IT IN A NUTSHELL

Max H. M. Costa
Unicamp

Summary

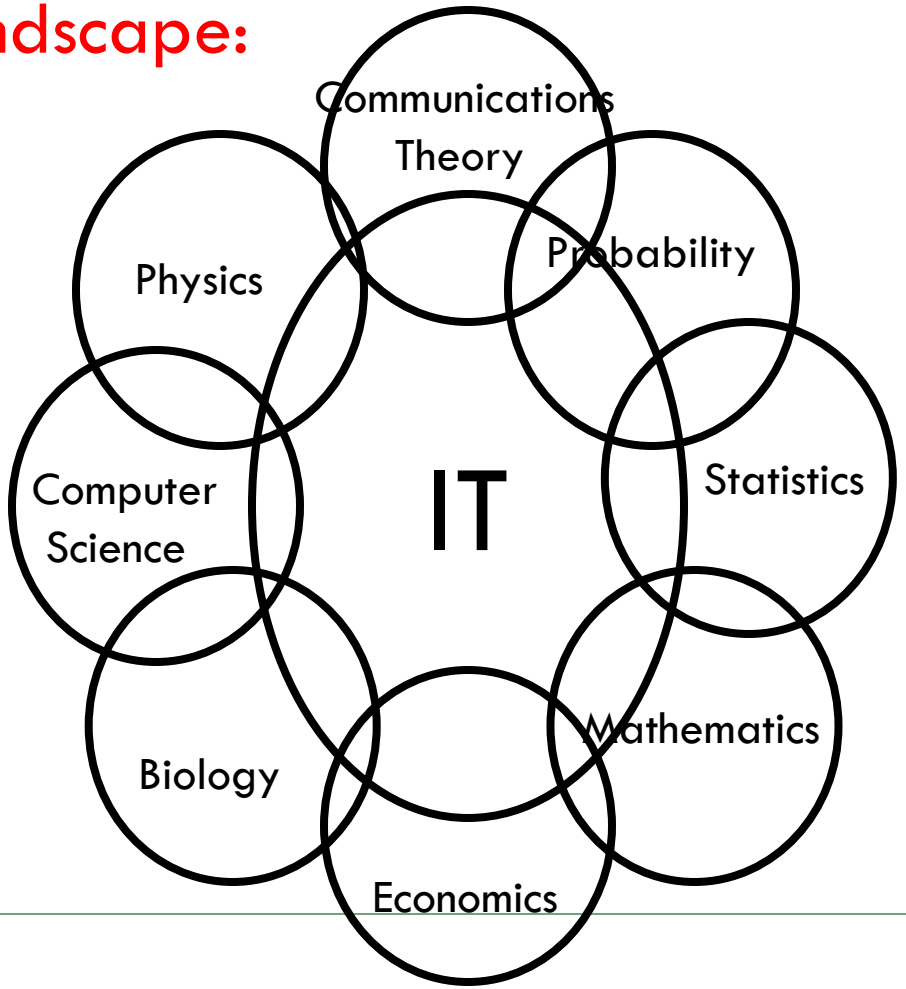
- Introduction
- Entropy, K-L Divergence, Mutual Information
- Data Compression (Source Coding)
- Transmission over Unreliable Channels (Channel Coding)
- Differential Entropy, Gaussian Channels
- Rate Distortion Theory
- Kolmogorov Complexity
- Applications in Biology, Economics
- Multiple User Information Theory
- Wrap up

Some References (Textbooks):

- [1] T. Cover and J. Thomas, Elements of Information Theory, Wiley, 2nd ed., 2006 (1991).
- [2] R. Ash, Information Theory, Dover, 1990.
- [3] R. Gallager, Information Theory and Reliable Communication, Wiley, 1968.
- [4] A. El Gamal and Y-H. Kim, Network Information Theory, Cambridge, 2011.

Information Theory and other Areas

- The IT landscape:



Entropy

- Definition: H(X) = The entropy of X
- Let X be a discrete random variable taking values in $\{x_1, x_2, ..., x_M\}$ with probabilities $\mathbf{p} = \{p_1, p_2, ..., p_M\}$.
- $H(X) = H(\mathbf{p}) = \sum_{k=1}^{M} p(x_k)\, log_2 \frac{1}{p(x_k)}$ (bits) =
- $= E\left(log_2 \frac{1}{p(x_k)} \right)$ bits

H(X) is a measure of the uncertainty of X.

How can H(X) arise naturally?

- Let X_1, X_2, … be independent and identically distributed (i.i.d.) according to p(x).
- Then
- $p(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2) \ldots p(x_n) = \prod_{i=1}^{n} p(x_i) =$

$$= 2^{\sum_{k=1}^{n} log_2 p(x_i)} = 2^{n \sum_{j=1}^{m} \frac{\nu(j)}{n} log_2 p(x_j)}$$

$$\rightarrow 2^{-nH(X)}$$

Asymptotic Equipartition Property

Change of base

- $H_b(X) = E\left(log_b \frac{1}{p(x_k)} \right) = log_b a \quad H_a(X)$

- Units of Entropy:
  - Base 2 → bits
  - Base 10 → dits or Hartleys
  - Base e → nats
  - Base 3 → trits (why not?)

Examples

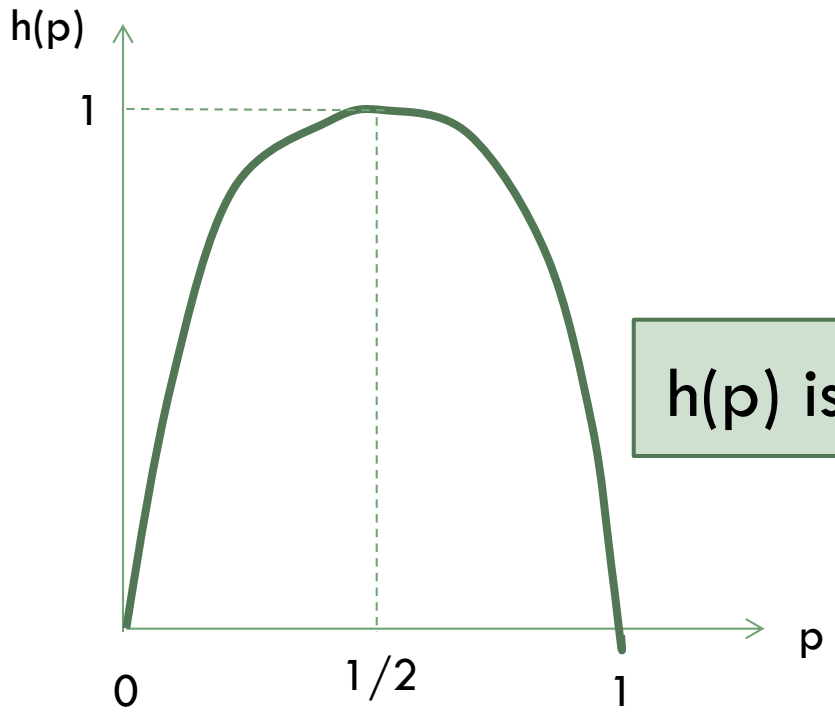
- Ex. 1) $X \in \{0,1\}$, $p(X=0)=0$, $p(X=1)=1$
- $H(X) = -\,0 \log 0 - 1 \log 1 = 0$
- Note: $\lim_{P \to 0} p \log p = 0$ by l'Hôpital's rule

No uncertainty !

X is deterministic

Examples (continued)

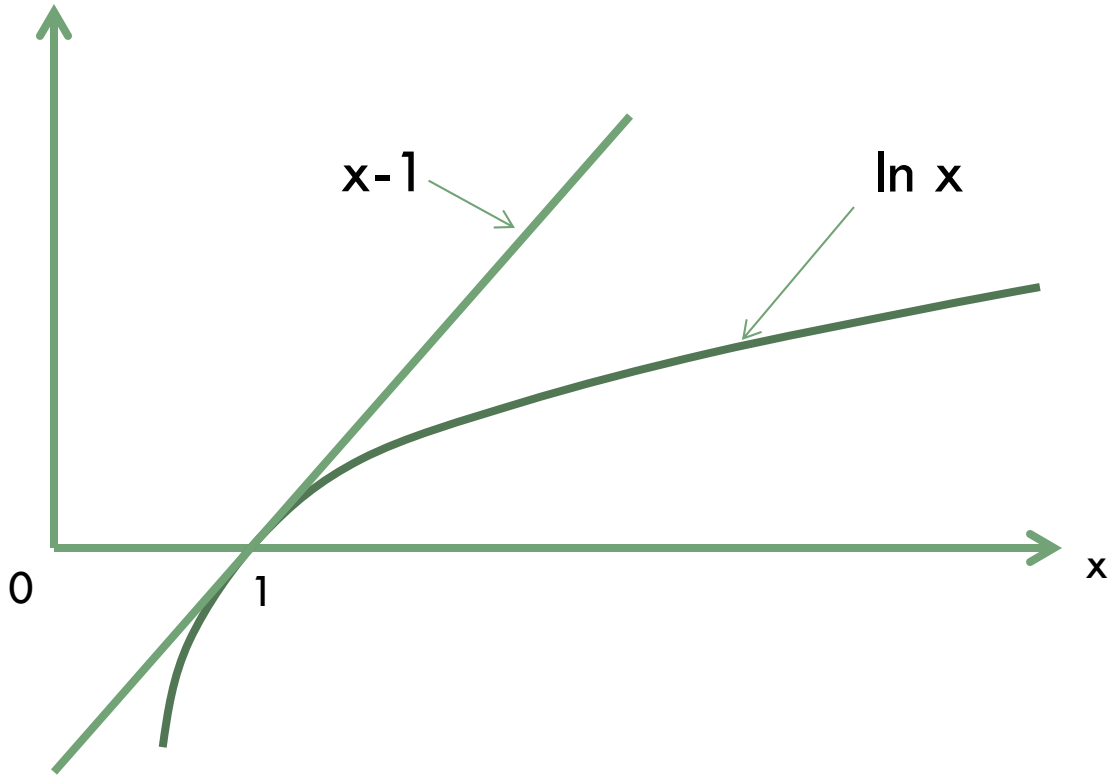
- Ex. 2) $X \in \{0,1\}$, $p(X=0)=p$, $p(X=1)=1-p$,
- $H(X) = -p \log p - (1-p) \log (1-p)$
- $= h(p)$



h(p) is the binary entropy function

Lemma

- $\ln x \leq x-1, \qquad x>0$
- Proof: Taylor series with remainder



Relative Entropy (Kulbach-Leibler divergence)

- Let p(x) and q(x) be two probability mass functions defined on alphabet $\mathcal{X}$.
- *The K-L divergence of p w.r.t. q is*
- $$D(p \| q) = \sum_{x} p(x) \log \frac{p(x)}{q(x)}$$

Proposition: Information Inequality

$D(p \| q) \geq 0$ with equality if and only if (iff) $p \equiv q$

- Proof: Let $A = \{x : p(x) > 0\}$
- Have $\ln x \leq x-1$ (Lemma)
- Thus $\ln \frac{q(x)}{p(x)} \leq \frac{q(x)}{p(x)} - 1$
- Multiply by p(x) and sum over $x \in A$
- $$\sum p(x) \log \frac{q(x)}{p(x)} \leq \sum p(x) \left(\frac{q(x)}{p(x)} - 1\right) \leq 0$$
- $-D(p \| q) \leq 0 \implies D(p \| q) \geq 0$
- equality iff $p = q$ in A, then $p \equiv q$. QED

Remark

- The K-L Divergence is very useful in IT,
- but it is not a metric.

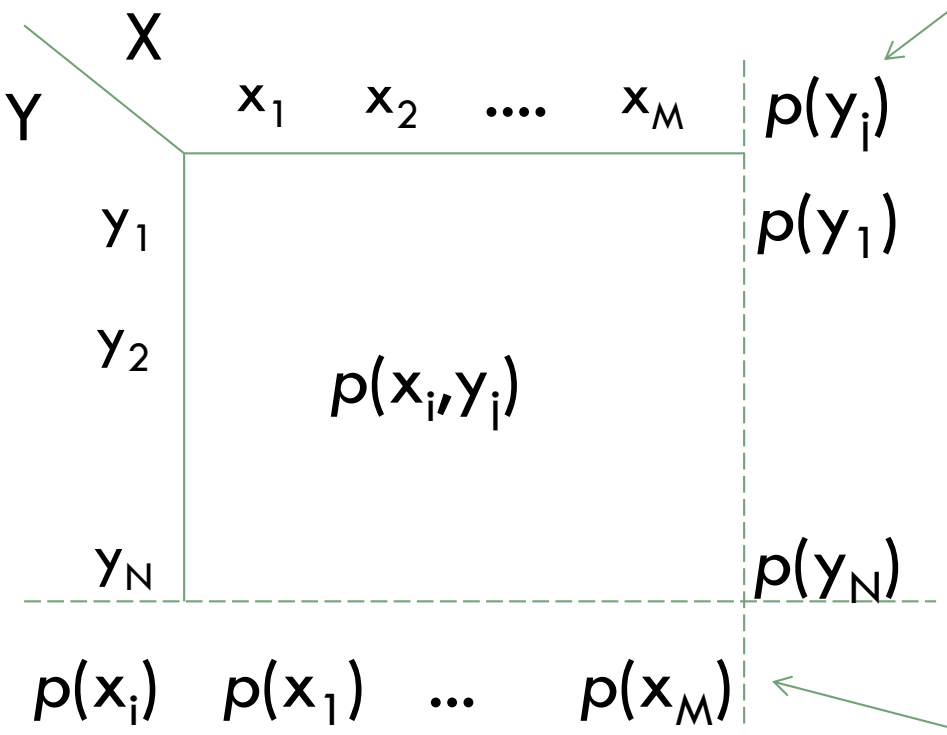
- It is not symmetric and does not satisfy the triangle inequality

Application

- Let q be the uniform distribution
- $q_i = 1/n$ for i=1,...,n
- $p = \{p_1, p_2, \ldots, p_n\}$
- Then $D(p \| q) \geq 0$
- $\sum p_i \log \frac{p_i}{q_i} \geq 0$
- $\sum p_i \log p_i \geq \sum p_i \log q_i = \sum p_i \log 1/n$
- Thus $H(p) \leq \log n$
- The uniform distribution has maximum entropy.

Joint, marginal and conditional distributions

- Joint Distribution:



Marginal distributions:

$$p(xi) = \sum_{y_j} p(x_i, yj)$$

$$p(y_j) = \sum_{x_i} p(x_i, yj)$$

Conditional Distributions:

- $$p(y_j|x_i) = \frac{p(xi,yj)}{p(xi)}$$

- $$p(x_i|y_j) = \frac{p(xi,yj)}{p(y_j)}$$

The joint distribution determines the marginal and the conditional distributions.

The opposite is not true.

Joint Entropy

- $H(X,Y) = H(\mathbf{p}(x,y)) = \sum_{(x_i, yj)} p(x_i, yj)\ log \frac{1}{p(x_i,yj)}$

Conditional Entropy

- $H(X|Y) = \sum p(x_i, yj)\ log \frac{1}{p(x_i|y_j)} = E\ (log\ p(x_i|yj)\ )$

- $H(Y|X) = \sum p(y_j, x_i)\ log \frac{1}{p(y_j|x_i)} = E\ (log\ p(y_j|xi)\ )$

Chain Rule (like peeling an onion):

$$H(X,Y) = H(X) + H(Y|X)$$

- $= H(Y) + H(X|Y)$
- Proof: Do it for homework.
- Simple algebraic manipulation.

- Corollary (conditional form):
- $H(X,Y|Z) = H(X|Z) + H(Y|X,Z)$
- $= H(Y|Z) + H(X|Y,Z)$

Mutual Information

- The Mutual information between X and Y is
- the K-L divergence of the joint distribution p(x,y) and the product of the marginals p(x) p(y).
- I(X;Y) = D(p(x,y) || p(x) p(y))
- $$= \sum_{x} \sum_{y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Properties of I(X;Y)

- 1) Non-negativity: $I(X;Y) \geq 0$, with equality
- iff X and Y are independent.
- Proof: I(X;Y) is a K-L divergence.

