


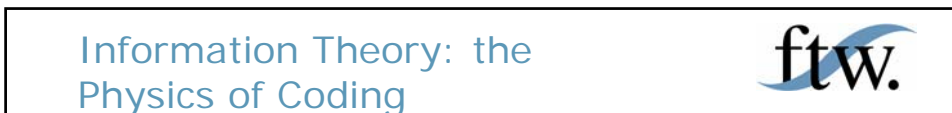
ftw.
Forschungszentrum Telekommunikation Wien

Theory and Design of
Turbo and Related Codes
Lecture 10

Jossy Sayir & Gottfried Lechner

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


Kompetenzzentrum-Programm



ftw.

Information Theory: the
Physics of Coding

- 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?

© ftw. 2005

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!

© ftw. 2005

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

© ftw. 2005

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)



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

© ftw. 2005

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

© ftw. 2005

Notation



- **random variable:** X (UPPER CASE)
- **value** of a random variable: x (lower case)
- $P_X(x)$ is short for $\text{Prob}(X = x)$
($X=x$ 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 $P_X(a)$

© ftw. 2005

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)

© ftw. 2005

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)

- $H(P_X)$ 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 $H(\cdot)$ is always a **function of a probability distribution!**

Entropy/Uncertainty



Examples:

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

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

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

$$H(X) = 1 \text{ bit}$$

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

$$H(X) = 0 \text{ bit}$$

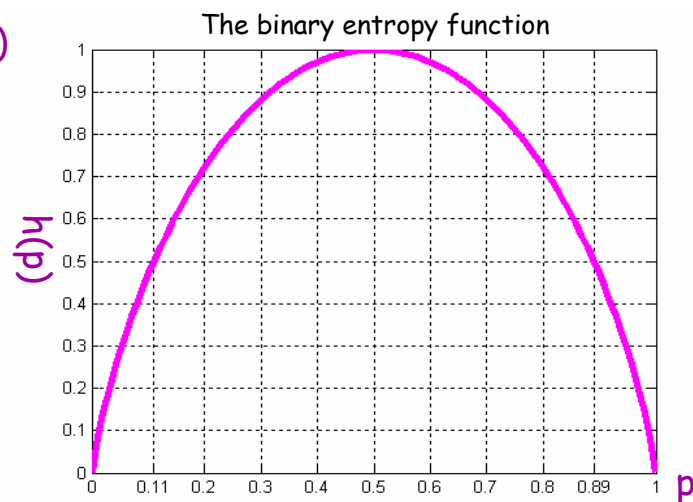
© ftw. 2005

Binary Entropy Function



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

$$H(X) = h(p)$$



© ftw. 2005

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!

- What is $H(XY)$?

$$\begin{aligned} \text{Answer: } H(XY) &= H(P_{XY}) \\ &= \sum_x \sum_y P_{XY}(x,y) \log_b P_{XY}(x,y) \end{aligned}$$

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

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

$$\begin{aligned} \text{Answer: } H(X|Y=y) &= H(P_{X|Y=y}) \\ &= \sum_x P_{X|Y=y}(x) \log_b P_{X|Y=y}(x) \end{aligned}$$

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

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} !

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

"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: $H(XY) = H(Y) + H(X|Y)$

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

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

The mutual information between X and Y is

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \end{aligned}$$

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

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}

© ftw. 2005

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

$$\begin{aligned} \text{where } P_X(x) &= \sum_{x'} P_{XY}(x',y) \\ \text{and } P_Y(y) &= \sum_{y'} P_{XY}(x,y') \end{aligned}$$

© ftw. 2005

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:

$$I(X;YZ) = I(X;Y) + I(X;Z|Y)$$

$$I(X;Y_1 \dots Y_N) = I(X;Y_1) + I(X;Y_2|Y_1) + \dots \\ \dots + I(X;Y_N|Y_1 \dots Y_{N-1})$$

