

Asymptotic Equipartition Principle
For an ensemble of N i.i.d. (independent identically distributed) random variables, with N sufficiently large, the outcome is almost certain to belong to a subset of all possible outcomes having only 2^{NH} members, each having propability close to 2^{-NH}.
(skip ϵ, δ,...)

Data Compression: Shannon's Source Coding Theorem

Compression (using fewer bits than an unencoded representation would use) helps reduce the consumption of expensive resources, such as hard disk space or transmission bandwidth.

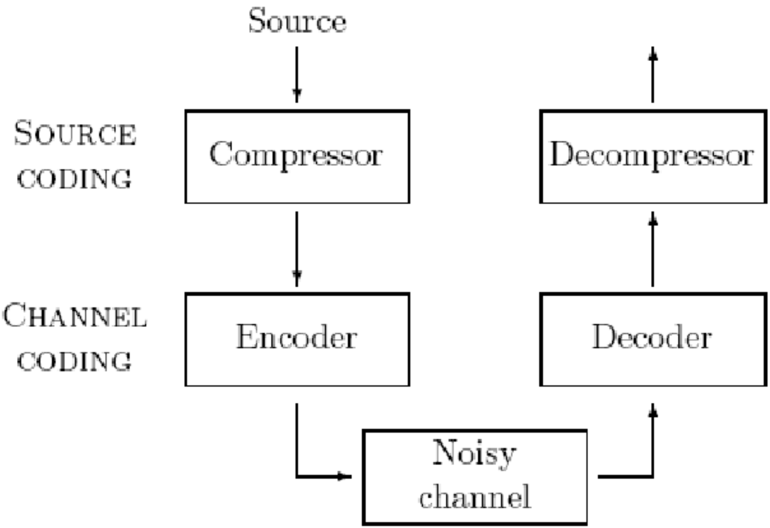
How much can the output of a source be compressed by use of the redundancy of the outcome?

What ist the minimum memory size from which the input can be recovered reliably?

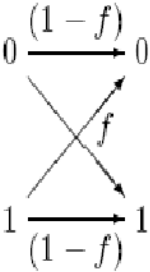
Data Compression: Shannon's Source Coding Theorem

Compression (using fewer bits than an unencoded representation would use) helps reduce the consumption of expensive resources, such as hard disk space or transmission bandwidth.
How much can the output of a source be compressed by use of the redundancy of the outcome?
What ist the minimum memory size from which the input can be recovered reliably?
Equivalent to the AEP is
Shannon's source coding theorem
N i.i.d. random variables each with entropy H can be compressed into more than NH bits with negligible risk of information loss, as $N \to \infty$; conversely if they are compressed into fewer than NH bits it is virtually certain that information will be lost.

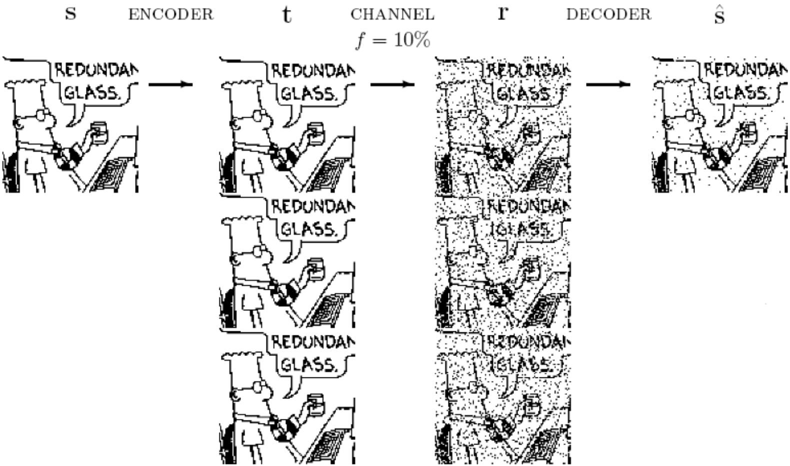
Channel Coding: The Noisy Channel



Channel Coding: The Binary Symmetric Channel



Channel Coding: The Binary Symmetric Channel



Channel Coding: The Binary Symmetric Channel

For a list of S codewords of length N, we define the *rate* to be the information in bits transmitted per use of the channel:
$R = \frac{logS}{N}$

Channel Coding: The Binary Symmetric Channel

For a list of S codewords of length N, we define the *rate* to be the information in bits transmitted per use of the channel:
$R = \frac{logS}{N}$
We may add redundancy in a controlled fashion to combat errors in the channel. However, by adding redundancy the rate of transmission decreases.