- 2) Symmetry:
- $I(X;Y) = I(Y;X)$
- Proof: Trivial ($p(x)p(y) = p(y)p(x)$)

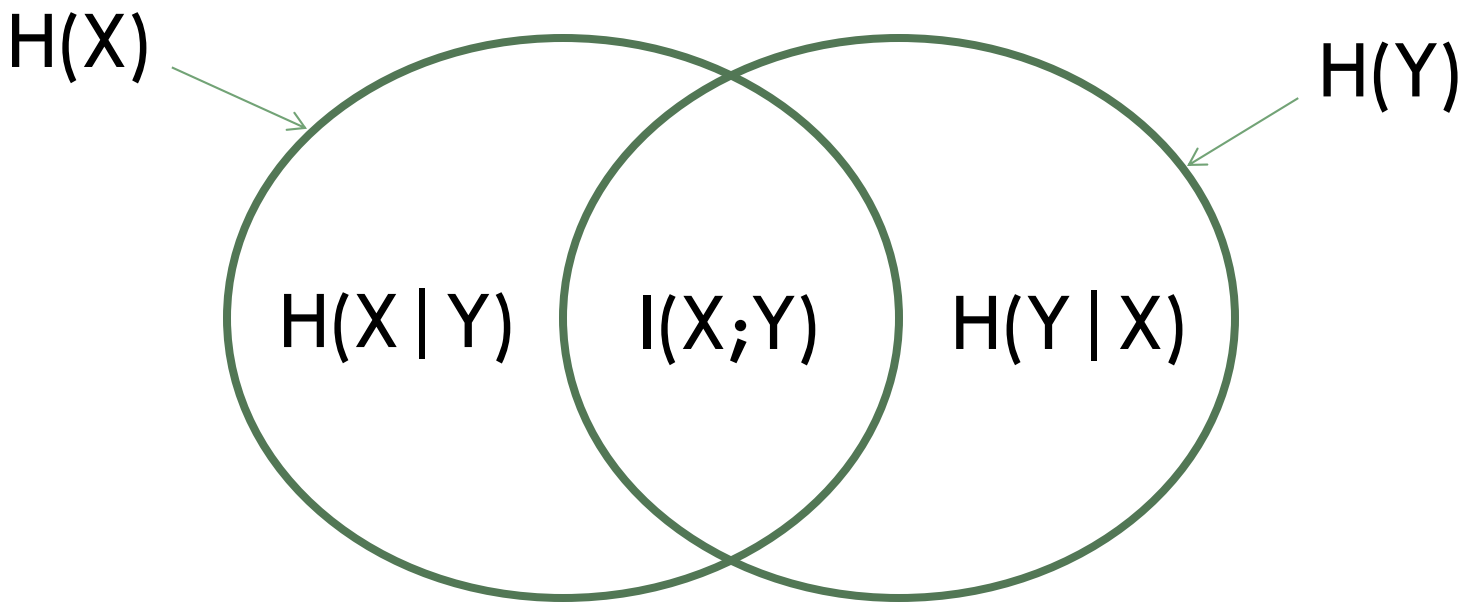
Mutual Information and Entropy

- $I(X;Y) = \sum \sum p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$
- $= H(X) + H(Y) - H(X,Y)$ (from above)
- $= H(X) - H(X \mid Y)$ (from chain rule)
- $= H(Y) - H(Y \mid X)$ (alternative form)

- Note: The Mutual Information between two random variables is the residual uncertainty about one r.v. after the other is revealed.

A Venn Diagram

- Works well for two random variables



Information can't hurt

- Conditioning reduces entropy:

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

- Proof: $I(X;Y) = H(X) - H(X|Y) \geq 0$

- On average the knowledge of Y cannot increase the uncertainty about X.

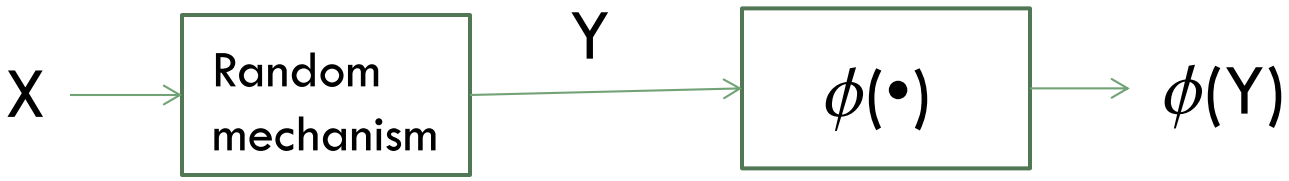
Entropy as self-information

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

- The entropy is the amount of information that a random variable conveys about itself.

Passing on Information

- Let X and Y be dependent r.v.'s



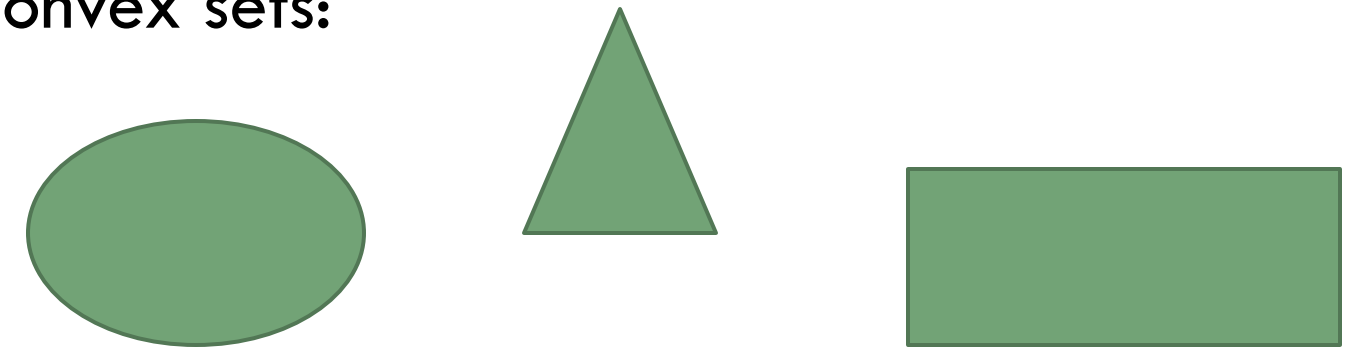
- Proposition: $I(X;Y) \geq I(X; \phi(Y))$
- Proof: $I(X;Y) = H(X) - H(X|Y)$
- $= H(X) - H(X|Y, \phi(Y))$
- $\geq H(X) - H(X|\phi(Y)) = I(X; \phi(Y))$

Conditioning reduces entropy

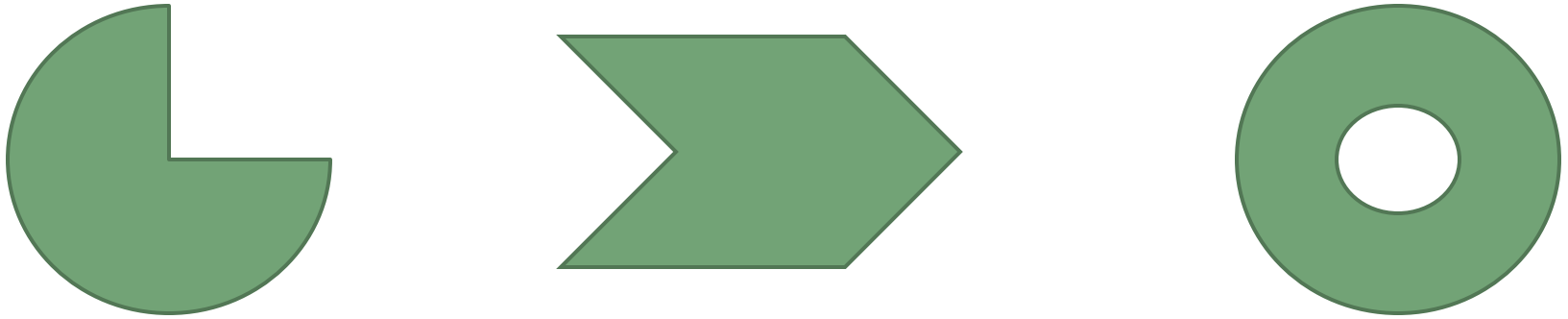
- This is a simple form of the Data Processing Inequality.

Convexity – quick review

- Convex sets:

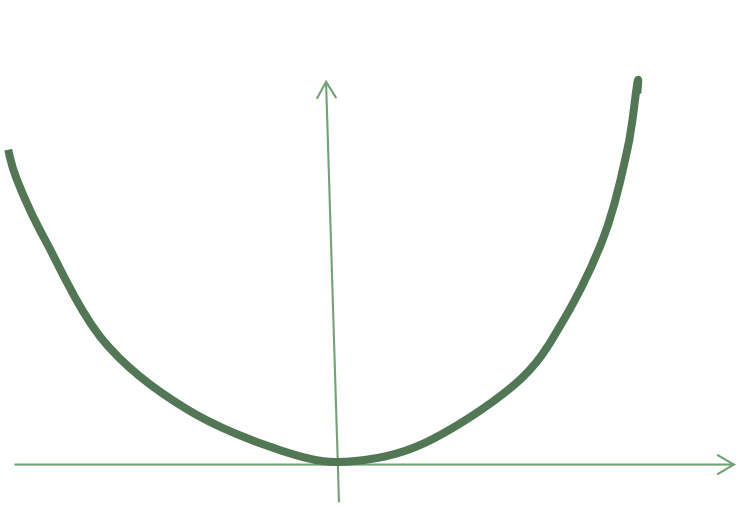


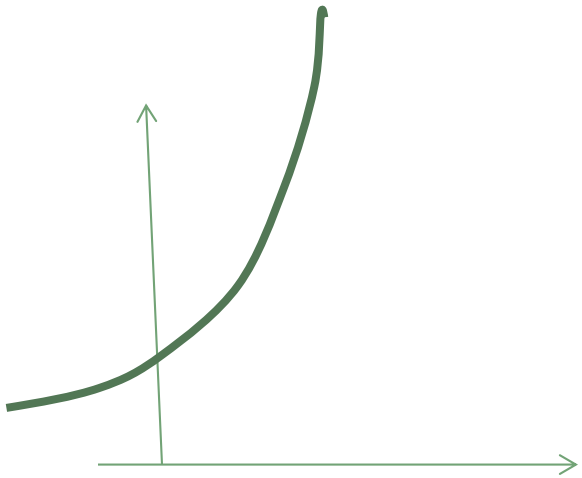
- Non-convex sets:



Convex Functions:

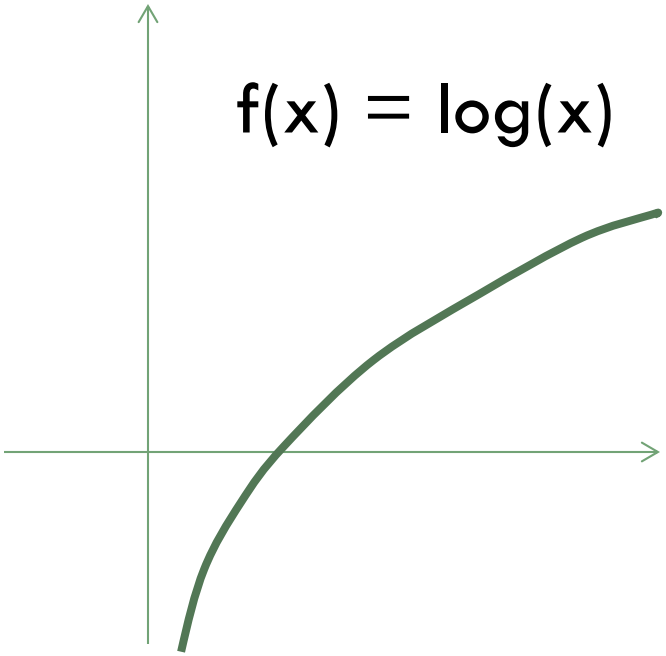
- A function f(x) is convex if the set of points above its graph is convex.
- Examples: $f(x) = x^2$ $f(x) = e^x$

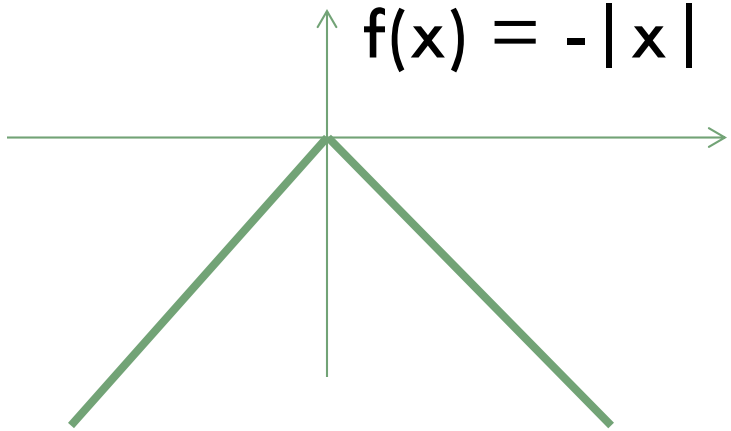




Mnemonic: The exponential function is conve^x

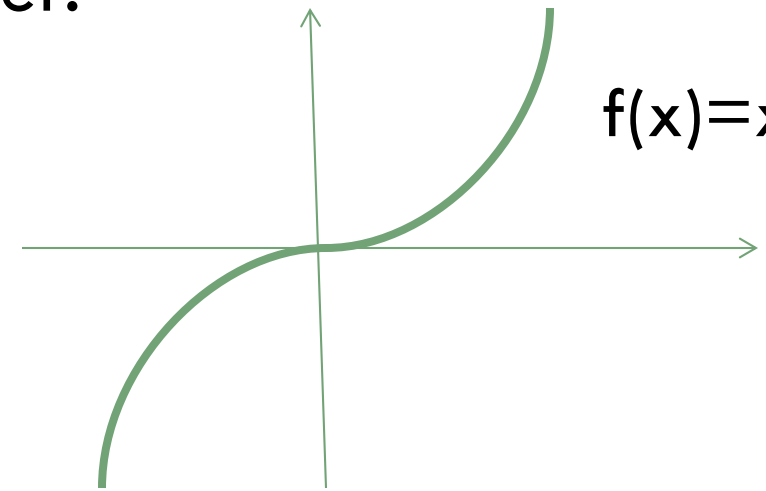
Concave functions

- f(x) is concave if {- f(x)} is convex.
- Examples:
- 



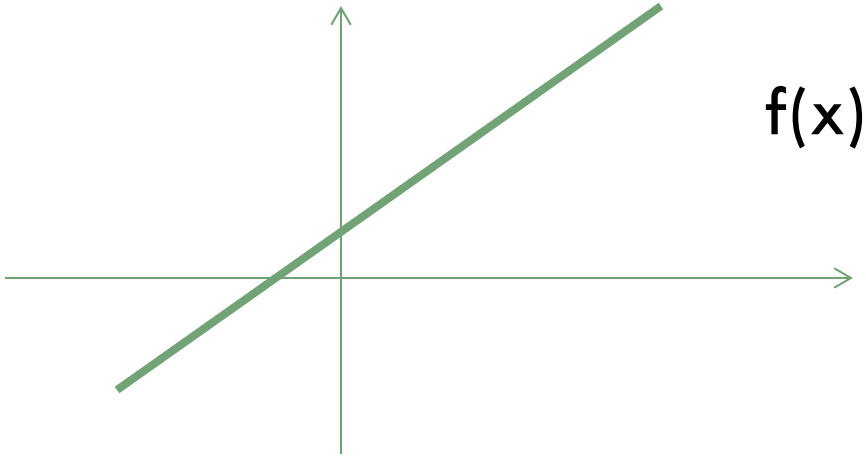
Some are neither, some are both

- Neither:



$f(x)=x^3$

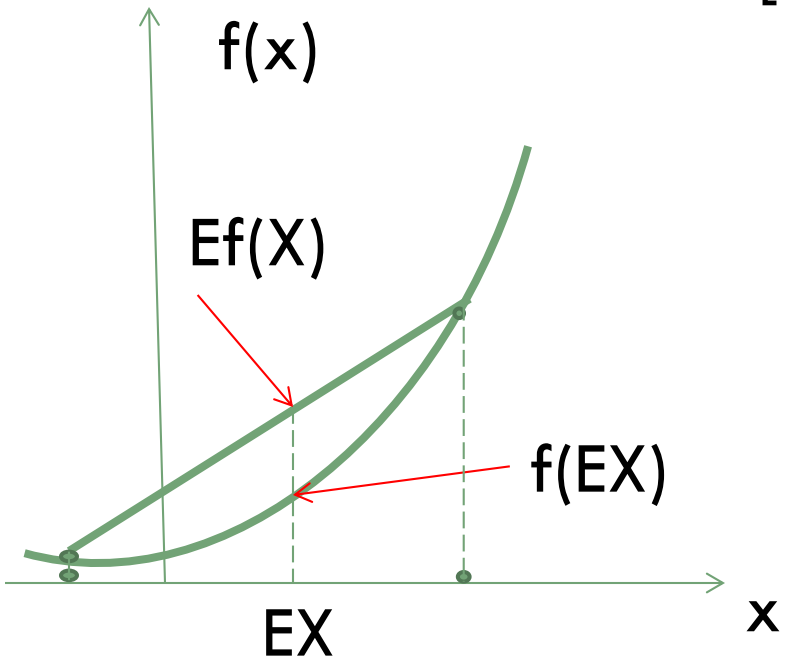
- Both:



$f(x)=ax+b$

Jensen's Inequality

- Let X be a random variable and f(x) a convex function.
- Then $E[f(X)] \geq f(EX)$



Proofs: Induction,
Taylor series (when f´´(x) exists).

Mnemonic: The chord is above the arc.