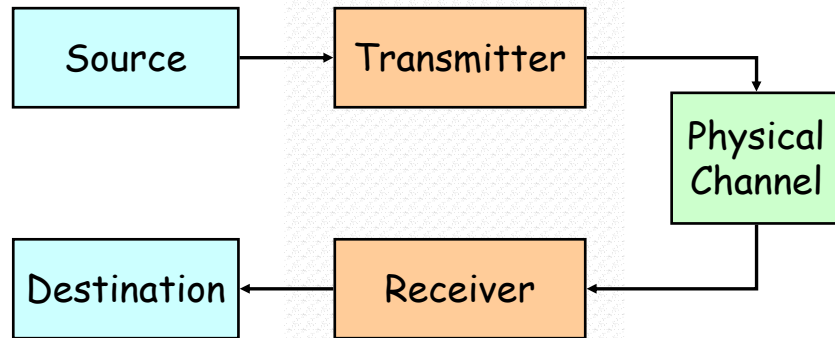
Warning: $I(X;Y|Z)$ can be smaller or larger than $I(X;Y)$

© ftw. 2005

- $I(X;Y)$ 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

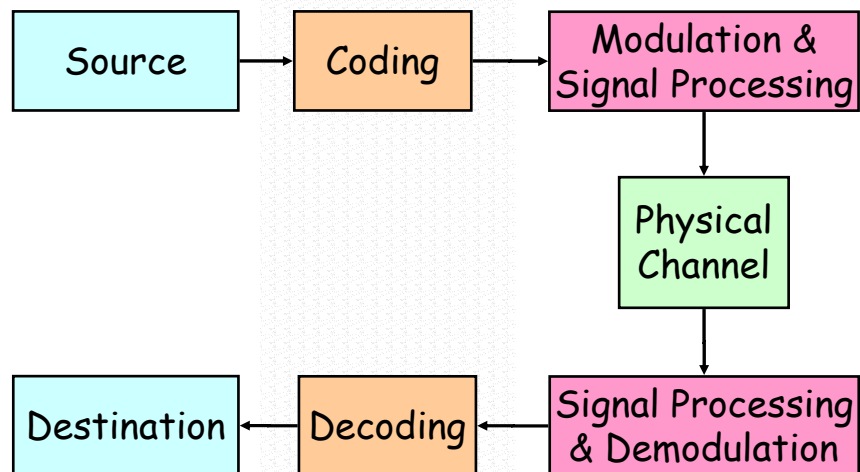
© ftw. 2005

Communication Model



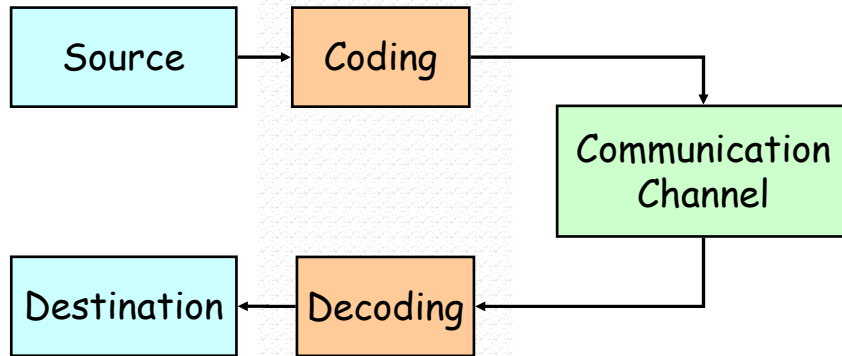
© ftw. 2005

