Theory and Design of

Turbo and Related Codes

Lecture 9

Jossy Sayir & Gottfried Lechner

http://userver.ftw.at/~jossy/turbo/index.html

Kplus
Kompetenzzentren-Programm

Information Theory: the Physics of Coding

ftw.

- We have studied various coding methods and compared them in terms of performance, complexity and optimizability
- What are the theoretical limits for the performance of coding systems?
- How do code rate, error probability and the required block length evolve in function of the channel characteristics?
- How good are turbo, LDPC and RA codes really?

Information Theory: the Physics of Coding

- We studied this for the Binary Erasure Channel with erasure probability δ in Lecture 1 and concluded the following:
  - for rates $R > 1 - \delta$, the error probability is lower bounded, no matter how good the coding (genie)
  - for rates $R < 1 - \delta$, the error probability can be made arbitrarily small by random linear coding of increasing block length N, decoding by matrix inversion
- $1 - \delta$ is the "capacity" of the BEC
- There is no rate-reliability tradeoff for rates below capacity!

Information Theory: the Physics of Coding

- Irregular LDPC codes can be optimized so that the threshold for iterative decoding is arbitrarily close to the capacity of the BEC
- This lecture: study the information theory of general discrete memoryless channels
- Next lectures: study the design of turbo, LDPC and RA codes to approach the capacity of general memoryless channels

Acknowledgements

This lecture is a summary of 9 two-hour lectures from Jim Massey's Applied Digital Information Theory I (ADIT I) course at ETH Zürich
(with very minor changes)

Photo by Jossy Sayir, Yokohama June 2003

Notes available at
http://www.isi.ee.ethz. ch/education/public/free_docs.en.html

Discrete Probability Theory

- Let Ω be the set of possible outcomes of a random experiment
- An event is any subset of Ω
- A probability measure P assigns real number between 0 and 1 to events, such that $P(\emptyset) = 0$, $P(\Omega) = 1$ and for A and $B \subset \Omega$, $P(A \cup B) = P(A) + P(B)$
- Events, and only events can have probabilities!
- A random variable is a mapping from Ω to a finite or countably finite set of real numbers

Notation

- random variable: X (UPPER CASE)
- value of a random variable: x (lower case)
- $P_X(x)$ is short for Prob(X = x)
  (X=x defines an event!)
- P_X is the discrete probability distribution of X, i.e., $P_X(x) \geq 0$ for all x, and $\sum_x P_X(x) = 1$
- P_{XY} is the joint probability distribution of X and Y, i.e., the probability distribution of the vector-valued random variable [X,Y]

- allowed abuse of notation: P(x) means $P_X(x)$
- forbidden abuse of notation: P(1), or P(a) for $P_X(a)$

Entropy / Uncertainty

How many b-ary symbols are needed to identify the value of an L-ary random variable X?

Hartley's measure of information:

$$I(X) = \log_b L$$

Example: L = 4, b = 2, I(X) = 2
L = 8, I(X) = 3
L = 6, I(X) = 2.58

R.V.L. Hartley, "Transmission of Information", Bell Syst. Tech. J., Vol. 3, July 1928, pp. 535 - 564)

Entropy / Uncertainty

What if the values of X are not equally probable, i.e., X is not uniformly distributed?

Shannon's measure of information:
$H(P_X) = -\sum_x P_X(x) \log_b P_X(x)$

(convention: 0 log 0 = 0)

Units: b=2, "bits"
b = e, "nats"
b = 10, "Hartley"

Claude E. Shannon, "A Mathematical Theory of Communication", Bell System Tech. Journal, Vol. 27, July and October 1948, pp. 379 - 423 and pp. 623 - 656)

Entropy/Uncertainty

- $H(P_X)$ is the entropy of the probability distribution of X
- The word "entropy" is used because of the similarity between the formula for H and the entropy in physics
- $H(P_X)$ is a measure for our uncertainty about the value of X
- Alternatively, we write H(X) for $H(P_X)$
- Keep in mind that H(.) is always a function of a probability distribution!