Concavity of H(p)

- Proposition: H(p) is a concave function of p.
- Proof: Let X_1 be distributed as p_1 and X_2 as p_2.
- Let index $\theta \in \{1,2\}$ with probabilities $(\lambda, 1-\lambda)$
- Let $Z = X_\theta$. Then Z is distributed as $\lambda\ p_1 + (1-\lambda)\ p_2$.
- Now since conditioning reduces uncertainty
- $H(Z) \leq H(Z|\theta)$. Equivalently
- $H(\lambda\ p_1 + (1-\lambda)\ p_2) \geq \lambda\ H(p_1) + (1-\lambda)\ H(p_2)$
- showing that h(•) is a concave function.
- Note: Mixing two gases of equal entropy results in a gas with higher entropy.

Additional topics:

- Log-Sum Inequality

- Convexity of D(p || q) in the pair (p,q)

- I(X;Y) as a function of $p(x,y) = p(x)\, p(y|x)$
  - is a concave function of p(x) and
  - a convex function of p(y|x).

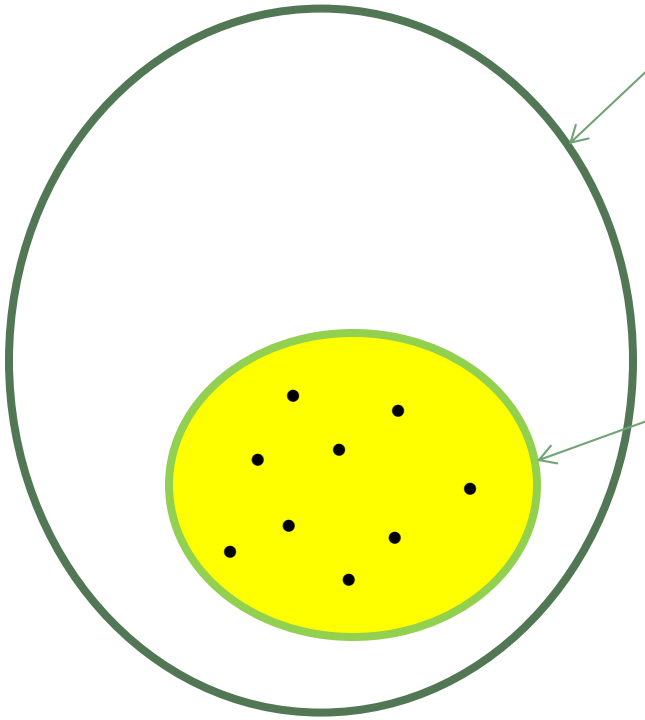
Additional topics (continued):

Markov Chains

- Data Processing Inequality
- Sufficient Statistics
- Fano´s Inequality

Asymptotic Equipartition Property

- Let $X_1, X_2, \ldots, X_n$ be i.i.d. according to $p(x)$



Sample space = set of all sequences $(x_1, x_2, \ldots, x_n)$

A_ϵ=Set of typical sequences

A.E.P.

$$\begin{cases} 1)\ \Pr\{A_\epsilon\} \geq 1-\epsilon \\ 2)\ p(x) \cong 2^{-nH(X)} \\ 3)\ \|A_\epsilon\| \cong 2^{nH(X)} \end{cases}$$

This is the DNA of IT !

An example of typical sequences

- Let X be a biased coin with
- P(Head)=0.9 and P(Tail) = 0.1

- Consider the set of 1000-long sequences of coin tosses.
- Typical sequences are those that have approximately 900 Heads and 100 Tails.

- Note: The most likely sequence, namely the one
- with 1000 Heads, is not Typical !

Conclusion

- Better bet on A_ϵ !

Entropy Rate of Random Processes

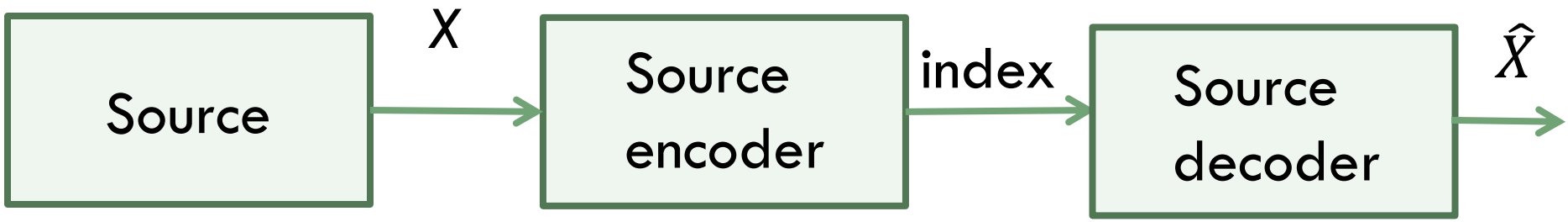
- The concept of Entropy can be extended to Random Processes:
- Let $X = \{X_i, i=1,2,\ldots,n\}$ be a collection of random variables forming a stationary process.
- Entropy Rate:
- $H(X) = \lim_{N\to\infty} H(X_n \mid X_{n-1}, X_{n-2}, \ldots, X_1)$ bits/symbol
- $= \lim_{N\to\infty} \frac{H(X_1, X_2, \ldots, X_n)}{n}$ bits/symbol
- For stationary X these limits exist and are equal.

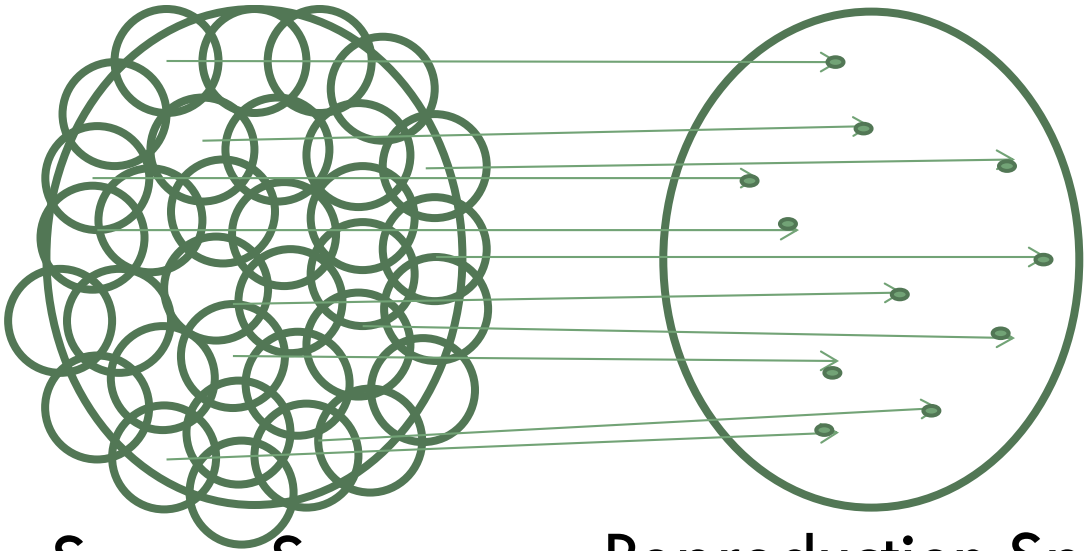
Exercise:

- Coin weighing:
- Suppose that we have 12 coins, among which there may or may not be one counterfeit coin. If there is one counterfeit coin it will be heavier or lighter than the legitimate coins. The coins are weighted by a two plate balance. What is a weighing strategy to determine to identify the odd coin, if there is one, with 3 weighings. Note: This problem can be solved with a ternary Hamming code.

Data Compression (Source Coding)

- Want to represent a source efficiently.





Source Code for a r.v. X

- A Source Code for X is a mapping from the alphabet of X to a finite sequence of a D-ary code alphabet.
- Examples:
- Source alphabet = $\mathcal{X}$ = {a, b, c, d, e}
- Code alphabet = $\mathcal{D}$ = {0, 1}
- Code example:
- C(a)=00, C(b)=01, C(c)=10, C(d)=11, C(e)=001

Note: This code is not instantaneous

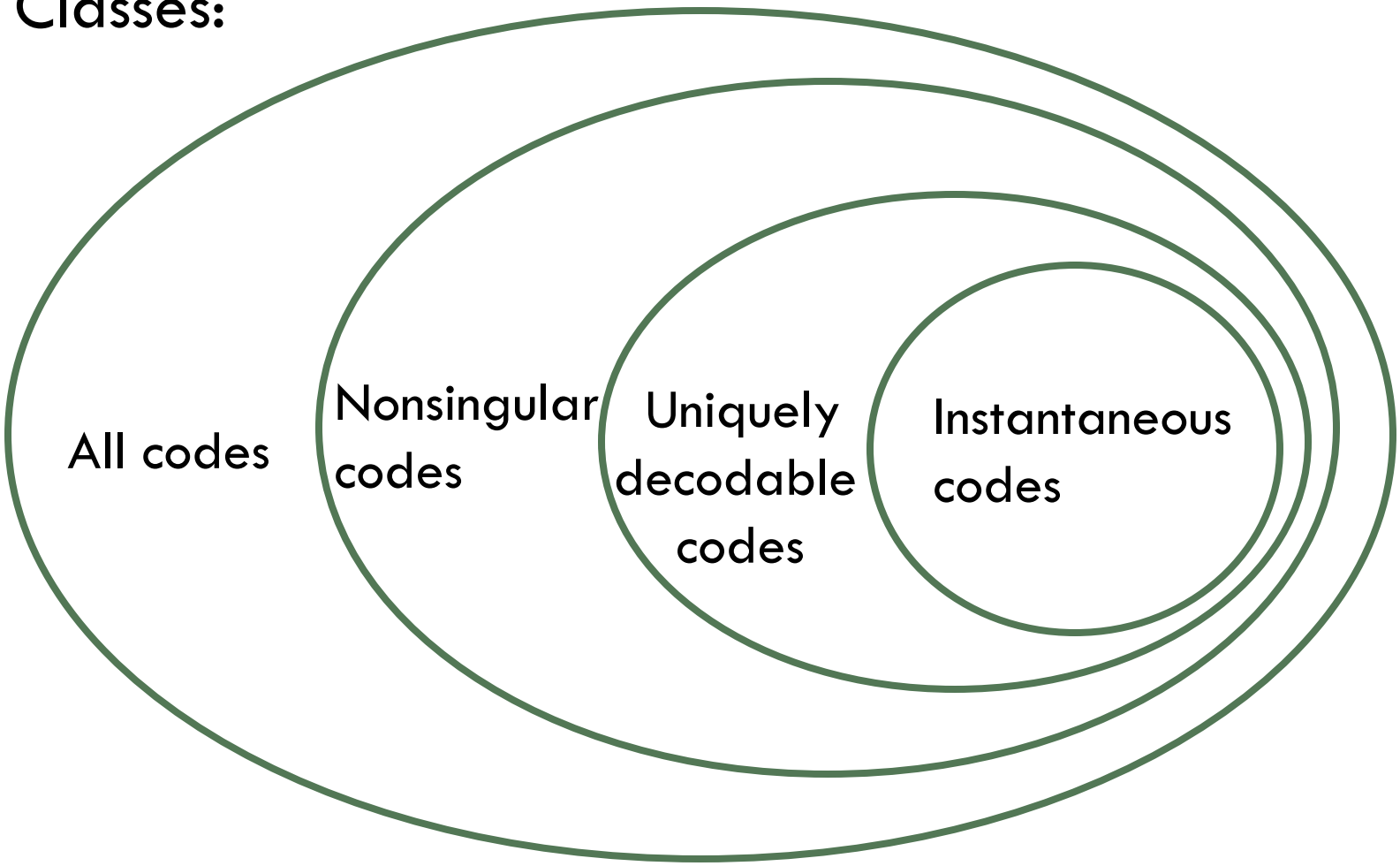
Instantaneous codes (Prefix-free)

- A code is a prefix-free code or a instantaneous code if no codeword is a prefix of another codeword.

- An instantaneous code is a self-punctuating code.
- We can add commas without looking ahead.

Classification of Codes

- Classes:



Classes of codes

Binary code alphabet

Source Alphabet $\mathcal{X}$	Singular	Nonsingular but not uniquely decodable	Uniquely Decodable but not Instantaneous	Instantaneous
A	0	0	10	0
B	1	010	00	10
C	0	01	11	110
D	1	10	101	111

Want small average code length per source symbol.
That depends on the source distribution.

Examples of codes

Binary code alphabet

X	P(X)	Code 1	Code 2 (a comma code)	Code 3
A	1/2	00	1	0
B	1/4	01	01	10
C	1/8	10	001	110
D	1/8	11	0001	111

Code 1: Average length = 2 bits per source symbol

Code 2: Average length = 1.875 bits/source symbol

Code 3: Average length = 1.75 bits/source symbol = H(X)

Kraft Inequality

- For any instantaneous code over a code alphabet of size D, the codeword lengths $\ell_1, \ell_2, \ldots, \ell_m$ must satisfy the inequality

$$\sum_i D^{-\ell_i} \leq 1$$

- Example for Code 3:
- $\ell_1 = 1, \ell_2 = 2, \ell_3 = 3, \ell_4 = 3$
- $\sum_i D^{-\ell_i} = 2^{-1} + 2^{-2} + 2^{-3} + 2^{-3} = 1$ → ok

Length Bounds for Optimal Codes

- First Shannon Theorem: The expected length L of any instantaneous D-ary code for a r.v. X is bounded below by the entropy $H_D(X)$:

$$L \geq H_D(X)$$

Proof: Expanding $L - H_D(X) = \sum p_i \ell_i - \sum p_i \log_D \frac{1}{p_i}$ have

$L - H_D(X) = D(p \,||\, r) + \log_D \frac{1}{c}$, where $r_i = \frac{D^{-\ell_i}}{c}$ and

$c = \sum D^{-\ell_i}$.

Thus $L - H_D(X) \geq 0$ with equality if $p \equiv r$ and $c = 1$,

i.e., the code meets Kraft inequality with equality.

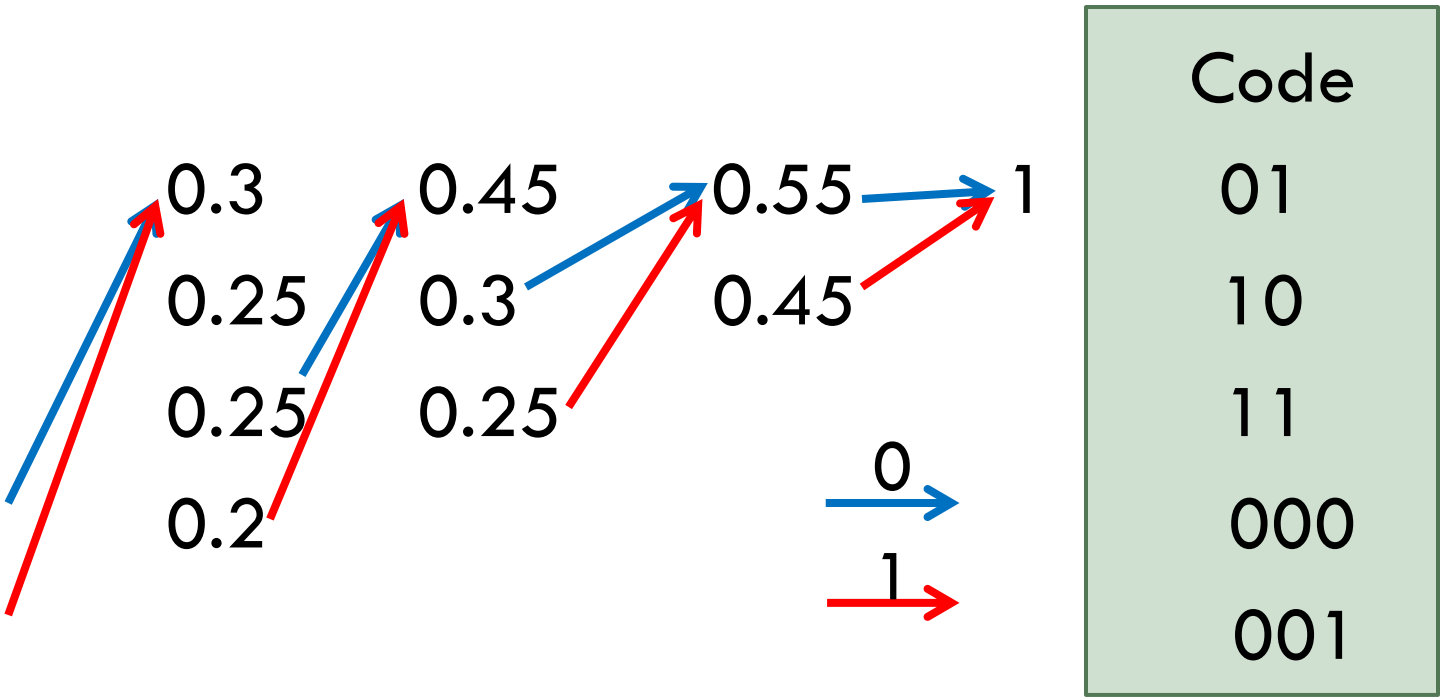
Efficient Codes: Shannon Code

- Choose $\ell_i = \left\lceil log_D \frac{1}{p_i} \right\rceil$,
- use a suitable codeword with this length.

- This code satisfies
- $$H_D(X) \leq L \leq H_D(X) + 1$$

Efficient Codes: Huffmann Codes (1952)

- The Huffman Code is the optimal prefix code (shortest expected length) for a given source distribution p(x).
- Example:



This code has average length 2.3 bits/source symbol.