Communication Model



© ftw. 2005

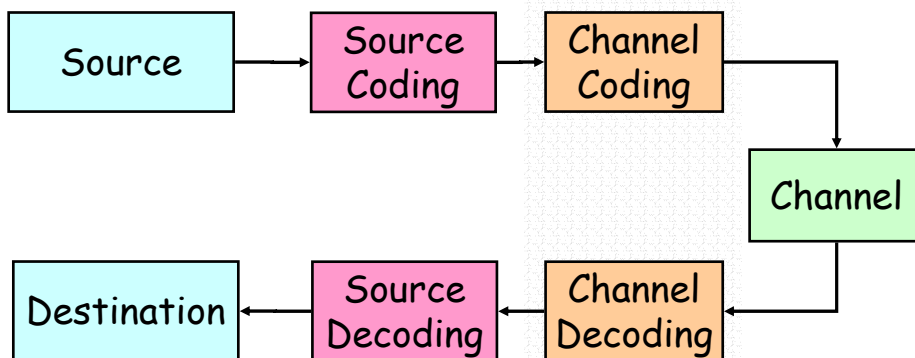
Communication Model



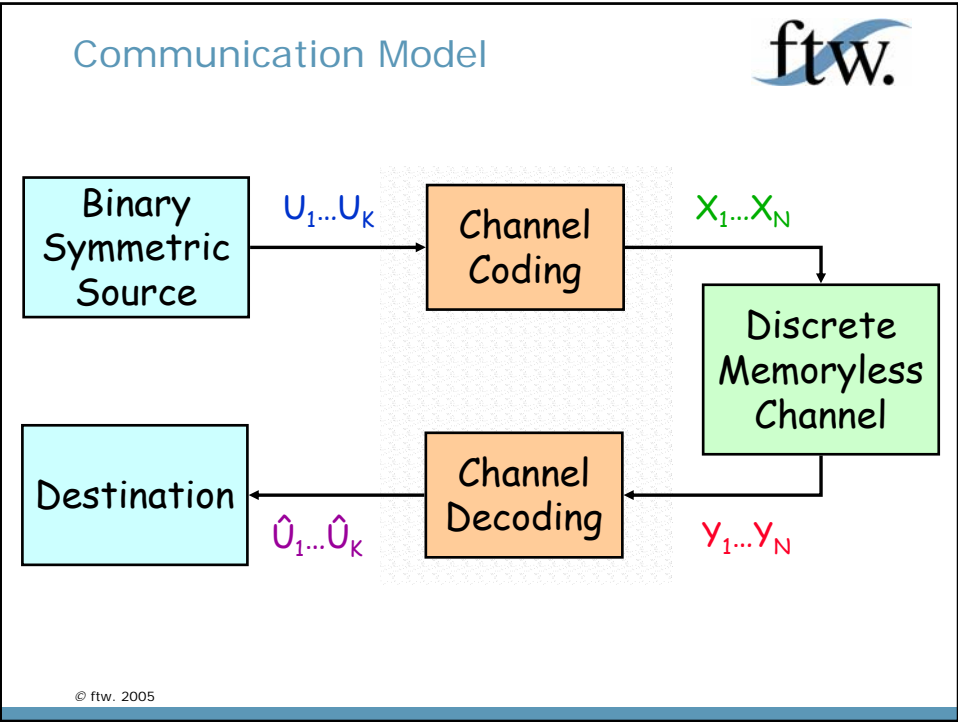
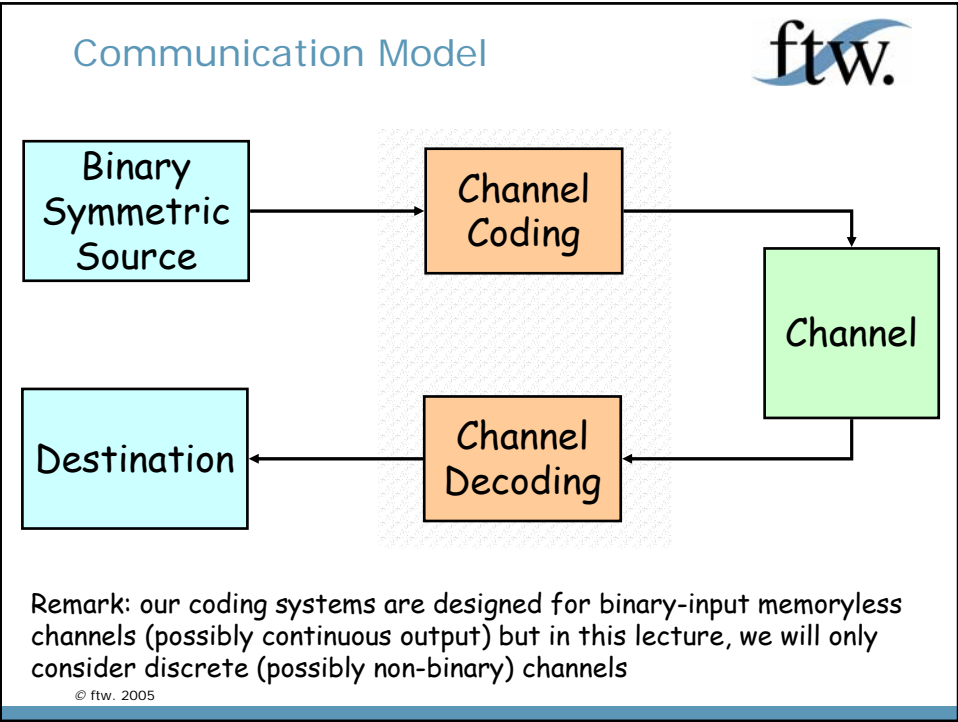
From a coding engineer's perspective, the **channel** is everything in a communication system that we are **unwilling/unable/not allowed to modify**