Entropy/Uncertainty

Examples:

$L = 3, P_X(0) = .5, P_X(1) = .25, P_X(2) = .25$

$H(X) = .5 \log_2 2 + .25 \log_2 4 + .25 \log_2 4$
$= 1.5$ bit

$L = 2, P_X(0) = P_X(1) = 1/2$

$H(X) = 1$ bit

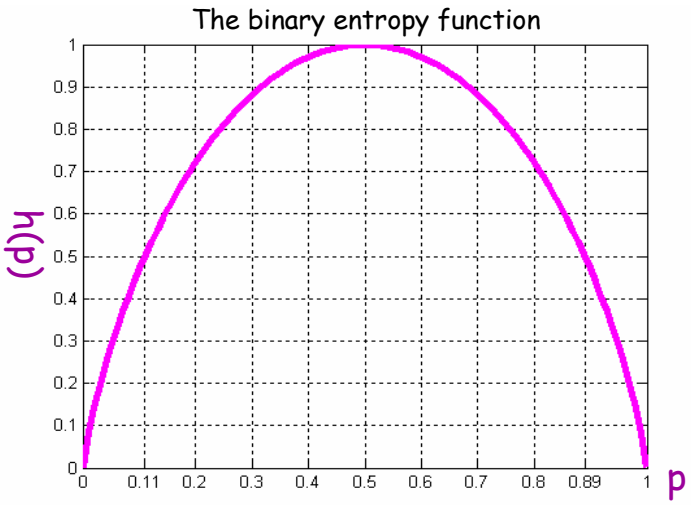
$L = 2, P_X(0) = 1, P_X(1) = 0$

$H(X) = 0$ bit

Binary Entropy Function

$L = 2, P_X(0) = p, P_X(1) = 1-p$

$H(X) = h(p)$



Entropy Bounds

For an L-ary random variable X,
$$0 \leq H(X) \leq \log_b L$$
with equality on the right when $P_X(x) = 1/L$ for all x,
with equality on the left when $P_X(x) = 1$ for one x
and $P(X) = 0$ for all other x.

Quiz: can you give an upper bound for the entropy of the set of all books in print? (independent of the choice of probability measure)

Answer: 10 Hartleys
Why? ➔ ISBN number!

Quiz

- What is H(XY)?

  Answer: $H(XY) = H(P_{XY})$
  $= \sum_x \sum_y P_{XY}(x,y) \log_b P_{XY}(x,y)$

- What about H(X|Y=y)?

  Warning: only defined if $P_Y(y) > 0$

  Answer: $H(X|Y=y) = H(P_{X|Y=y})$
  $= \sum_x P_{X|Y=y}(x) \log_b P_{X|Y=y}(x)$

Both $H(P_{XY})$ and $H(P_{X|Y=y})$ are just plain entropies of probability distributions!

Equivocation

The equivocation, or average conditional entropy, of X given Y is defined as

$$H(X|Y) = \sum_y P_Y(y)\, H(P_{X|Y=y})$$

Warning: do not confuse with $H(X|Y=y) = H(P_{X|Y=y})$

$H(X|Y=y)$ is an entropy conditioned on an event, whereas the equivocation $H(X|Y)$ is an "entropy" conditioned on a random variable

In reality, $H(X|Y)$ is not an entropy at all but an average entropy. It is a function of P_{XY}!

Equivocation

$$\begin{aligned} H(X|Y) &= -\sum_y P_Y(y) \sum_x P_{X|Y=y}(x) log_b P_{X|Y=y}(x) \\ &= -\sum_x \sum_y P_{XY}(x,y) log_b \frac{P_{XY}(x,y)}{P_Y(y)} \\ &= -\sum_x \sum_y P_{XY}(x,y) log_b \frac{P_{XY}(x,y)}{\sum_{x'} P_{XY}(x',y)} \end{aligned}$$

Properties of the equivocation

"Conditioning can only reduce entropy"

$$0 \leq H(X|Y) \leq H(X)$$

equality on the left if Y essentially determines X
equality on the right if X and Y are independent

Warning: $H(X|Y=y)$ can be larger than $H(X)$!!
"Conditioning on events can increase entropy"

Chain rule: $H(XY) = H(Y) + H(X|Y)$

$$H(X_1 \ldots X_N) = H(X_1) + H(X_2|X_1) + \ldots + H(X_N|X_1 \ldots X_{N-1})$$

Example

Random experiment: pick a student uniformly at random in class

r.v. X sex of student (0 male, 1 female)
r.v. Y row number where student is sitting