Efficient Codes:

Other efficient codes:

- Shannon-Fano-Elias Codes
- Arithmetic Codes
- Lempel-Ziv Codes (Universal – learns source distribution)
- Run-length codes + Golomb codes (very simple code)

Run-length + Golomb Codes

- Run-length codes used to encode long binary sequences where 0's are (much) more likely than 1's (or vice versa) .
- Initial step: Run length code
- Represent the zero runs as integers.
- Example:
- Input = 0000010000001000100000000001001…
- Runs: 5 6 3 11 2 … Now use Golomb code.

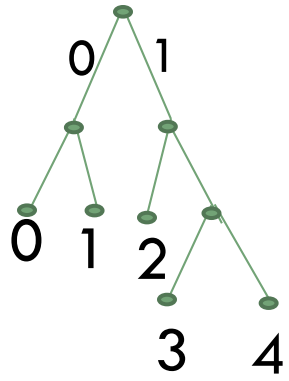
Golomb code of Order m

- Variable length code for integers
- Choose suitable order m.
- Represent each integer as n = qm+r
- (integer divide n by m to get quocient q and remainder r)
- Use unary code for q (sequence of q 1's), use comma = 0;
- Use prefix code (with m leaves) for r

Example: m=5

Code for r:

rϵ{0,1,2,3,4}



r	output
0	00
1	01
2	10
3	110
4	111

Golomb code (m=5)

- Example: Runs = {5 6 3 11 2}
- Sequence of (q,r)'s = {(1,0) (1,1) (0,3) (2,1) (0,2)}

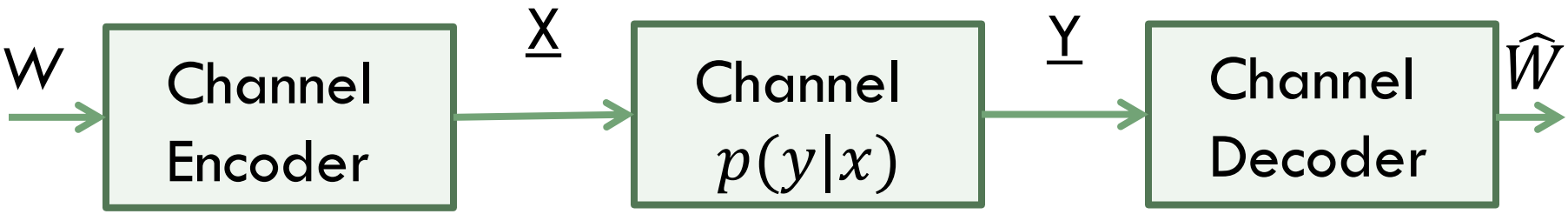
- Code sequence = {(1-0-00)(1-0-01) (0-110) (11-0-01)
- (0-10)}
- Input length = 32
- Output length = 20
- Compression ratio = 20/32 = 0.625
- Typically can get excess rates of less than 1%.

Selecting the order m

- Approximate solution: Find m such that
- $$pm = \frac{1}{2}, \quad \text{where p} = \Pr\{X=0\}.$$
- Thus m $\cong \left\lceil \frac{-1}{log_2 p} \right\rceil$.
- Remarks: Very simple implementation
- Need good tuning of order parameter m
- Can be made adaptive
- Variable to variable length code
- Terminate input sequence with a 1

Transmission over Unreliable Channels

- The Channel Coding Problem:



- $W \in \{1,2,\ldots,2^{nR}\}$ = message set of rate R
- $\underline{X} = (x_1\ x_2\ \ldots\ x_n)$ = codeword input to channel
- $\underline{Y} = (y_1\ y_2\ \ldots\ y_n)$ = codeword output from channel
- $\widehat{W}$ = decoded message P(error) = $P\{W \neq \widehat{W}\}$

Simple examples

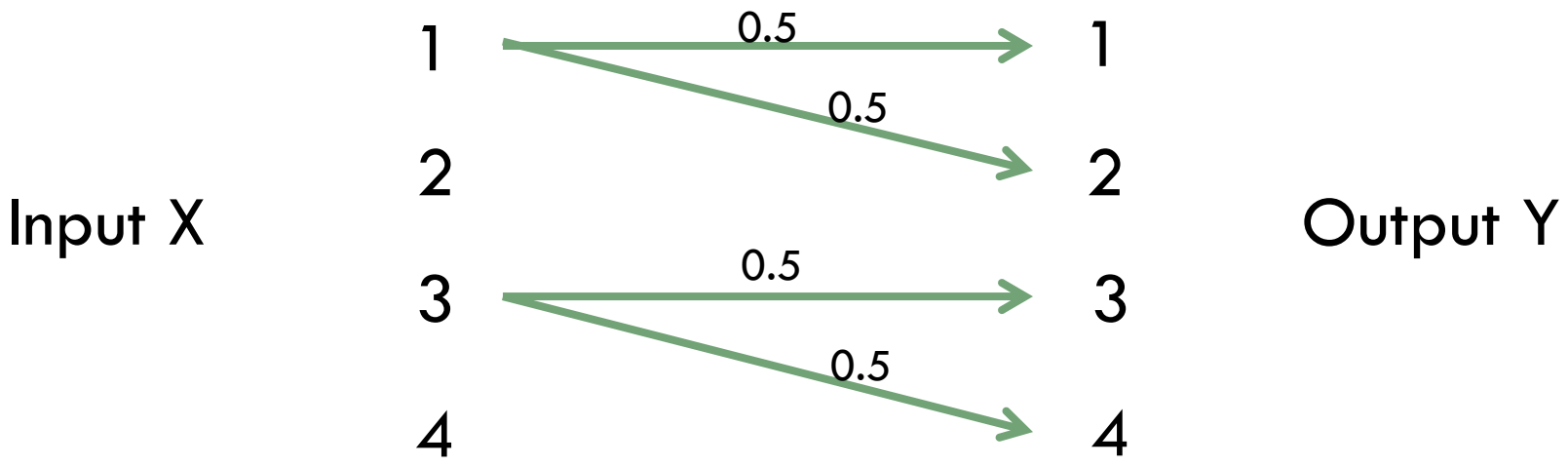
- Noiseless typewriter:

Number of noise free symbols = 4

Can transmit R = $log_2\ 4 = 2$ bits/transmission

Simple examples

- Noisy typewriter (type 1):

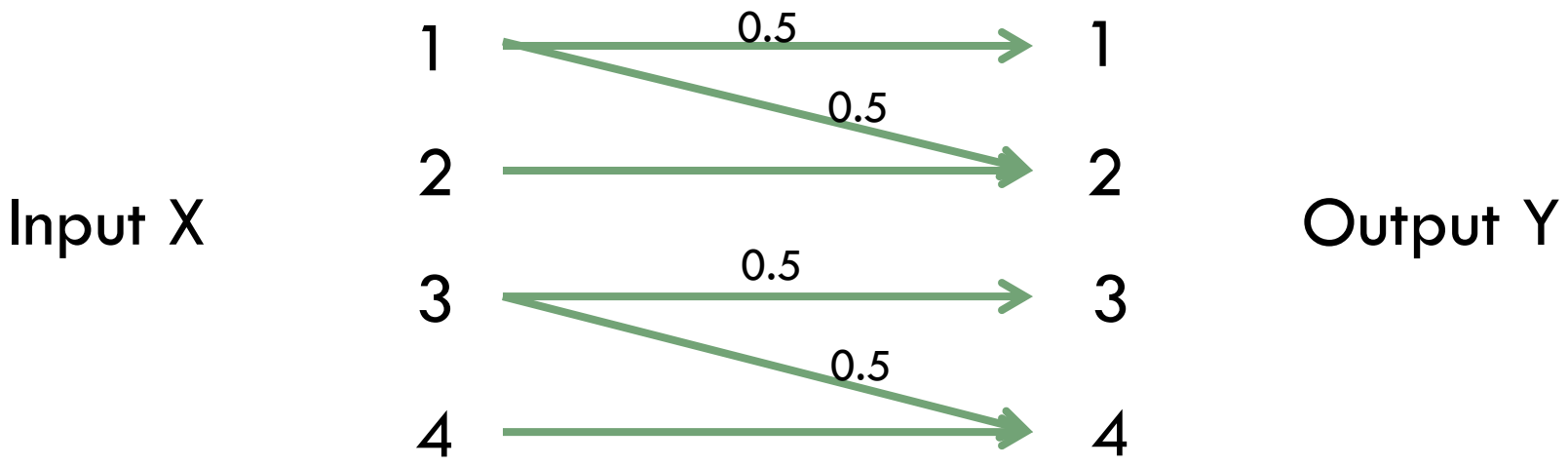


Number of noise free symbols = 2

Can transmit R = $log_2\ 2 = 1$ bit/transmission

Simple examples

- Noisy typewriter (type 2):

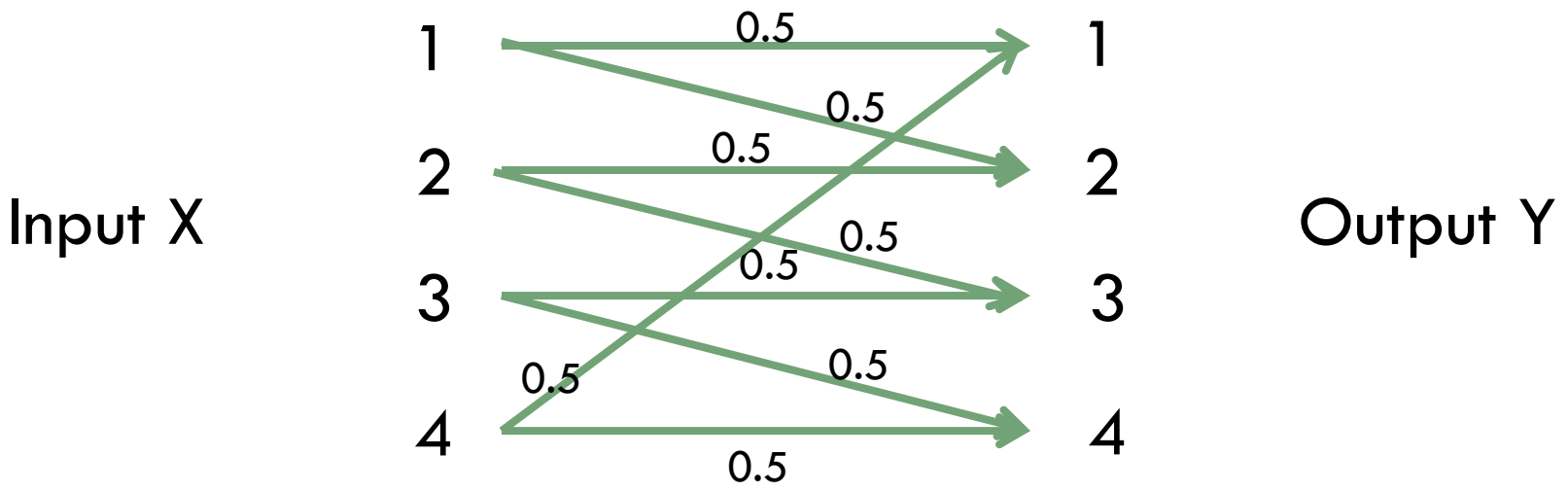


Number of noise free symbols = 2

Can transmit R = $log_2\ 2 = 1$ bit/transmission

Simple examples

- Noisy typewriter (type 3):

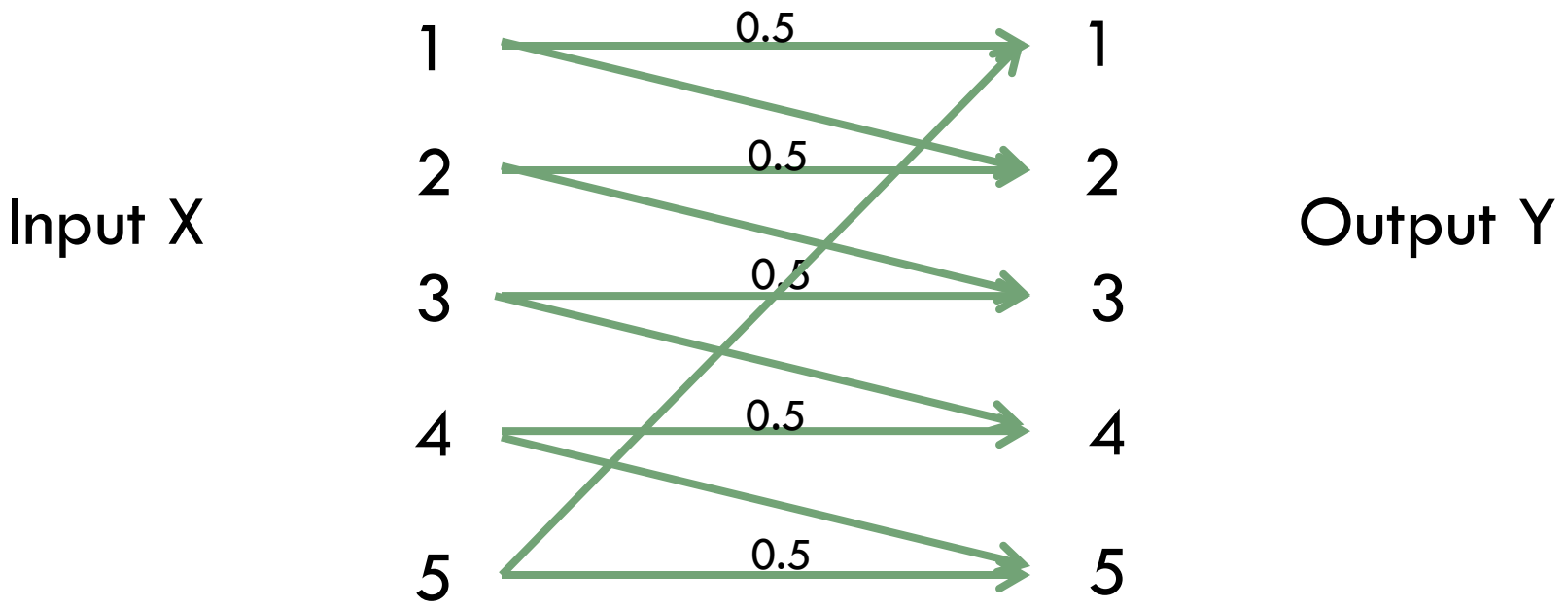


Number of noise free symbols = 2 Use X=1 and X=3

Can transmit R = log_2 2 = 1 bit/transmission

Simple examples

- A tricky typewriter:

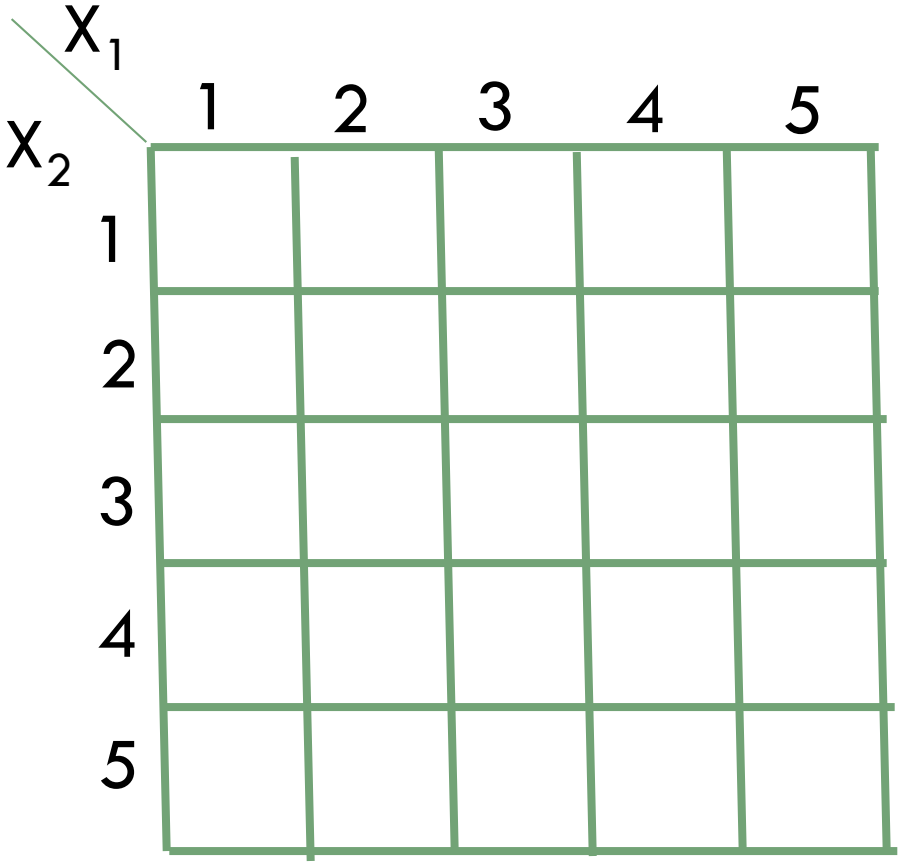


How many noise free symbols?