© ftw. 2005

Communication Model



© ftw. 2005



Data Processing Theorem



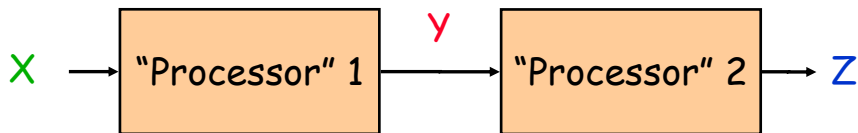
Let X, Y, Z form a Markov Chain, i.e.,

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

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 $Y, X, Z...$)

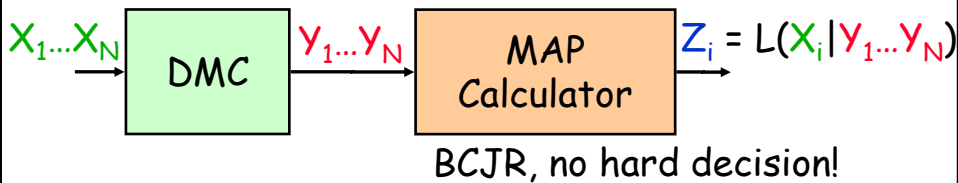


© ftw. 2005

Data Processing Lemma



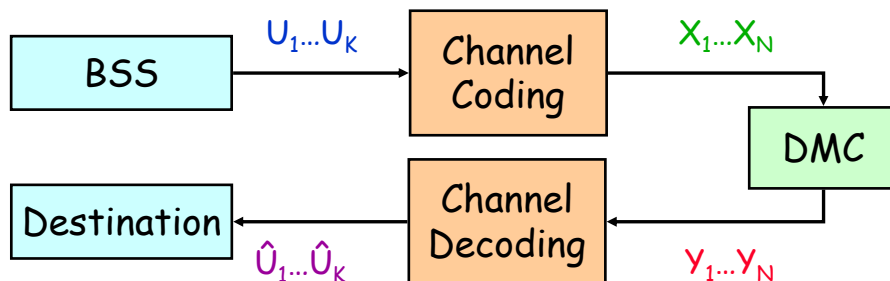
Interesting fact:



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

© ftw. 2005

Data Processing Theorem



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

© ftw. 2005

Data Processing Lemma



We have

$$\begin{aligned}
 H(U_1 \dots U_K | \hat{U}_1 \dots \hat{U}_K) &= \sum_i H(U_i | \hat{U}_1 \dots \hat{U}_K U_1 \dots U_{i-1}) \\
 &\quad \text{(chain rule)} \\
 &\leq \sum_i H(U_i | \hat{U}_i) \\
 &\quad \text{(conditioning reduces entropy)}
 \end{aligned}$$

We also have

$$\begin{aligned}
 H(U_1 \dots U_K | \hat{U}_1 \dots \hat{U}_K) &= H(U_1 \dots U_K) - I(U_1 \dots U_K; \hat{U}_1 \dots \hat{U}_K) \\
 \text{where } H(U_1 \dots U_K) &= K \text{ bits} \\
 &\quad \text{(BSS)}
 \end{aligned}$$

Therefore

$$\begin{aligned}
 H(U_1 \dots U_K | \hat{U}_1 \dots \hat{U}_K) &\geq K - I(X_1 \dots X_N; Y_1 \dots Y_N) \\
 &\quad \text{(data processing lemma)}
 \end{aligned}$$