For some rows, $H(X|Y=y) > H(X)$

but $H(X|Y) < H(X)$

Mutual Information

The mutual information between X and Y is

$$I(X;Y) = H(X) - H(X|Y)$$
$$= H(Y) - H(Y|X)$$

$$0 \leq I(X;Y) \leq \min[H(X),H(Y)]$$

equality on left if X and Y independent
equality on right if X essentially determines Y or vice-versa

I(X;Y) is a function of the joint distribution P_{XY}

Mutual Information

$$\begin{aligned} I(X;Y) &= -\sum_x P_X(x) \log_b P_X(x) - H(X|Y) \\ &= -\sum_x \sum_y P_{XY} \log_b P_X(x) - H(X|Y) \\ &= -\sum_x \sum_y P_{XY}(x,y) \log_b \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} \end{aligned}$$

where $P_X(x) = \sum_{x'} P_{XY}(x',y)$
and $P_Y(y) = \sum_{y'} P_{XY}(x,y')$

Mutual Information

Conditioning:

$$I(X;Y|Z=z) = H(X|Z=z) - H(X|Y,Z=z)$$
$$I(X;Y|Z) = H(X|Z) - H(X|YZ)$$

Chain rule:

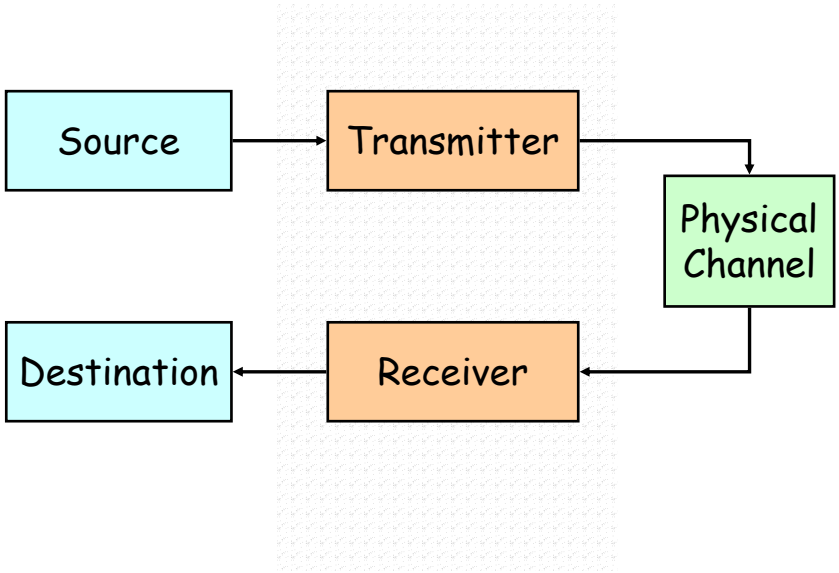
$$I(X;YZ) = I(X;Y) + I(X;Z|Y)$$
$$I(X;Y_1 \ldots Y_N) = I(X;Y_1) + I(X;Y_2|Y_1) + \ldots$$
$$\ldots + I(X;Y_N|Y_1 \ldots Y_{N-1})$$

Warning: $I(X;Y|Z)$ can be smaller or larger than $I(X;Y)$ (as Ralf and I found recently to our surprise…)