Clearly at least 2, hopefully more.

Simple examples

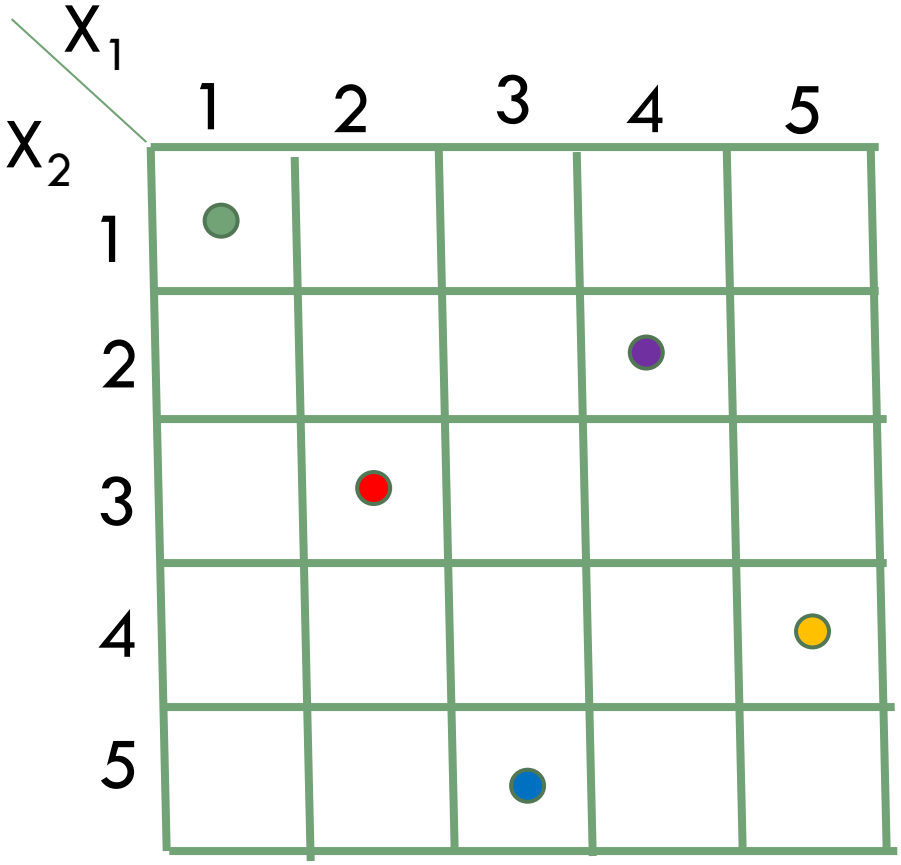
- Consider the n=2 extension of the channel:



Which code squares to pick?

Simple examples

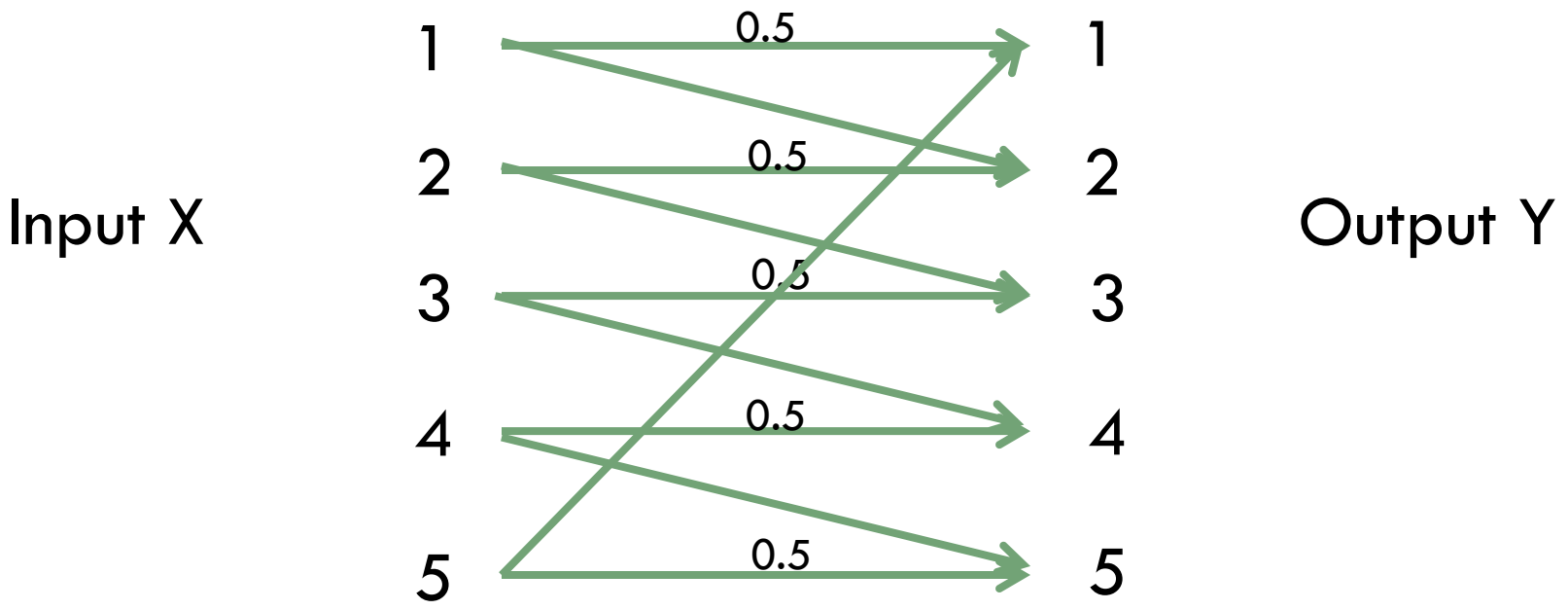
- Consider the n=2 extension of the channel:



Let $\{X_1, X_2\}$ be $\{(1,1), (2,3), (3,5), (4,2), (5,4)\}$

Reminder of the channel

- A tricky typewriter:

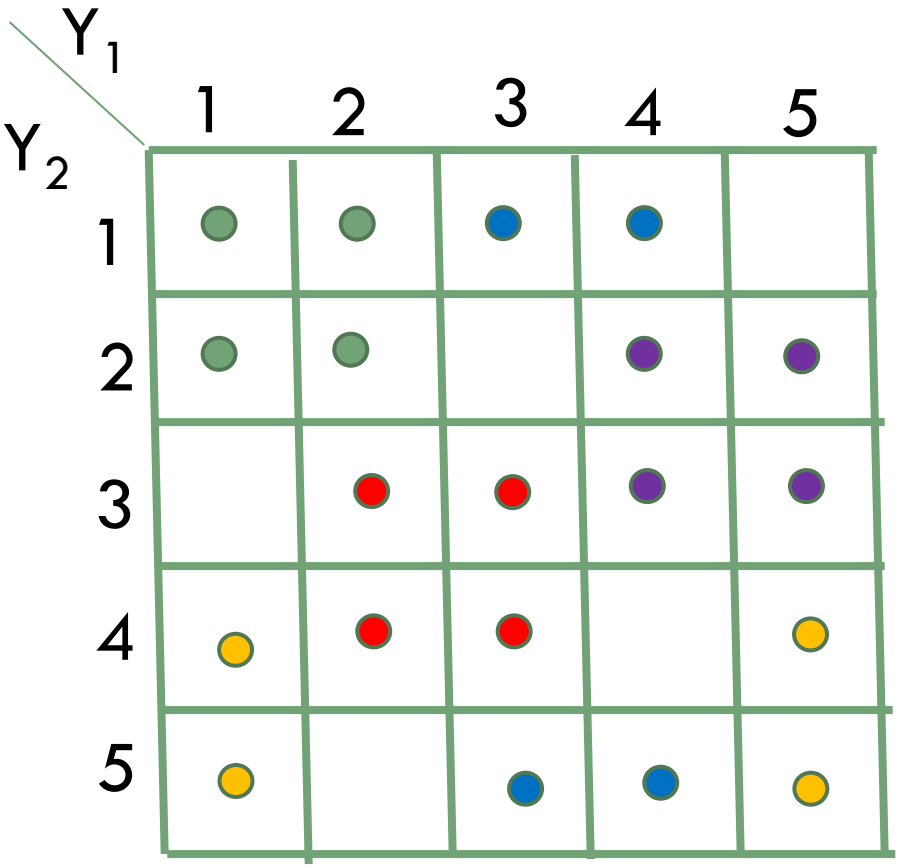


How many noise free symbols?

Clearly at least 2, hopefully more.

Simple examples

- Looking at the outputs:



Let $\{X_1, X_2\}$ be $\{(1,1), (2,3), (3,5), (4,2), (5,4)\}$

Simple examples - observations

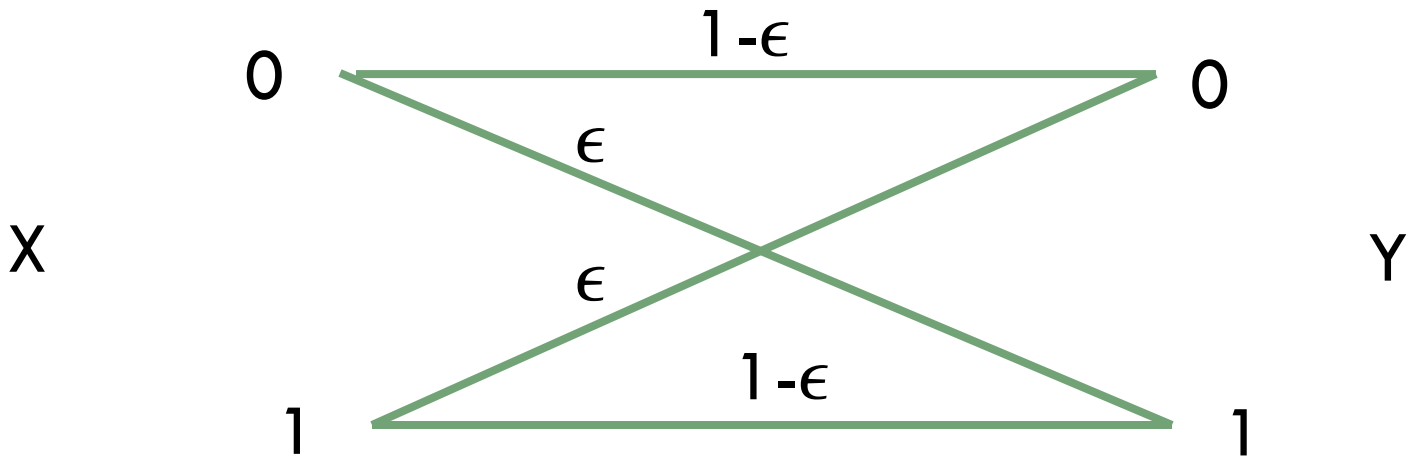
- Note that we get 5 noise-free symbols in n=2 transmissions.

- Thus achieve rate $\frac{log_2 5}{2} = 1.16$ bits/transmission
- with P(error) = 0.

- For arbitrarily small P(error) can use long codes (n→ ∞) to achieve 1.32 bits/transmission, the channel capacity.

The Binary Symmetric Channel (BSC)

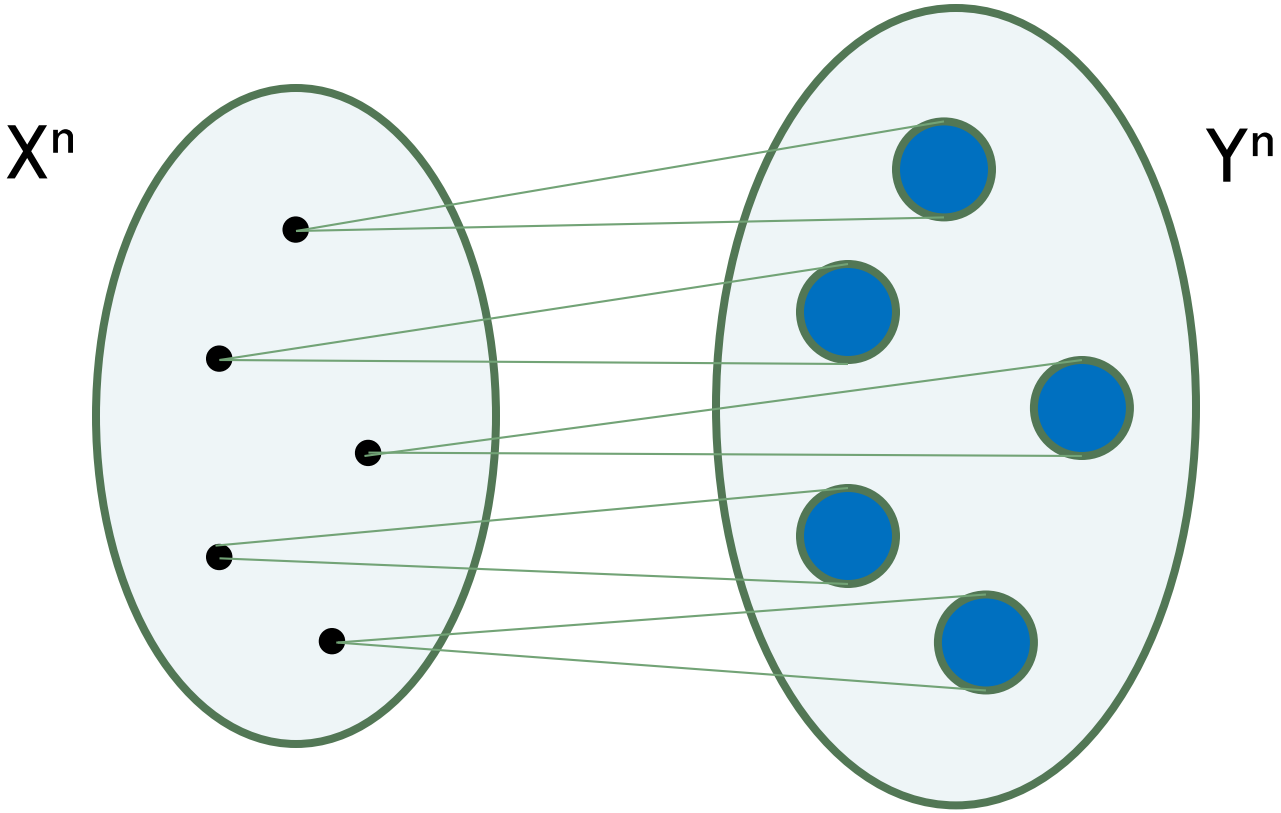
- How many noise free symbols?



A.: Clearly for n=1 there are none.
How about using n large?

Shannon's Second Theorem

- Using the channel n times:



Shannon's Second Theorem

- The Information Channel Capacity of a discrete memoryless channel is

- $$C = \max_{p(x)} I(X;Y).$$

- Note: $I(X;Y)$is a function of $p(x,y) = p(x)p(y|x)$.
- But $p(y|x)$ is fixed by the channel.

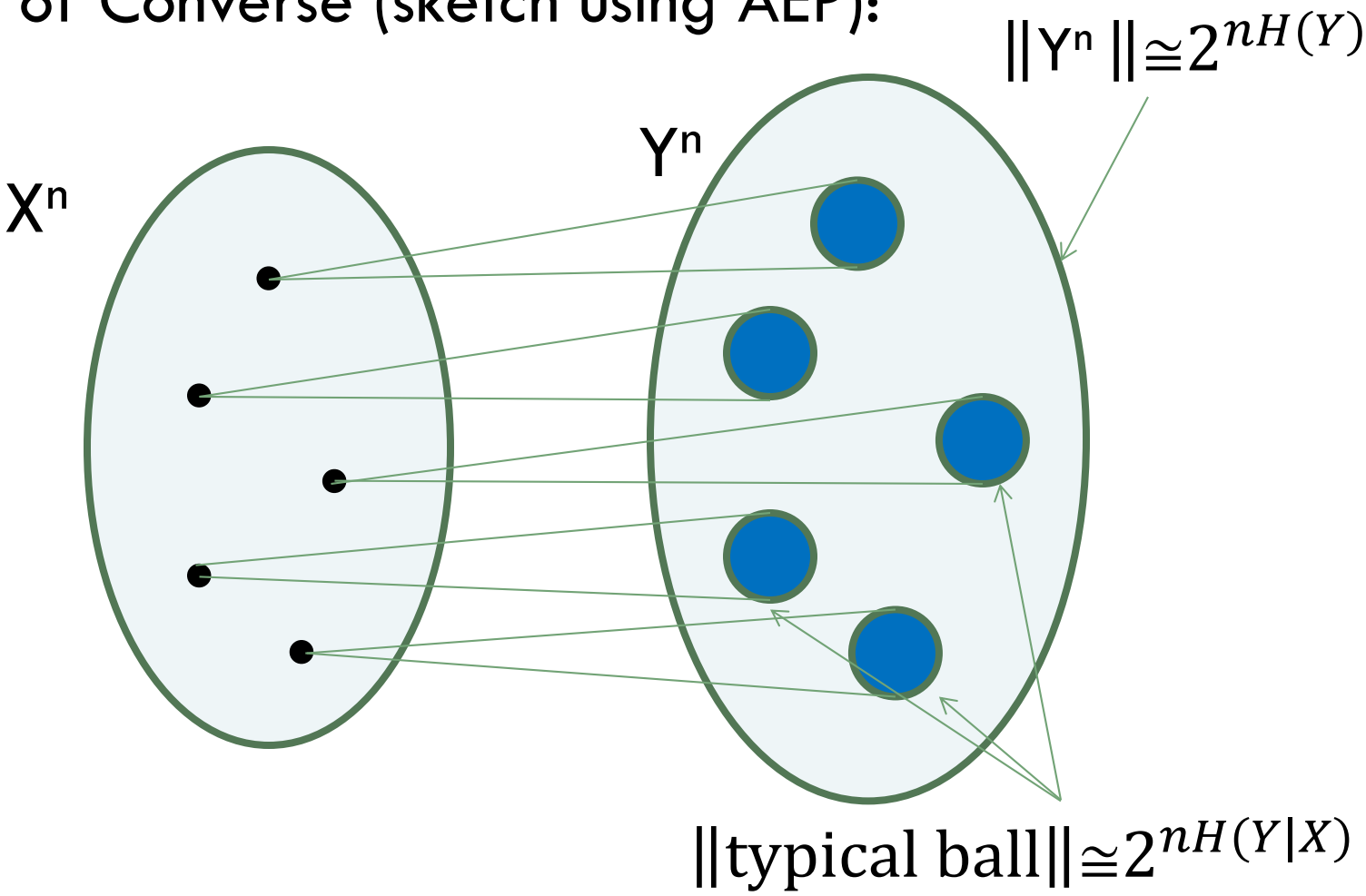
Shannon's Second Theorem

- **Theorem:** For a discrete memoryless channel, all rates R below the information channel capacity C are achievable with maximum probability of error arbitrarily small. Conversely, if the rate is above C, the probability of error is bounded away from zero.