© ftw. 2005

Data Processing Lemma



Therefore

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

Furthermore,

$$H(Y_1 \dots Y_N | X_1 \dots X_N) = \sum_i H(Y_i | X_i)$$

(DMC)

$$\text{and } H(Y_1 \dots Y_N) = \sum_i H(Y_i | Y_1 \dots Y_{i-1}) \leq \sum_i H(Y_i),$$

(chain rule) (conditioning reduces entropy)

therefore $I(X_1 \dots X_N; Y_1 \dots Y_N) \leq \sum_i I(X_i; Y_i)$

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

© ftw. 2005

Converse coding theorem



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

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)

© ftw. 2005

Converse coding theorem



We now write

$$\begin{aligned}\sum_i H(U_i | \hat{U}_i) &\leq \sum_i h(P_{e_i}) && \text{(Fano's inequality)} \\ &\leq K h(1/K \sum_i P_{e_i}) = K h(P_e) && \text{(convexity of } h(\cdot)\text{)}\end{aligned}$$

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_{x_i} to maximize $I(X_i; Y_i)$ for any given channel to minimize the error probability.

© ftw. 2005

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)

© ftw. 2005

Capacity

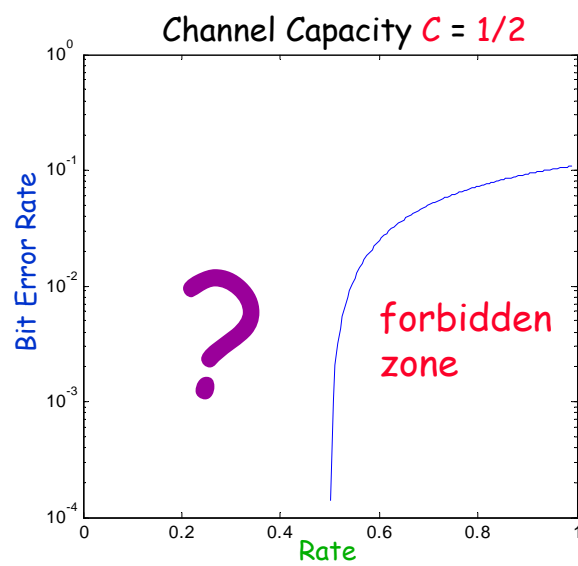


- of a BEC: $C = 1 - \delta$
- of a BSC: $C = 1 - h(\epsilon)$
- of a general binary-input symmetric channel:

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

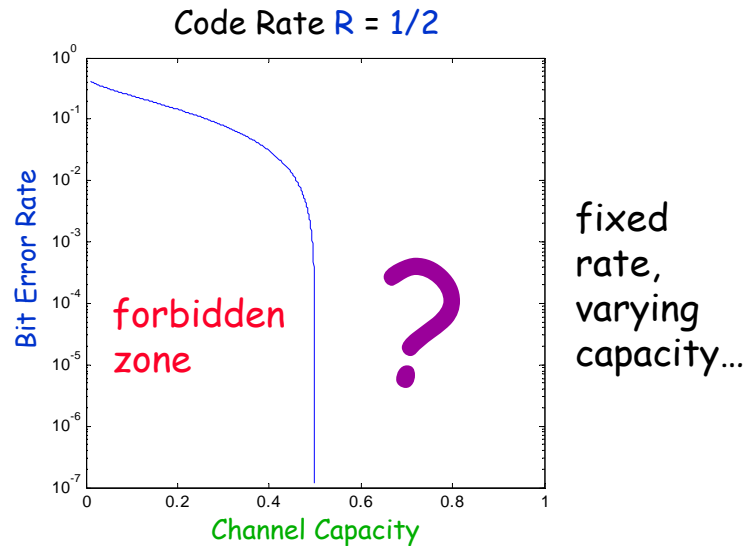
© ftw. 2005

Converse Coding Theorem



© ftw. 2005

Converse Coding Theorem



© ftw. 2005

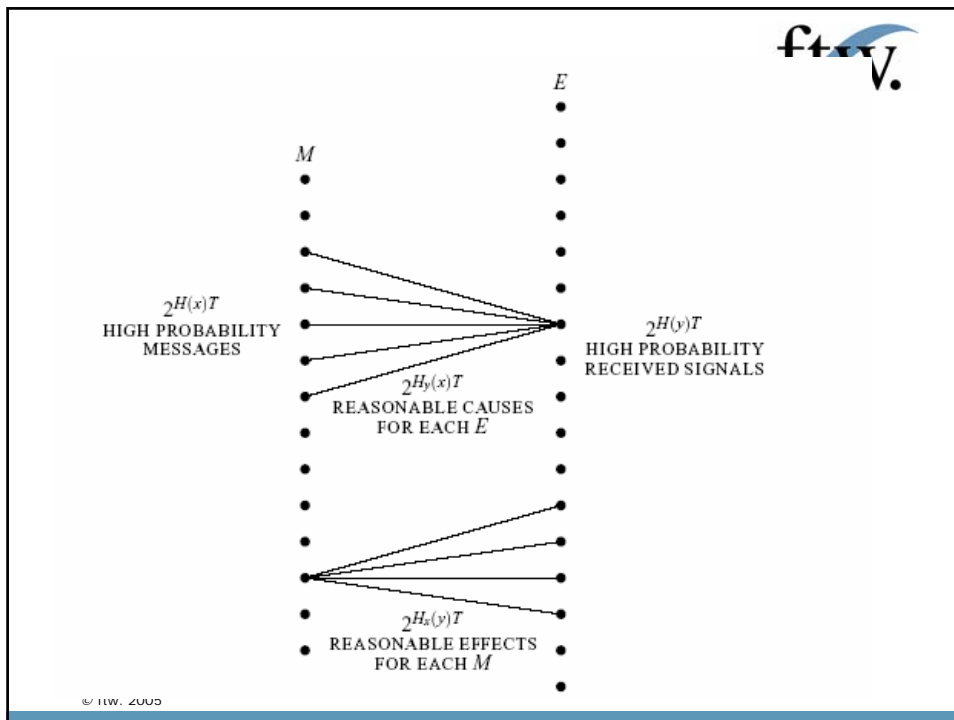
Coding Theorem



For any $\varepsilon > 0$ and any $R < C$, by choosing N sufficiently large, it is possible to design an (N, K) block code with $K/N \geq R$ such that the block error rate P_B for maximum likelihood decoding satisfies:

$$P_B < \varepsilon$$

© ftw. 2005



Bhattacharyya Bounding

ftw.

$$\underline{y} \in \mathcal{D}_i \Rightarrow P_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x}_i) \geq P_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x}_k)$$

$$\Rightarrow \sqrt{P_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x}_i)} \geq \sqrt{P_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x}_k)}$$

$$\Rightarrow P_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x}_k) \leq \sqrt{P_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x}_i)P_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x}_k)}$$

...

$$\Rightarrow P_{B|i} \leq \sum_{\substack{j=1 \\ j \neq i}}^M \sum_{\underline{y}} \sqrt{P_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x}_i)P_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x}_j)}$$

$$\Rightarrow P_{B|i} \leq \sum_{\substack{j=1 \\ j \neq i}}^M \prod_{n=1}^N \sum_{\underline{y}} \sqrt{P_{Y|X}(y|x_{in})P_{Y|X}(y|x_{jn})}$$

$M = 2^K$

© ftw. 2005

Gallager Bounding



$\underline{y} \notin \mathcal{D}_i \Rightarrow P_{Y|X}(\underline{y}|\underline{x}_j) \geq P_{Y|X}(\underline{y}|\underline{x}_i)$ for some $j \neq i$

$$\Rightarrow \left\{ \sum_{\substack{j=1 \\ j \neq i}}^M \left[\frac{P_{Y|X}(\underline{y}|\underline{x}_j)}{P_{Y|X}(\underline{y}|\underline{x}_i)} \right]^s \right\}^\rho \geq 1, \text{ for all } s \geq 0, \rho \geq 0$$

because at least one term in the sum will be at least 1.