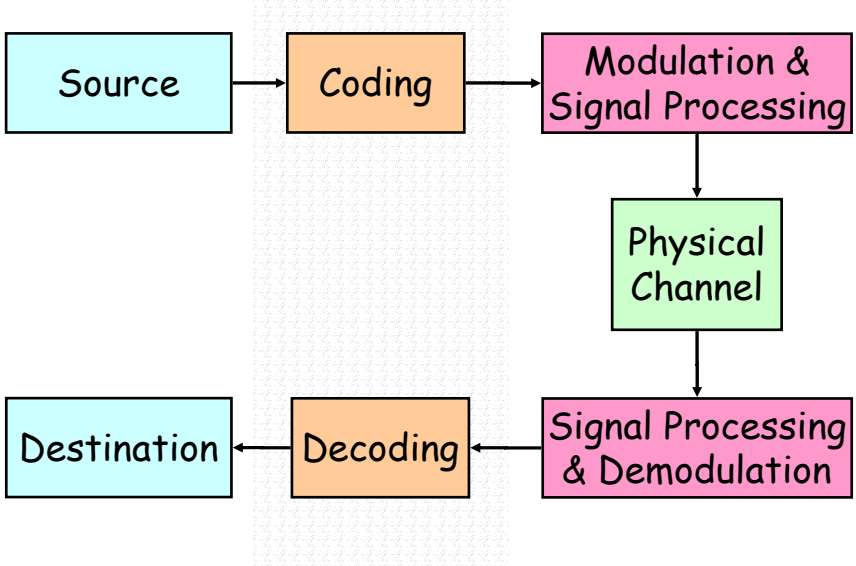
The essence of Mutual Information

- $I(X;Y)$ tells us how much uncertainty is reduced about X by knowing Y (or vice-versa)
- $I(X;Y)$ tells us how much information X gives about Y (or vice-versa)
- $I(X;Y)$ is a very general type of correlation measure: it is 0 when X and Y are independent (and thus uncorrelated) and maximized when X is a function of Y or vice-versa

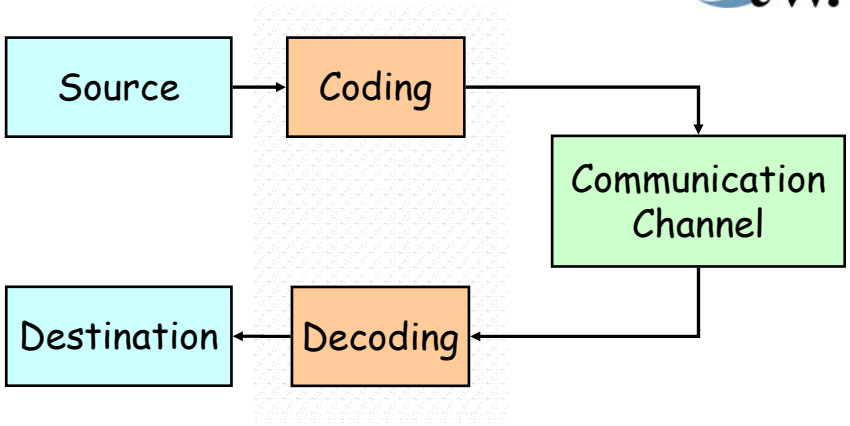
Communication Model



Communication Model

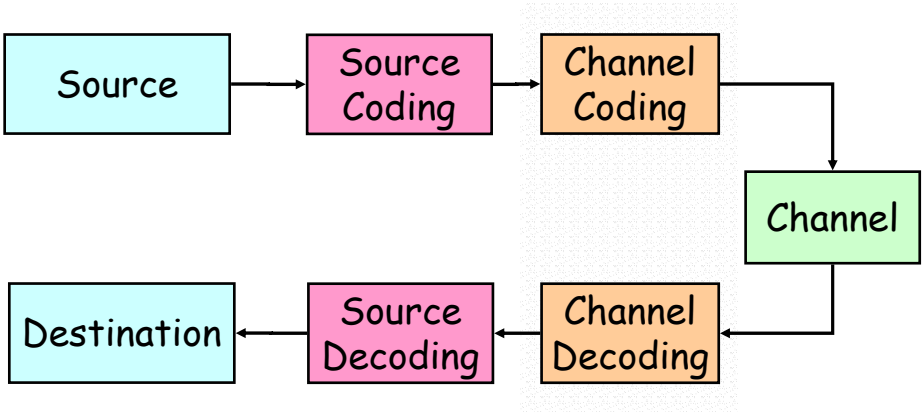


Communication Model

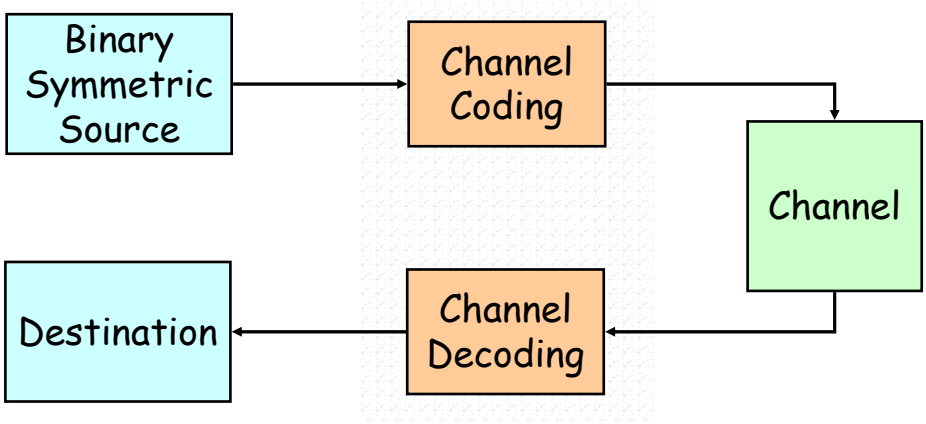


From a coding engineer's perspective, the channel is everything in a communication system that we are unwilling/unable/not allowed to modify