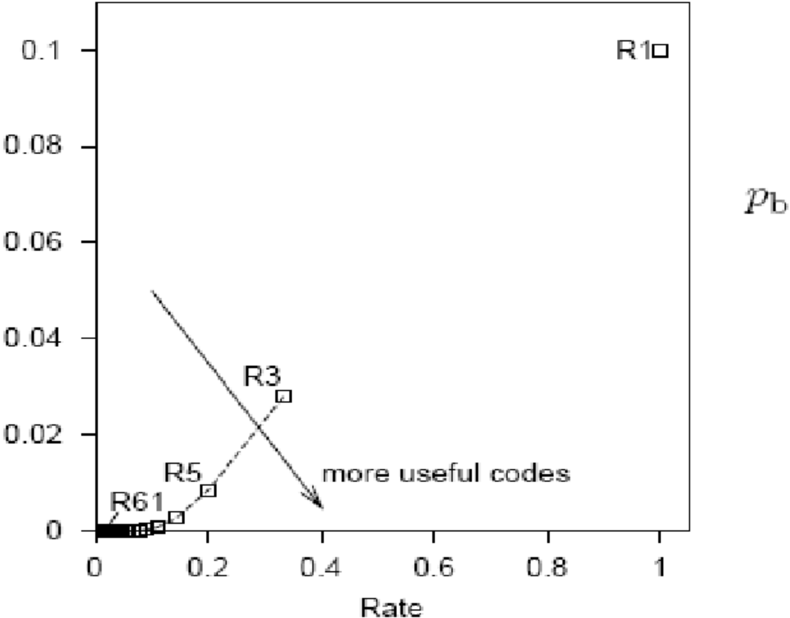
Channel Coding: The Binary Symmetric Channel

For a list of S codewords of length N, we define the *rate* to be the information in bits transmitted per use of the channel:
$R = \frac{logS}{N}$
We may add redundancy in a controlled fashion to combat errors in the channel. However, by adding redundancy the rate of transmission decreases.
We would expect that to make the probability of error approach zero, the redundancy of the encoding must increase indefinitely, and the rate of transmission therefore approach zero.

Channel Coding: The Binary Symmetric Channel

For a list of S codewords of length N, we define the *rate* to be the information in bits transmitted per use of the channel:
$R = \frac{logS}{N}$
We may add redundancy in a controlled fashion to combat errors in the channel. However, by adding redundancy the rate of transmission decreases.
We would expect that to make the probability of error approach zero, the redundancy of the encoding must increase indefinitely, and the rate of transmission therefore approach zero.
This is by no means true!

Channel Coding: The Binary Symmetric Channel



Channel Coding: Channel Coding Theorem

The *channel capacity* of a discrete memoryless channel is defined as
$C = max_{p(x)} I(X; Y)$.
where the maximum is taken over all possible input distributions
p(x).
(X: sender, Y: receiver)

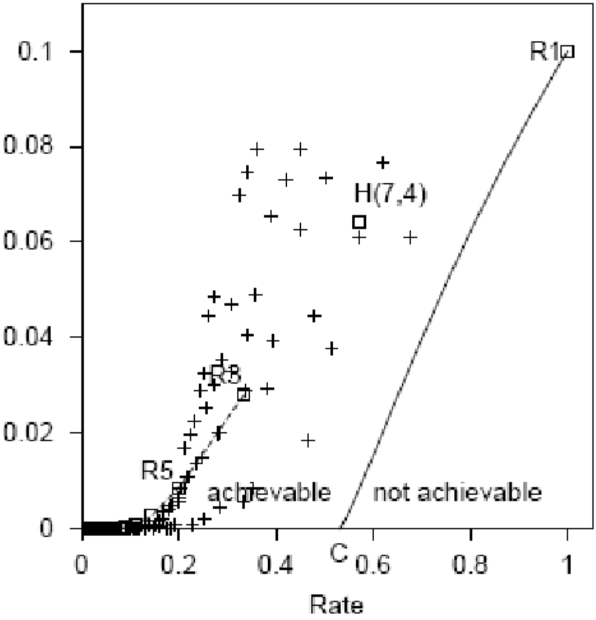
Channel Coding: Channel Coding Theorem

The *channel capacity* of a discrete memoryless channel is defined as
$C = max_{p(x)} I(X;Y)$.
where the maximum is taken over all possible input distributions
p(x).
(X: sender, Y: receiver)
For the binary channel we have $0 \leq C \leq 1$.