We choose $s = 1/(1+\rho)$ and obtain:

$$P_{B|i} \leq \sum_{\underline{y}} P_{Y|X}(\underline{y}|\underline{x}_i)^{\frac{1}{1+\rho}} \left[\sum_{\substack{j=1 \\ j \neq i}}^M P_{Y|X}(\underline{y}|\underline{x}_j)^{\frac{1}{1+\rho}} \right]^\rho, \text{ all } \rho \geq 0$$

$$P_{B|i} \leq \prod_{n=1}^N \sum_y P_{Y|X}(y|x_{in})^{\frac{1}{1+\rho}} \left[\sum_{\substack{j=1 \\ j \neq i}}^M P_{Y|X}(y|x_{jn})^{\frac{1}{1+\rho}} \right]^\rho, \text{ all } \rho \geq 0$$

© ftw. 2005

Random Coding

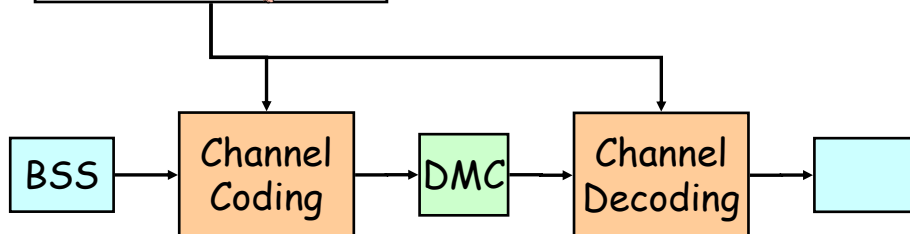


Random
Code
Selector



Chooses codewords according
to a distribution

$$Q(\underline{x}) = \prod_i Q(x_i)$$



What is the error rate **on average** over all choices of codes (good and bad...)??

© ftw. 2005

Random Coding Union / Bhattacharyya Bound



$$E[P_B] < 2^{-N(R_0 - R)}$$

where

$$R_0 = -\log_2 \left\{ \min_Q \sum_y \left[\sum_x \sqrt{P(y|x)Q(x)} \right]^2 \right\}.$$

is called the **cutoff rate**.

There exist codes whose error probability diminishes exponentially in N for rates R up to the cutoff rate R_0 .

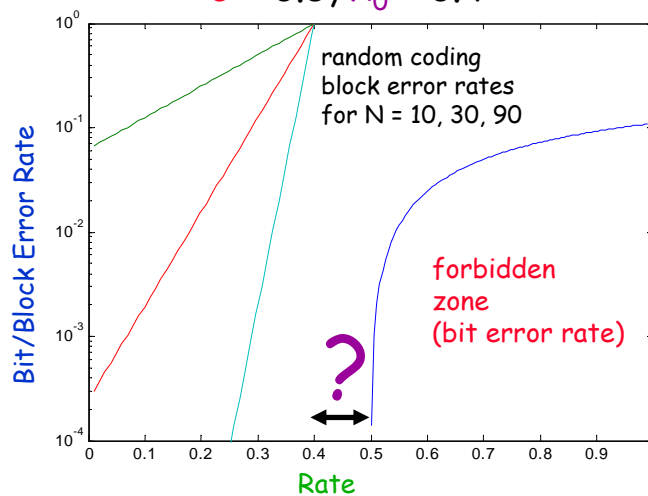
but what happens between R_0 and C ?

© ftw. 2005

Random Coding Union / Bhattacharyya Bound

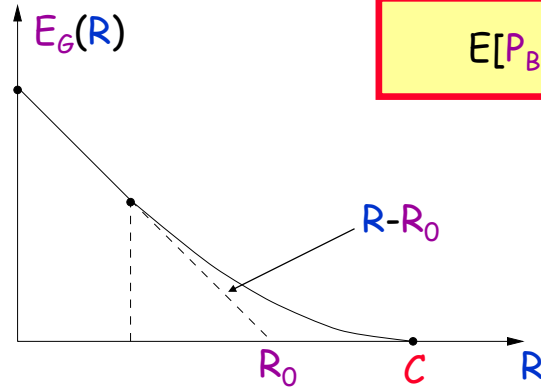


$$C = 0.5, R_0 = 0.4$$



© ftw. 2005

Random Coding Gallager Bound



There exist codes whose error probability diminishes exponentially in N for rates R up to the capacity C .