- **Proof:** Achievability: Use random coding to generate codes with a particular $p(x)$distribution in the codewords. Then show that the average *P(error)* tends to zero with $n \rightarrow \infty$ if $R < C$. Then expurgate bad codewords to get a code with small maximum *P(error)*.

Shannon's Second Theorem

- Proof of Converse (sketch using AEP):



Shannon's Second Theorem

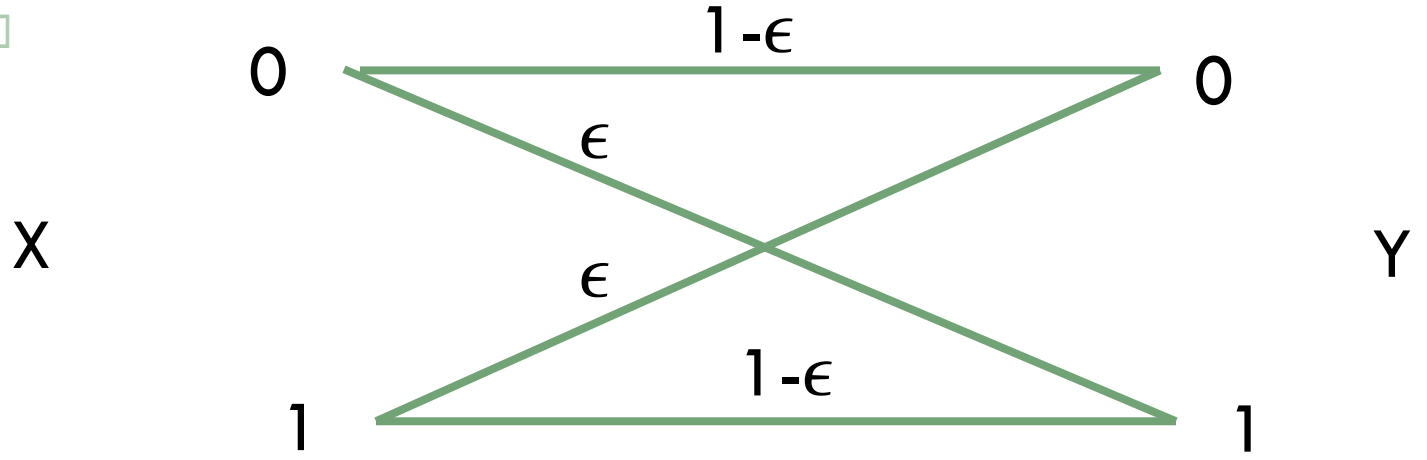
- Proof of Converse (sketch using AEP):
- Recall the sphere packing problem.

  Maximum number of non-overlapping balls is bounded by

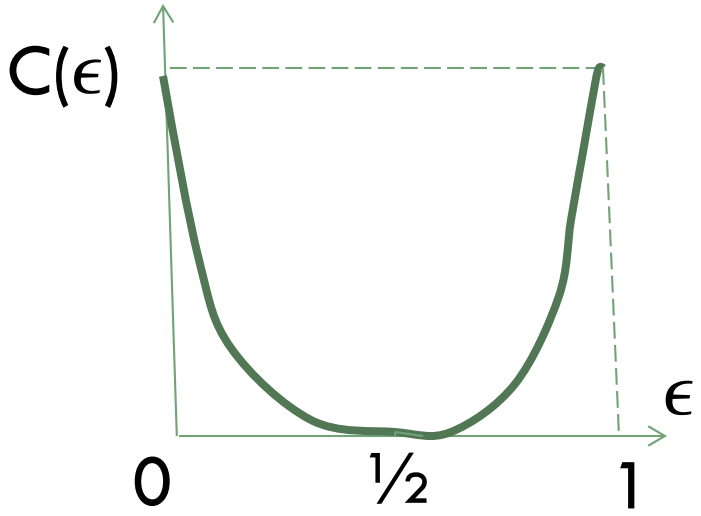
$$\frac{2^{nH(Y)}}{2^{nH(Y|X)}} = 2^{I(X:Y)} \leq 2^{C}$$

- Thus $2^{R} \leq 2^{C}$ and $R \leq C$.
- A formal proof uses Fano's inequality.

Example: The Binary Symmetric Channel

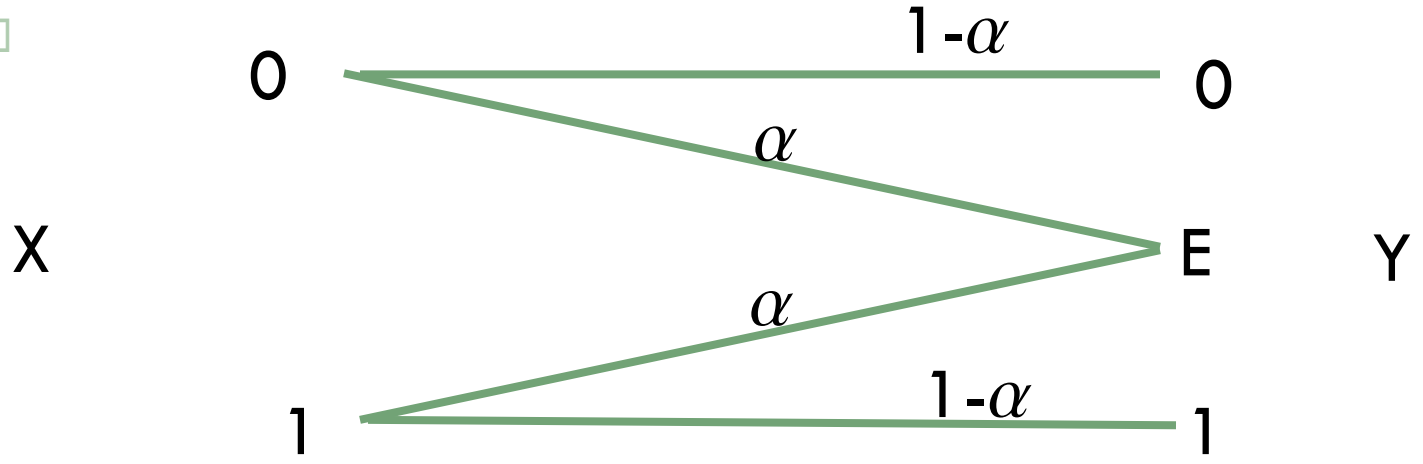


- $C = \max (H(Y) - H(Y|X))$
- $= 1 - h(\epsilon)$ bits/transmission

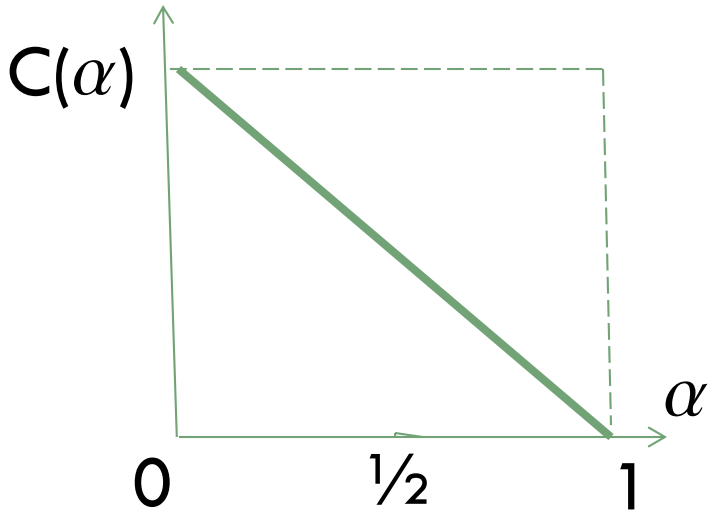


- Note: C=0 for $\epsilon = ½$.

Example: The Binary Erasure Channel



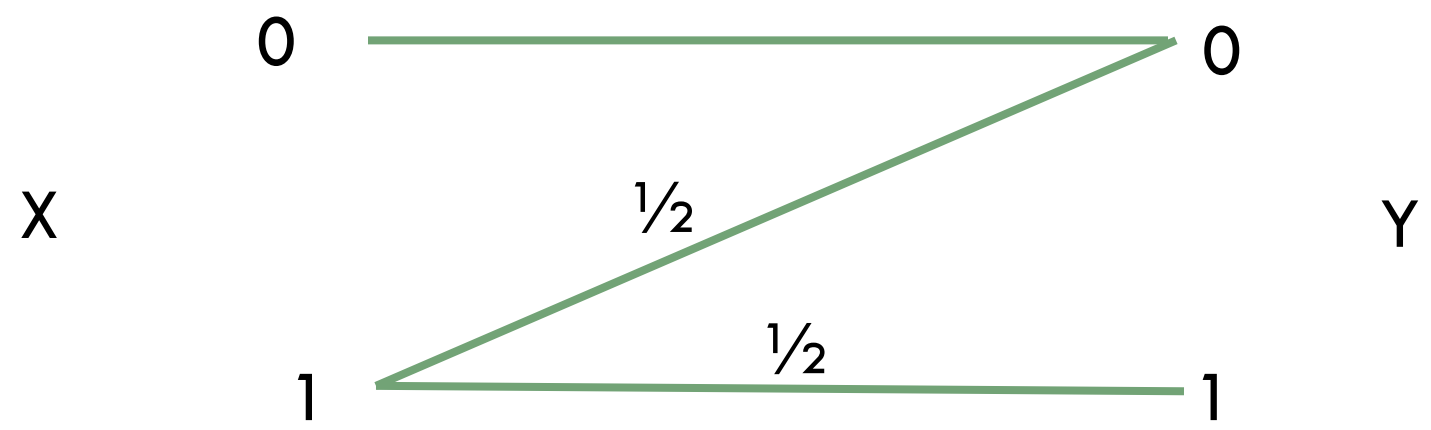
- C = max (H(Y) – H(Y|X))
- = 1 – α bits/transmission



Note: C=0 for $\alpha = 1$.

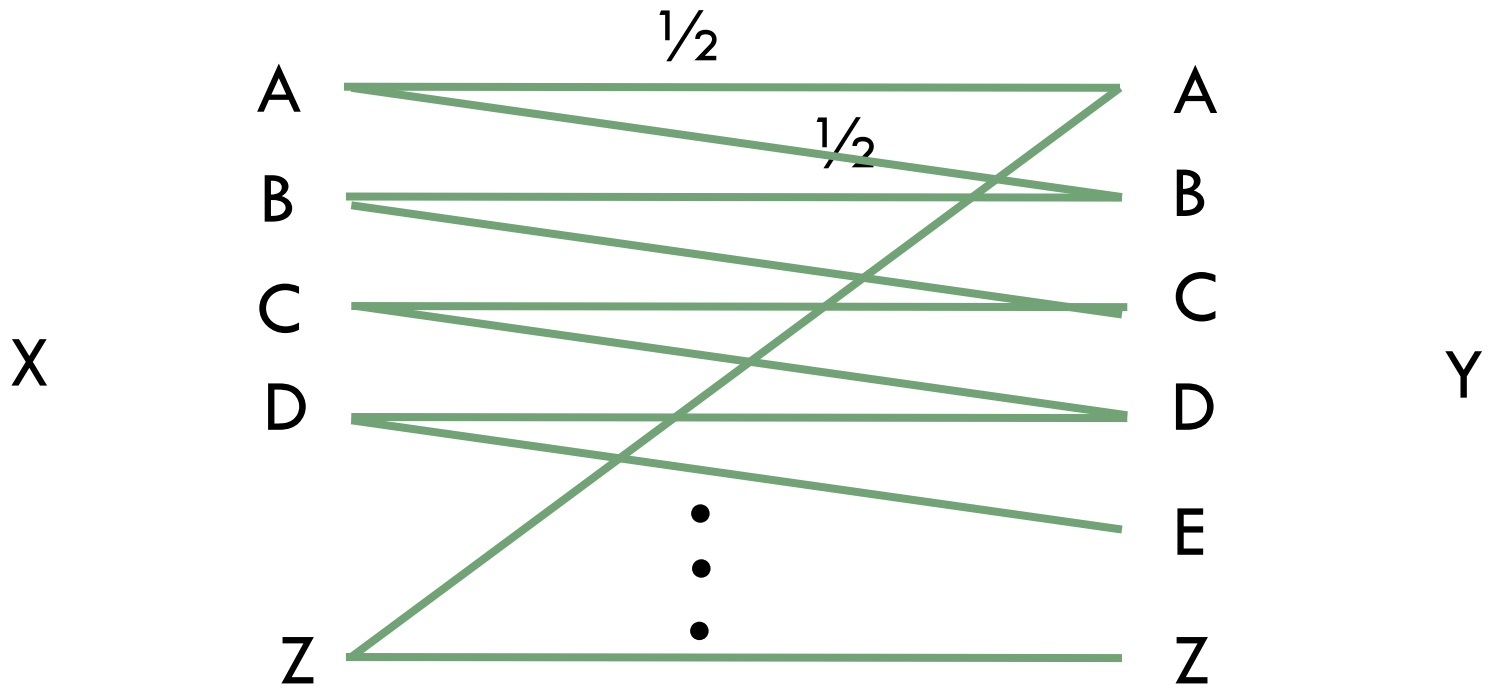
Capacity is achieved with $p(X = 0) = p(X = 1) = ½$.

Example: The Z Channel



- $C = \max_{p(x)} (H(Y) - H(Y|X)) = log_2(5) - 2 = 0.322 \frac{bit}{tr.}$
- Note: Maximizing $p(X = 1) = \frac{2}{5}$.
- Homework: Obtain this capacity.

Example: Noisy typewriter



- C = max (H(Y) – H(Y | X))
- $= log_2 26 - log_2 2 = log_2 13$ bits/transmission
- Achieved with uniform distribution on the inputs.

Remark:

- For this example, we can also achieve
- $C = log_2 13$ bits/transmission with *P(error)=0* and
- $n = 1$ by transmitting alternating input symbols, i.e.,
- $\mathcal{X}$ = {A C E … Z}.

Differential Entropy

- Let X be a continuous random variable with density $f(x)$ and support S. The differential entropy of X is

$$h(X) = -\int_S f(x) \log f(x) dx \qquad \text{(if it exists).}$$

Note: Also written as $h(f)$.

Examples: Uniform distribution

- Let X be uniform in the interval $[0, a]$. Then
- $f(x) = \frac{1}{a}$ in the interval and $f(x) = 0$ outside.

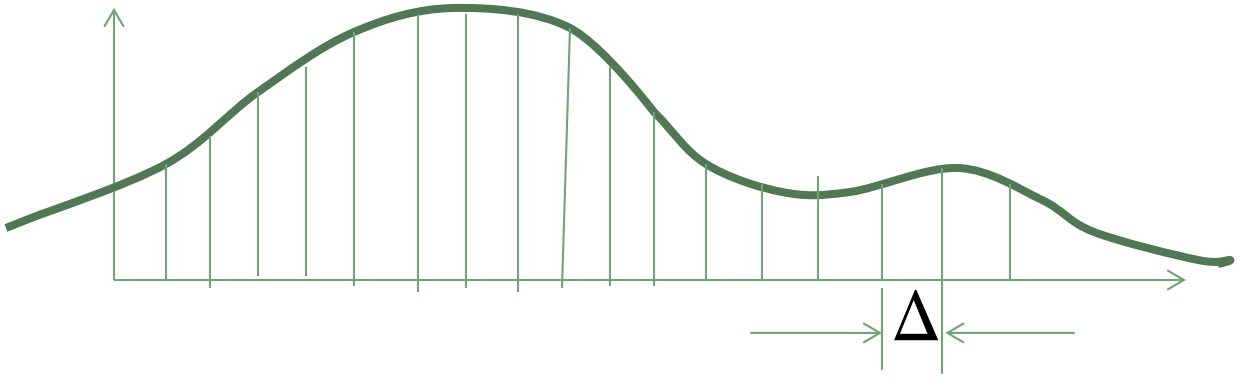
- $h(X) = -\int_0^a \frac{1}{a}\, log\, \frac{1}{a}\, dx = \log a$
- Note that $h(X)$ can be negative for $a < 1$.
- However, $2^{h(f)} = 2^{\log a} = a$ is the size of the support set, which is non-negative.