Communication Model

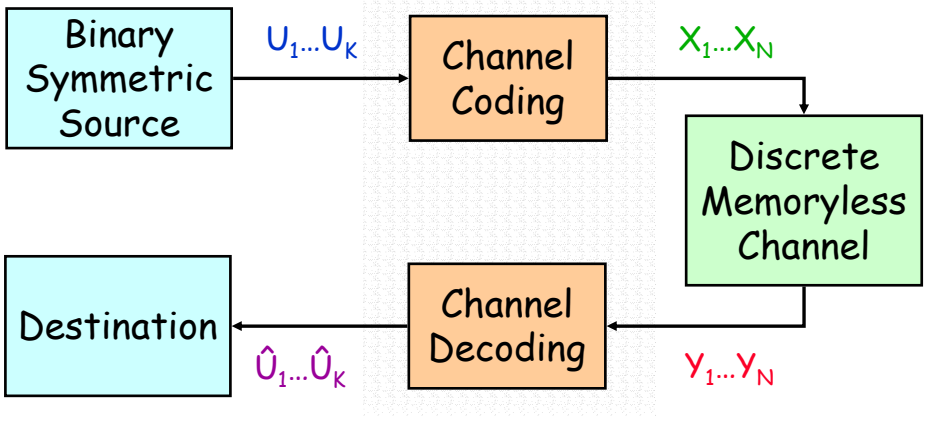


Communication Model



Remark: our coding systems are designed for binary-input memoryless channels (possibly continuous output) but in this lecture, we will only consider discrete (possibly non-binary) channels

Communication Model



Data Processing Theorem

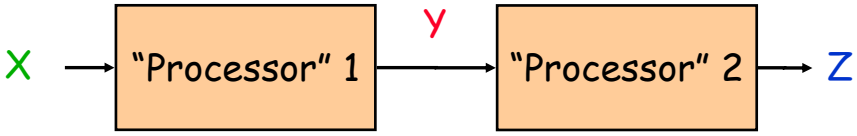
Let X, Y, Z form a Markov Chain, i.e.,

$$I(X;Z|Y) = 0$$

or $H(X|YZ) = H(X|Y)$ or $H(Z|XY) = H(Z|Y)$

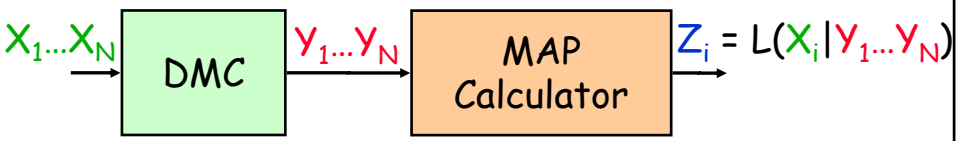
$I(X;Y) \geq I(X;Z)$ and $I(Y;Z) \geq I(X;Z)$
Information cannot be increased by processing!

When "=" in the left inequality, we say that Z is a sufficient statistic for X and then X, Z, Y also form a Markov Chain (but not necessarily Y,X,Z…)



Data Processing Lemma

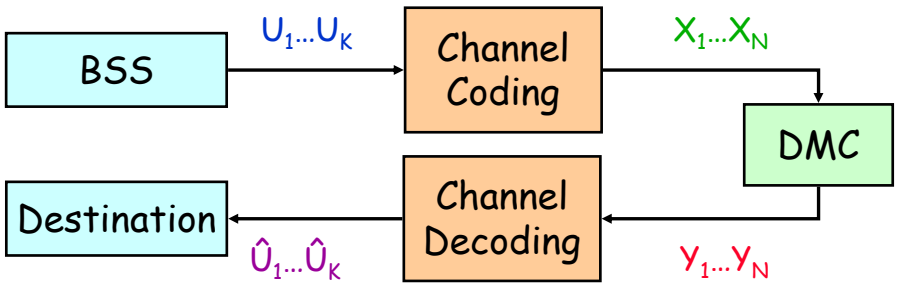
Interesting fact:



BCJR, no hard decision!