© ftw. 2005

Bit vs. Block Error Rate



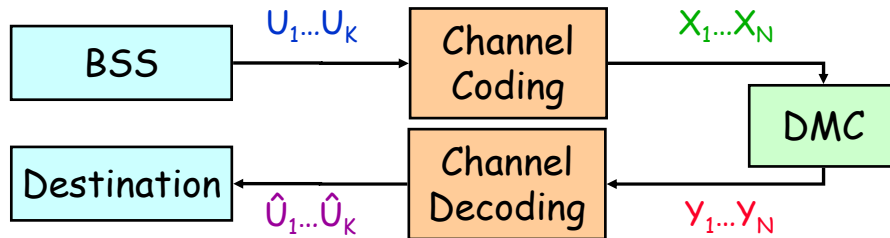
The bit error rate is always smaller or equal than the block error rate: every bit error results in a block error...

$$h^{-1}(1-C/R) \leq P_b \leq P_B \leq 2^{-N E_G(R)}$$

- the **converse to the coding theorem** also gives a lower bound on the **block error rate**
- the **coding theorem** also gives an upper bound for the **bit error rate (for block decoding!)**
- for **bitwise decoding**, the **bit error rate** can only be **better**, the **block error rate cannot** be better.

© ftw. 2005

Mutual Information



$$\lim_{N \rightarrow \infty} P_B = 0 \Rightarrow \lim_{N \rightarrow \infty} I(\underline{U}; \underline{\hat{U}}) = K$$

$$\Rightarrow \lim_{N \rightarrow \infty} I(\underline{U}; \underline{\hat{U}})/N = \lim_{N \rightarrow \infty} I(\underline{X}; \underline{Y})/N = R$$

$$\lim_{N \rightarrow \infty} P_b = 0 \Rightarrow \lim I(U_i; \hat{U}_i) = \lim I(U_i; Y)$$

$$= \lim I(X_i; Y)$$

$$= 1, \text{ for all } i$$

(for binary codes)

© ftw. 2005

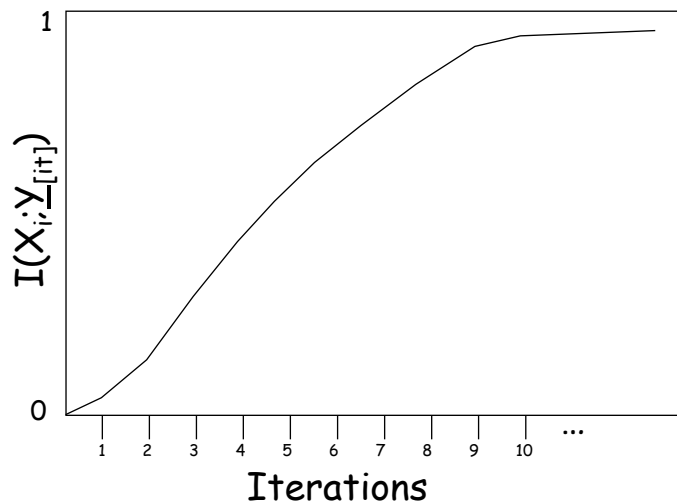
Iterative Decoding



- How does the mutual information evolve in an **iterative decoding** algorithm?
- We have learned that it is possible to **optimize** LDPC codes and RA codes so as to **maximize their threshold**
- We will see that we can design **capacity-achieving, iteratively decodable** families of LDPC & RA codes!! (i.e., threshold \rightarrow capacity)
- What is the implication in terms of **mutual information**?

© ftw. 2005

Mutual Information Trajectory



© ftw. 2005

Mutual Information Trajectory



- The L-values calculated in the tree are optimal in the sense of a **MAP-calculator**, i.e., $L(X_i | Y_{[i+1]})$ is a sufficient statistic for $Y_{[i+1]}$:
$$I(X_i; L(X_i | Y_{[i+1]})) = I(X_i; Y_{[i+1]})$$
- We can also draw the trajectory at **half-iterations** (after variable nodes & after check nodes)
- **But:** the output messages of variable nodes and check nodes are **extrinsic** L-values, whereas the mutual information trajectory we consider now is for **a-posteriori** L-values

© ftw. 2005

Next time: Extrinsic Information
Transfer Charts (Stephan ten Brink)



Photo by Jesse Squire

Stephan did his PhD at the U of Stuttgart, then worked for Bell Labs in the U.K., then in New Jersey. He is currently with RealTek. He is a regular visitor of ftw. and TU Wien.

(Stephan is the guy on the right, not the clown on the left)

© ftw. 2005