Example: Gaussian distribution

- Let $X \sim \phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(\frac{-x^2}{2\sigma^2}\right)$
- Then $h(X) = h(\phi) = -\int \phi(x)[-\frac{x^2}{2\sigma^2} - ln\sqrt{2\pi\sigma^2}]\, dx$
- $= \frac{EX^2}{2\sigma^2} + \frac{1}{2} ln\ 2\pi\sigma^2$
- $= \frac{1}{2} ln\ 2\pi e\sigma^2$ nats
- Changing the base we have $h(X) = \frac{1}{2} log\ 2\pi e\sigma^2$ bits

Relation of Differential and Discrete Entropies

- Consider a quantization of X, denoted by X^{Δ}



- Let $X^{\Delta} = x_i$ inside the ith interval.

Then $H(X^{\Delta}) = -\sum_i p_i \log p_i$

$= -\sum_i \Delta f(x_i) \log f(x_i) - \log \Delta$

$\cong h(f) - \log \Delta$

Differential Entropy

- So the two entropies differ by the log of the quantization level Δ.

- We can define joint differential entropy, conditional differential entropy, K-L divergence and mutual information with some care to avoid infinite differential entropies.

K-L divergence and Mutual Information

- ??
- $D(f \quad g) = \int f \log \frac{f}{g}$
- $I(X;Y) = \int f(x,y) \log \frac{f(x,y)}{f(x)g(y)} dx\, dy$
- Thus, $I(X;Y) = h(X) + h(Y) - h(X,Y)$.

Differential entropy of a Gaussian vector

- Theorem: Let $\boldsymbol{X}$ be a Gaussian *n*-dimensional vector with mean μ and covariance matrix K. Then

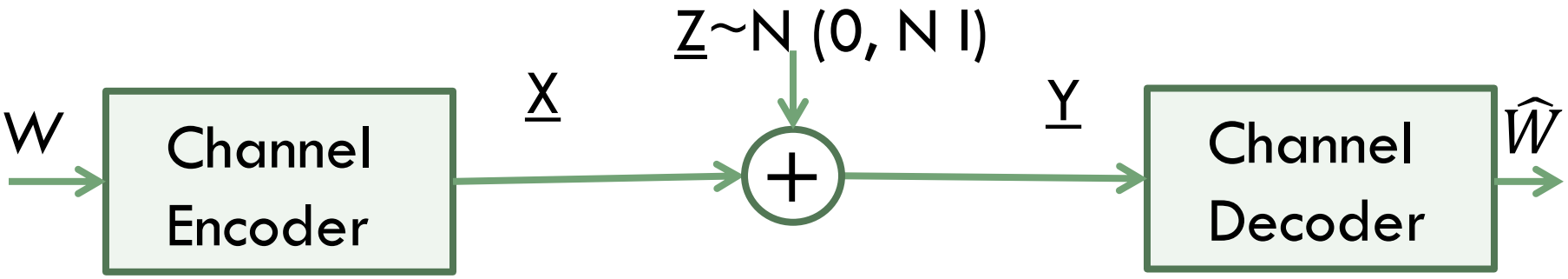
$$h(\boldsymbol{X}) = \frac{1}{2}\log(2\pi e)^n |K|$$

- where $|K|$ denotes the determinant of K.

- Proof: Algebraic manipulation.

The Gaussian Channel

- The Gaussian Channel Problem:



Power constraint: $EX^2 \leq P$

- $W \in \{1,2,\ldots,2^{nR}\}$ = message set of rate R
- $\underline{X} = (x_1\ x_2\ \ldots\ x_n)$ = codeword input to channel
- $\underline{Y} = (y_1\ y_2\ \ldots\ y_n)$ = codeword output from channel
- $\widehat{W}$ = decoded message P(error) = $P\{W \neq \widehat{W}\}$

The Gaussian Channel

-
- $\mathrm{C}apacity \qquad C = \max_{f(x):\, EX^2 \le P} I(X;Y)$

- $I(X;Y) = h(Y) - h(Y|X) = h(Y) - h(X+Z|X)$
- $\quad = h(Y) - h(Z) \le \frac{1}{2}\log 2\pi e(P+N) - \frac{1}{2}\log 2\pi eN$
- $\quad = \frac{1}{2}\log\left(1+\frac{P}{N}\right)$ bits/transmission

The Gaussian Channel

- The capacity of the discrete time additive Gaussian channel:

$$C = \frac{1}{2}\log\left(1+\frac{P}{N}\right) \text{ bits/transmission}$$

- achieved with X ~ N(0 , P).

Bandlimited Gaussian Channel

- Consider the channel with continuous waveform inputs $x(t)$ with power constraint $(\frac{1}{T}\int_0^T x^2(t)dt \leq P)$ and Bandwidth limited to W. The channel has white Gaussian noise with power spectral density $N_0/2$ watt/Hz.
- In the interval $(0,T)$ we can specify the code waveform by 2WT samples (Nyquist criterion). We can transmit these samples over discrete time Gaussian channels with noise variance $N_0/2$. This gives
-

$$C = W\log(1+\frac{P}{N_0W}) \text{ bit/second}$$

Bandlimited Gaussian Channel

- $$C = W \log(1 + \frac{P}{N_0 W}) \text{ bit/second}$$
- Note: If W→ ∞
- we have C = $\frac{P}{N_0} log_2 e$ bits/second.

Bandlimited Gaussian Channel

- Let $\frac{R}{W}$ be the spectral density ν in bits per second per Hertz. Also let $P = E_b R$ where E_b is the available energy per information bit.
- We get
- $\frac{R}{W} \leq \frac{C}{W} = \log(1+\frac{E_b R}{N_0 W})$ bit/second.
- Thus
-

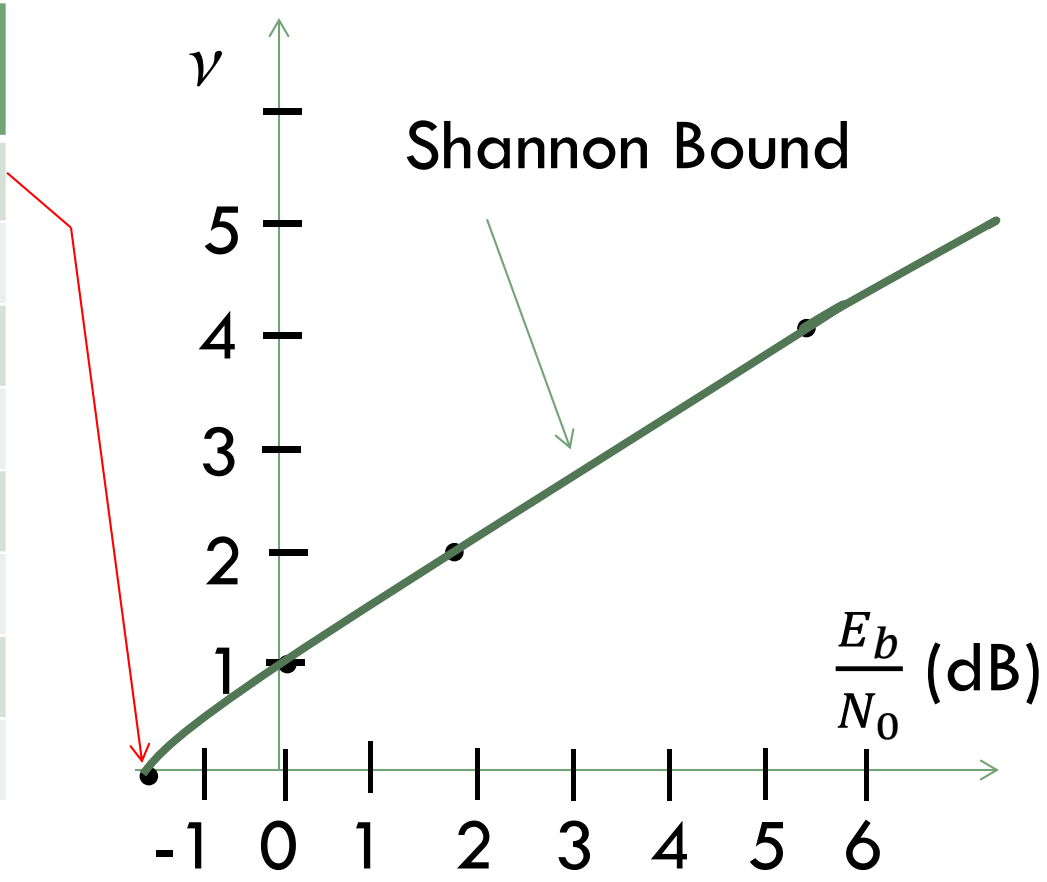
$$\frac{E_b}{N_0} \geq \frac{2^{\nu}-1}{\nu}$$

This relation defines the so called Shannon Bound.

The Shannon Bound

$$\frac{E_b}{N_0} \geq \frac{2^{\nu}-1}{\nu}$$

ν	$\frac{E_b}{N_0}$	$\frac{E_b}{N_0}$ (dB)
→0	0.69	-1.59
0.1	0.718	-1.44
0.25	0.757	-1.21
0.5	0.828	-0.82
1	1	0
2	1.5	1.76
4	3.75	5.74
8	31.87	15.03



Shannon's Water Filling Solution

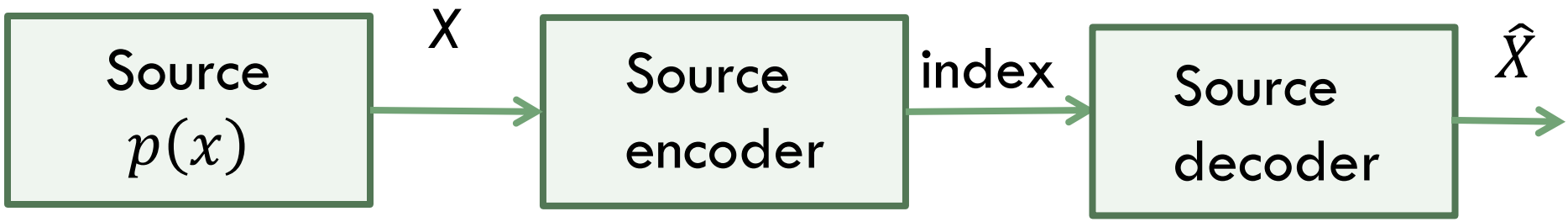
Parallel Gaussian Channels

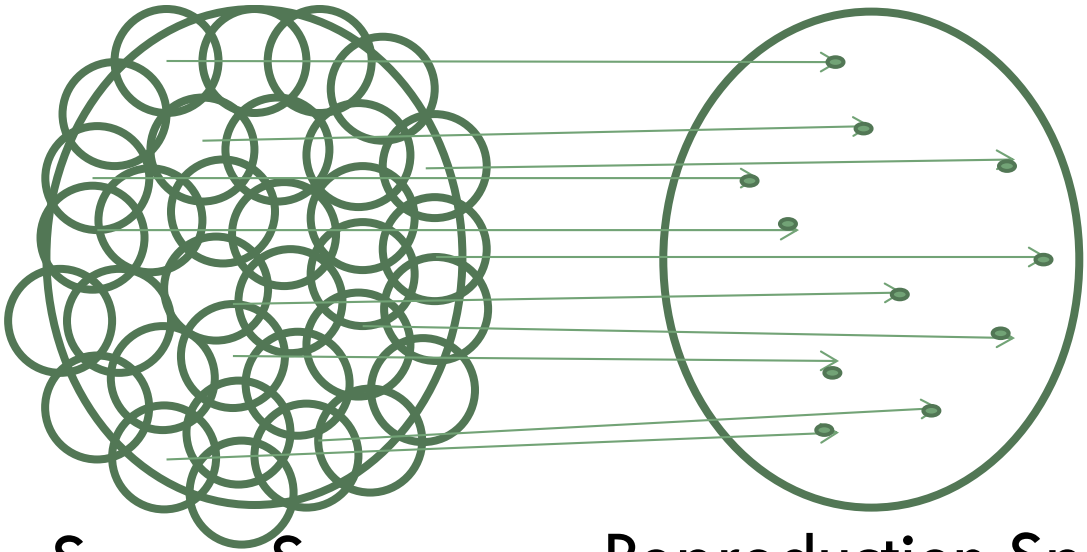
Example of Water Filling

- Channels with noise levels 2, 1 and 3.
- Available power = 2
- Capacity= $\frac{1}{2}\log\left(1+\frac{0.5}{2}\right)+\frac{1}{2}\log\left(1+\frac{1.5}{1}\right)+\frac{1}{2}\log\left(1+\frac{0}{3}\right)$
- Level of noise + signal power = 2.5
- No power allocated to the third channel.

Rate Distortion Theory

- Want to represent a source efficiently.





Rate Distortion Theory

- Define a distortion $d(x, \hat{x}) = \frac{1}{n}\sum_{i=1}^{n} d(x_i, \hat{x}_i)$
- where $d(.,.)$: $\chi \times \hat{\chi} \rightarrow \mathbb{R}^+$
- Want to find 2^{nR} $\hat{x}$'s (points in representation space) and a mapping $\hat{X}(x) : \chi^n \rightarrow \hat{\chi}^n$ such that
- $Ed\left(X, \hat{X}(X)\right) \leq D.$ Eq. ♣
- Theorem: This is possible iff (Shannon, 1959)
- $$R \leq R(D) = \min_{p(\hat{x}|x) \text{ satisfying Eq. } \clubsuit} I(X; \hat{X})$$

All depends on $p(\hat{x}|x)$

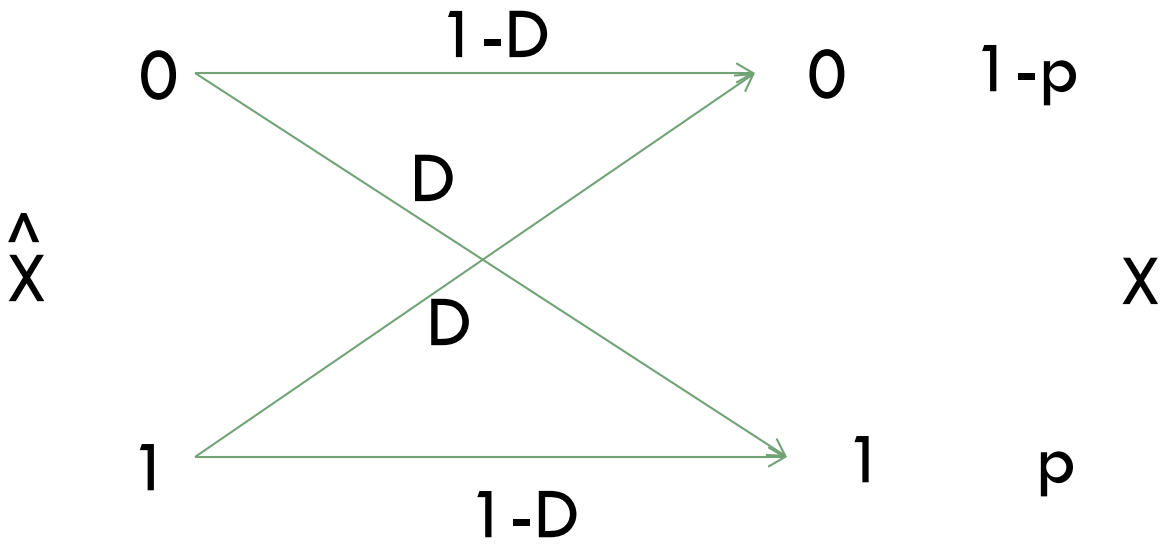
- Note:
- $E\ d\left(X, \hat{X}(X)\right)$ as well as $I(X;\hat{X})$ are functions of
- $p(x,\hat{x}\ ) = p(x)\ p(\hat{x}|x)$, but $p(x)$ is fixed by the source.