$I(X_i;Y_1\ldots Y_N) = I(X_i;Z_i)$
Z_i is a sufficient statistic for X_i

Data Processing Theorem



Consequence: $I(X_1 \ldots X_N; Y_1 \ldots Y_N) \geq I(U_1 \ldots U_K; \hat{U}_1 \ldots \hat{U}_K)$

Data Processing Lemma

We have

$$H(U_1 \ldots U_K | \hat{U}_1 \ldots \hat{U}_K) = \sum_i H(U_i | \hat{U}_1 \ldots \hat{U}_K U_1 \ldots U_{i-1})$$

(chain rule)

$$\leq \sum_i H(U_i | \hat{U}_i)$$

(conditioning reduces entropy)

We also have

$$H(U_1 \ldots U_K | \hat{U}_1 \ldots \hat{U}_K) = H(U_1 \ldots U_K) - I(U_1 \ldots U_K; \hat{U}_1 \ldots \hat{U}_K)$$

where $H(U_1 \ldots U_K) = K$ bits

(BSS)

Therefore

$$H(U_1 \ldots U_K | \hat{U}_1 \ldots \hat{U}_K) \geq K - I(X_1 \ldots X_N; Y_1 \ldots X_N)$$

(data processing lemma)

Data Processing Lemma

Therefore

$$\Sigma_i H(U_i|\hat{U}_i) \geq H(U_1 \ldots U_K|\hat{U}_1 \ldots \hat{U}_K) \geq K - I(X_1 \ldots X_N; Y_1 \ldots Y_N)$$

Furthermore,

$$H(Y_1 \ldots Y_N|X_1 \ldots X_N) = \Sigma_i H(Y_i|X_i)$$

(DMC)

and $H(Y_1 \ldots Y_N) = \Sigma_i H(Y_i|Y_1 \ldots Y_{i-1}) \leq \Sigma_i H(Y_i)$,

(chain rule) (conditioning reduces entropy)

therefore $I(X_1 \ldots X_N; Y_1 \ldots Y_N) \leq \Sigma_i I(X_i; Y_i)$

$$\Sigma_{i=1\ldots K} H(U_i|\hat{U}_i) \geq K - \Sigma_{i=1\ldots N} I(X_i; Y_i)$$

Converse coding theorem

We define the error probabilities $P_{ei} = P(\hat{U}_i \neq U_i)$ and $P_e = 1/K \Sigma_{i=1\ldots K} P_{ei}$

The optimal symbol decoder will choose $\hat{u}_i$ according to the maximum of $P(U_i|\text{observations})$.
Therefore, $P(U_i = 0|\hat{U}_i = 1) = P(U_i = 1|\hat{U}_i = 0) = P_{ei}$
and $P(U_i = 0|\hat{U}_i = 0) = P(U_i = 1|\hat{U}_i = 1) = 1 - P_{ei}$

We conclude that

$$H(U_i|\hat{U}_i) \leq h(P_{ei})$$

with equality for the optimal decoder

This is known as Fano's inequality

(given here only for the binary case)

Converse coding theorem

We now write

$$\sum_i H(U_i|\hat{U}_i) \leq \sum_i h(P_{ei})$$

(Fano's inequality)

$$\leq K\, h(1/K \sum_i P_{ei}) = K\, h(P_e)$$

(convexity of h(.))

We obtain: $h(P_e) \geq 1 - 1/K \sum_i I(X_i;Y_i)$

$I(X_i;Y_i)$ is a function of P_{XY}. The conditional distribution $P_{Y|X}$ is given by the channel. Therefore, we can choose P_{Xi} to maximize $I(X_i;Y_i)$ for any given channel to minimize the error probability.

Converse coding theorem

Let us define

$$C = \max_{P_X} I(X;Y)$$

C is called the capacity of the discrete memoryless channel

Converse coding theorem: if $R > C$,

$$P_e > h^{-1}(1 - C/R)$$

(Shannon, 1948)

Capacity

- of a BEC: $C = 1 - \delta$
- of a BSC: $C = 1 - h(\varepsilon)$
- of a general binary-input symmetric channel:

$$C = 1 - H(P_{Y|X=0})$$