Channel Coding: Channel Coding Theorem

The *channel capacity* of a discrete memoryless channel is defined as
$C = max_{p(x)} I(X;Y)$.
where the maximum is taken over all possible input distributions
p(x).
(X: sender, Y: receiver)
For the binary channel we have $0 \leq C \leq 1$.
The Noisy-Channel Coding Theorem
For every rate R below the channel capacity C, for large enough N, there exists a code of length N and rate R and a decoding algorithm, with maximal probability of block error as small as desired.

Channel Coding: Channel Coding Theorem



Literature

Literature:

- Cover, Thomas M.: *Elements of information theory*
- David J.C. MacKay: *Information Theory, Inference, and Learning Algorithms*
- C.E. Shannon: *A Mathematical Theory of Communication*

Summarization

Summarization:

Summarization

Summarization:

- Shannon information content of an outcome x: $h(x) = log \frac{1}{p(x)}$

Summarization

Summarization:

- Shannon information content of an outcome x: $h(x) = log \frac{1}{p(x)}$
- Shannon's Formula for the entropy of an outcome:
  $H(X) = -\sum_{x \in \Omega} p(x) logp(x) \leq log \mid \Omega \mid$

Summarization

Summarization:

- Shannon information content of an outcome x: $h(x) = log \frac{1}{p(x)}$
- Shannon's Formula for the entropy of an outcome:
  $H(X) = -\sum_{x \in \Omega} p(x) log p(x) \leq log \mid \Omega \mid$
- Chain Rule: $H(X, Y) = H(X) + H(Y \mid X)$

Summarization

Summarization:

- Shannon information content of an outcome x: $h(x) = log\frac{1}{p(x)}$
- Shannon's Formula for the entropy of an outcome: $H(X) = -\sum_{x\in\Omega} p(x)logp(x) \leq log \mid \Omega \mid$
- Chain Rule: $H(X, Y) = H(X) + H(Y \mid X)$
- Conditioning reduces entropy: $H(X \mid Y) \leq H(X)$

Summarization

Summarization:

- Shannon information content of an outcome x: $h(x) = log\frac{1}{p(x)}$
- Shannon's Formula for the entropy of an outcome: $H(X) = -\sum_{x\in\Omega} p(x)logp(x) \leq log \mid \Omega \mid$
- Chain Rule: $H(X, Y) = H(X) + H(Y \mid X)$
- Conditioning reduces entropy: $H(X \mid Y) \leq H(X)$
- For large N the outcome is almost certain to belong to a set of 2^{NH} members, each having probability close to 2^{-NH}.

Summarization

Summarization:

- Shannon information content of an outcome x: $h(x) = log\frac{1}{p(x)}$
- Shannon's Formula for the entropy of an outcome: $H(X) = -\sum_{x\in\Omega} p(x)logp(x) \leq log \mid \Omega \mid$
- Chain Rule: $H(X,Y) = H(X) + H(Y \mid X)$
- Conditioning reduces entropy: $H(X \mid Y) \leq H(X)$
- For large N the outcome is almost certain to belong to a set of 2^{NH} members, each having probability close to 2^{-NH}.
- N i.i.d. random variables each with entropy H can be compressed into more than NH bits with negligible risk of information loss.

Summarization

Summarization:

- Shannon information content of an outcome x: $h(x) = log\frac{1}{p(x)}$
- Shannon's Formula for the entropy of an outcome: $H(X) = -\sum_{x \in \Omega} p(x) log p(x) \leq log \mid \Omega \mid$
- Chain Rule: $H(X, Y) = H(X) + H(Y \mid X)$
- Conditioning reduces entropy: $H(X \mid Y) \leq H(X)$
- For large N the outcome is almost certain to belong to a set of 2^{NH} members, each having probability close to 2^{-NH}.
- N i.i.d. random variables each with entropy H can be compressed into more than NH bits with negligible risk of information loss.
- There is a positive maximal rate at which information can be transmitted over a noisy channel with a propability of error as small as desired: The capacity of the channel.