Example: Binary Source

- Binary source with Hamming distortion:
- $R(D) = \begin{cases} h(p) - h(D), & 0 \le D \le \min(p, 1-p) \\ 0 & D > \min(p, 1-p) \end{cases}$

- $I(X;\hat{X}) = H(X) - H(X|\hat{X})$
- $= h(p) - H(X \oplus \hat{X} | \hat{X})$
- $\geq h(p) - H(X \oplus \hat{X})$
- $\geq h(p) - h(D)$

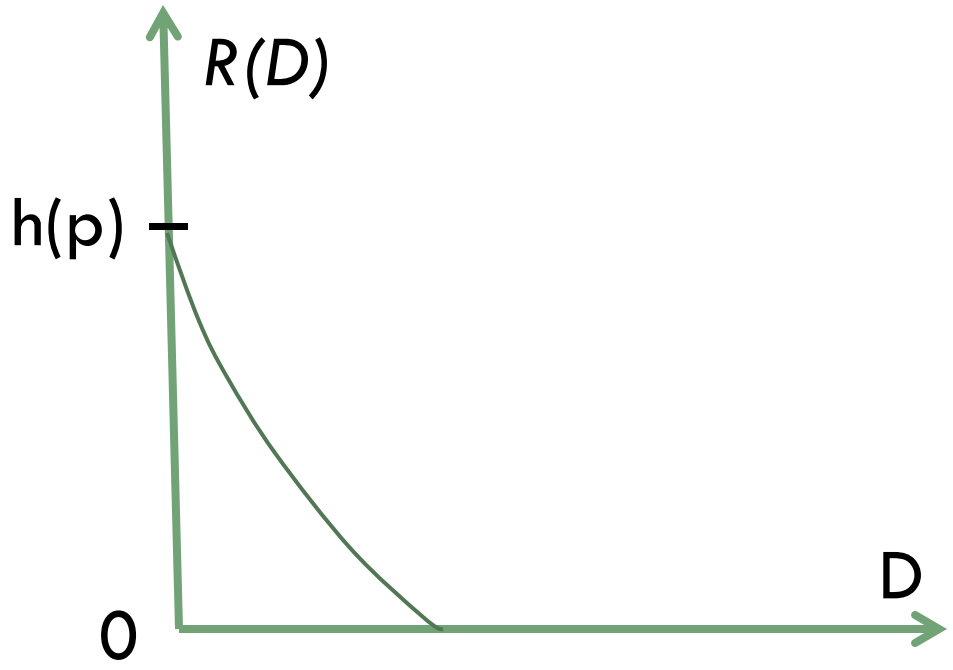
Test channel for binary source



with $P(\hat{X} = 0) = \frac{1-p-D}{1-2D}$ and $P(\hat{X} = 1) = \frac{p-D}{1-2D}$

This satisfies the bound with equality

- Thus $R(D) = h(p)-h(D)$, for $0 \le D \le \min(p,1-p)$, and
  $R(D)=0$, otherwise.



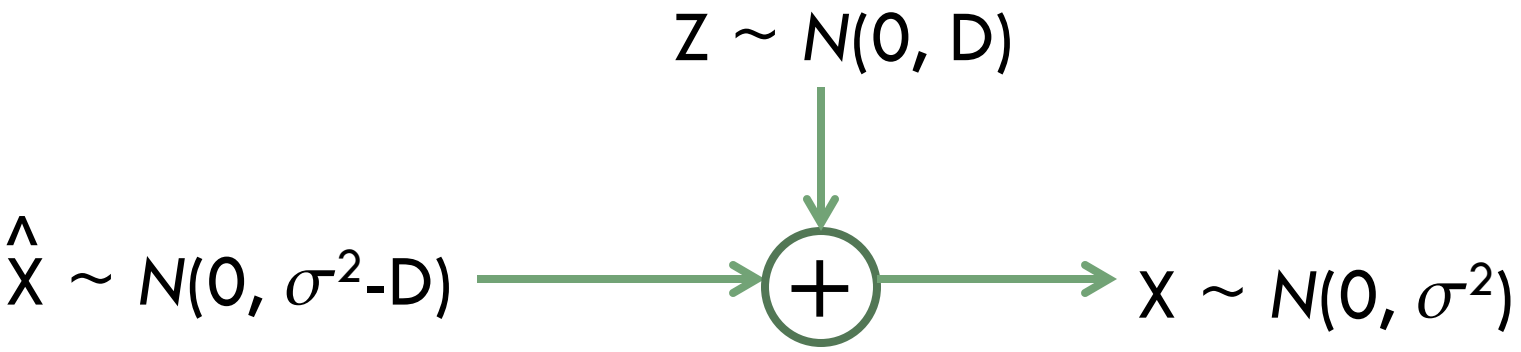
Example: Gaussian Source, MSE Distortion

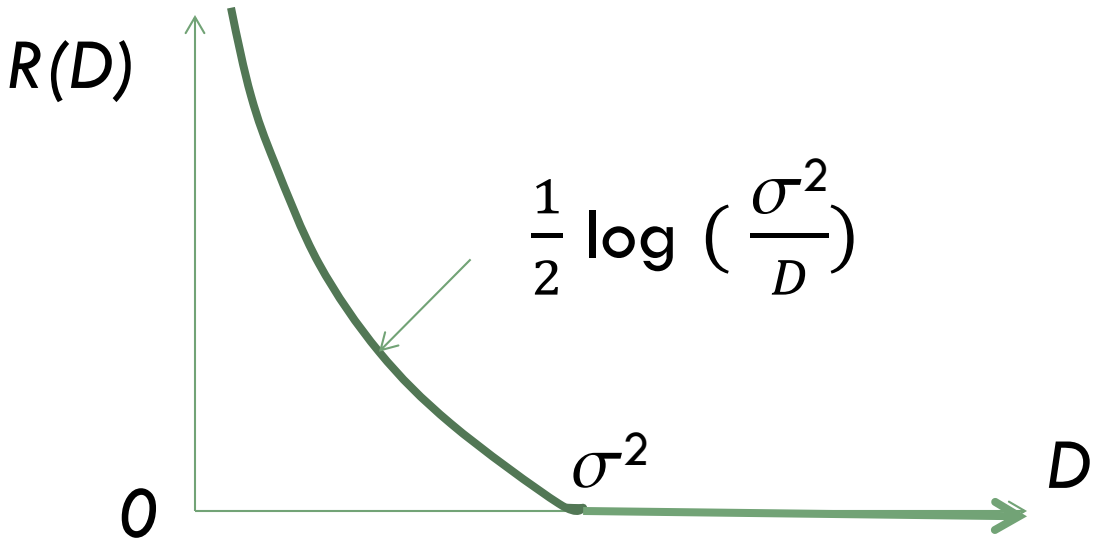
- Have $R(D) = \frac{1}{2} log \frac{\sigma^2}{D}$, for $0 \leq D \leq \sigma^2$, and
  $R(D)=0$, otherwise.

- $I(X;\hat{X}) = h(X) - h(X|\hat{X})$
- $\geq \frac{1}{2} \log (2\pi e \sigma^2) - \frac{1}{2} \log (2\pi e D)$
- $= \frac{1}{2} \log \left(\frac{\sigma^2}{D} \right)$

Test channel and R(D) curve

- This is achievable with the test channel





Useful Rule of thumb:

- Inverting R(D) we have
- $D(R) = 2^{-2R}\,\sigma^2$ (Distortion x Rate function)
- Thus Max SNR (dB) $= 10\,log_{10}\left(\frac{\sigma^2}{D(R)}\right) = 20R\;\; log_{10}2$
- $\cong\; 6R$ (Note: R in bits/source sample)
- → For example an audio system with 16 bits/sample
  - can give you ~ 96 dB of SNR !

Kolmogorov Complexity

- The intrinsic descriptive complexity of an object.

- The Kolmogorov complexity of a string x with respect to a universal computer U is defined as

- $$K_U(x) = \min_{p:U(p)=x} \ell(p),$$

- the minimum length over all programs that print x and stop.

Universality

- For any computer a there is a constant c(A) such that
- $$K_U(x) \leq K_A(x) + c(A)$$
- Note: The constant c(A) depends on A but not on x.

Examples:

- Ex. 1) Repeating sequence x=01010101010101…
- $K(x \mid \ell(x)) = c$
- Program: Print alternating binary digits of length $\ell(x)$.

- Ex. 2) Some arbitrary sequence x
- $K(x \mid \ell(x)) \leq \ell(x) + c$
- Program: Print this sequence: x.

Examples

- Ex. 3) Upper bound on K(x)
- $K(x) \leq K(x \mid \ell(x)) + 2 \log \ell(x) + c$

- Describe $\ell(x)$ by repeating every bit in the binary representation and terminating with 01.

- Alternatively use $\log^* n$ ($= \log n + \log \log n + \ldots$)

Examples:

- Ex. 4) $K(n) \leq \log n + 2 \log \log n + c$
- Ex. 5) $K(n_1 + n_2) \leq K(n_1) + K(n_2) + c$
- Ex. 6) $K(\pi) = c$
- Ex. 7) K (n digit representation of π) $= K(n) + c$

Monkeys in the computer

- What is the probability that a monkey typing on a computer will produce:
- a) 0^n followed by an arbitrary string;
- b) 0^n1 followed by an arbitrary string;
- c) The works of Shakespeare followed by an arbitrary string
- Reasoning: The probability that a computer with random input will type x followed by an arbitrary sequence is the sum of the probabilities of programs that print x y summed over all y.

Universal probability

- Thus $p_U(x\ \ldots) = \sum_y\ p_U(x\ y)$

- where $p_U(x) = \sum_{p:U(p)=x} 2^{-l(p)}$
- This sum is approximated by the largest term, corresponding to the simplest x y concatenation.
- Answers: a) the smallest program for 0^n y with y arbitrary is “print zeros forever”.
- $$p_U(\ 0^n\ \ldots) \cong 2^{-c}$$

Universal probability

- b) $p_U(\ 0^n 1 \ldots) \cong 2^{-\log^* n - c}$
- c)) $p_U(\text{Shakespeare} \ldots) \cong 2^{-n\, H(English)}$
- Note: We can define a Universal probability
- $P_U(x) \cong 2^{-K(x)}$

Lemma

- For any computer U
- $$\sum_{p:U(p)halts} 2^{-l(p)} \leq 1$$
- Proof: If the computer halts on any program it does not look ahead. So no halting program is the prefix of another, like instantaneous codes. Their lengths satisfy Kraft's inequality.

Kolmogorov Complexity and Entropy

- Let X be i.i.d. ~ f(x) on a finite alphabet χ.
- Then $E \frac{1}{n} K(X^n | n) \rightarrow H(X)$
- Proof outline: Uses Kraft inequality, Jensen´s inequality and the concavity of H(.)

Applications to Biology

- BCH error correcting codes have been found in DNA sequences generated by BCH codes over GF(4)
- L.C.B. Faria, A.S.L. Rocha, J.H. Kleinschmidt, R. Palazzo Jr. and M.C. Silva-Filho
- The question raised by researchers in the field of mathematical biology regarding the existence of error-correcting codes in the structure of the DNA sequences is answered positively. It is shown, for the first time, that DNA sequences such as proteins, targeting sequences and internal sequences are identified as codewords of BCH codes over Galois fields.
- Electronics Letters, vol 46, No. 3, 4/Feb/2010

Applications to Economics

- Stock Market:
- Portfolio $b=(b_1\ b_2\ \ldots\ b_m)$, $b_i \geq 0$, $\sum b_i = 1$
- Stock vector $\mathbf{X}= (x_1, x_2, \ldots x_m)$
- Stocks $\mathbf{X}_i \geq 0$, i = 1,2,…, n.
- x_i represent the relative final price w.r.t. initial price in day i. For example, $x_i = 1.03$ represent a 3% variation that day.
- The wealth after n days using portfolio b is
- $$S_n = \prod_{i=1}^{n} b^T X_i$$

Optimal portfolio

- Def.: The growth rate of a stock portfolio b w.r.t. to a stock market distribution F(x) is
- $$W(b,F) = E \log b^T X.$$
- Def. The optimal growth rate W*(F) is
- $$W^*(F) = \max_b W(b,F)$$
- Theorem: The optimal wealth after n days behaves as $S_n^* \approx 2^{nW^*}$ with probability 1.

Proof

- By the strong Law of Large Numbers,
- $\frac{1}{n} \log S_n^* = \frac{1}{n} \sum_{i=1}^{n} \log b^{*T} X_i$
- $\rightarrow W^*$ with probability 1.
- Thus
- $S_n^* \approx 2^{nW^*}$ with probability 1.

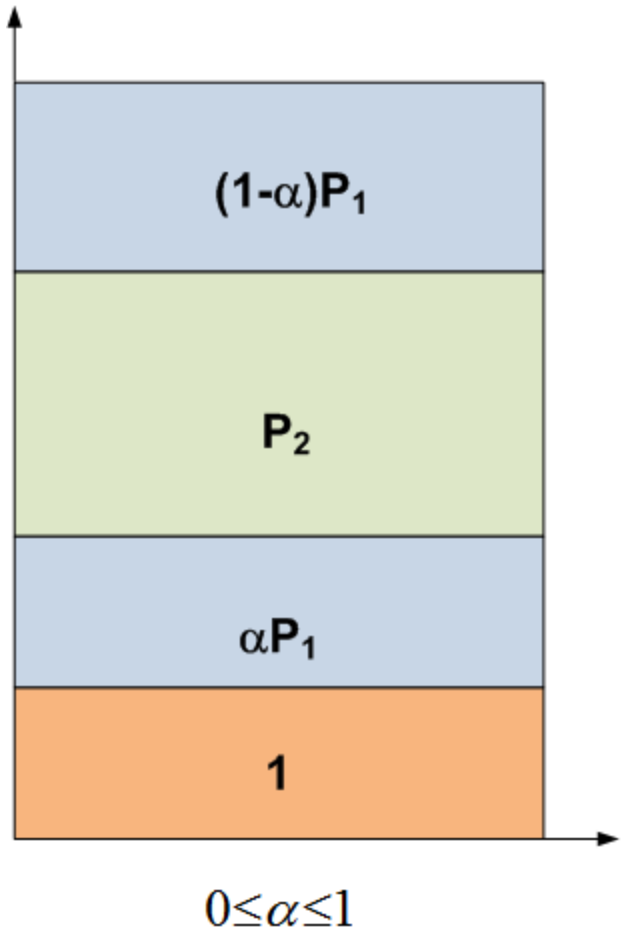
Multiple User Information Theory

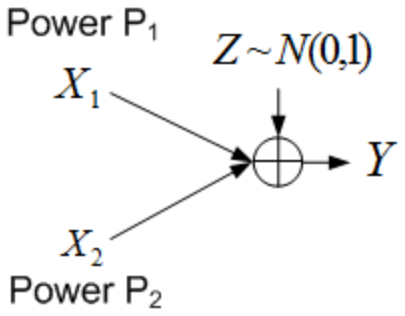
- Building Blocks:

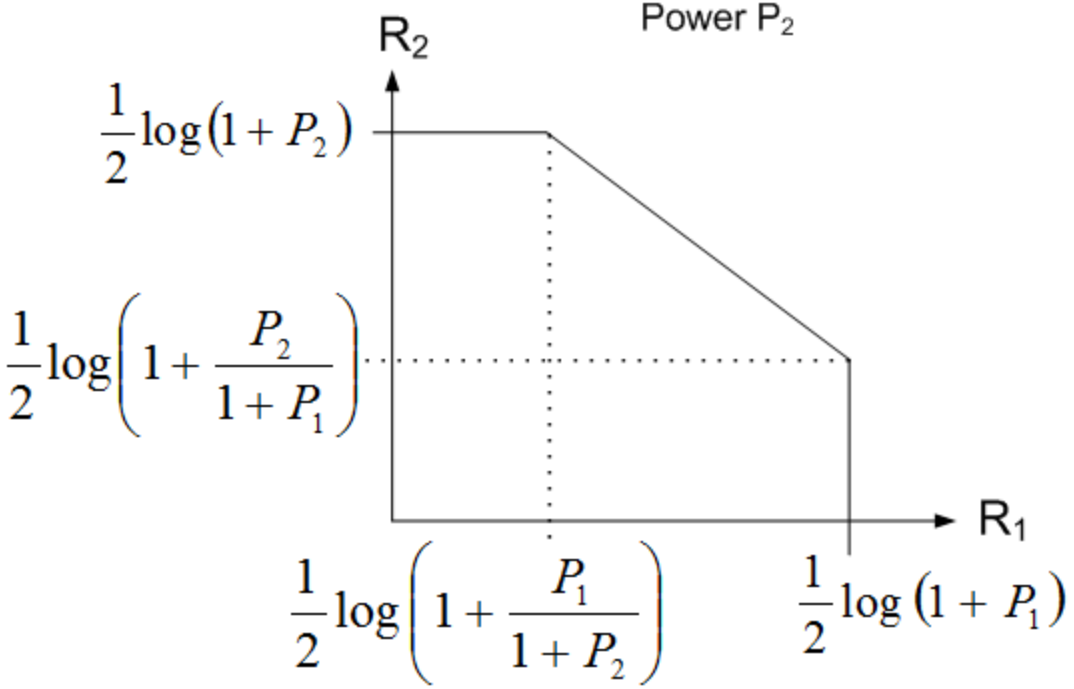
- Multiple Access Channels (MACs)
- Broadcast Channels (BCs)
- Interference Channels (IFCs)
- Relay Channels (RCs)

- Note: These channels have their discrete memoryless and Gaussian versions. For simplicity we will look at the Gaussian models.

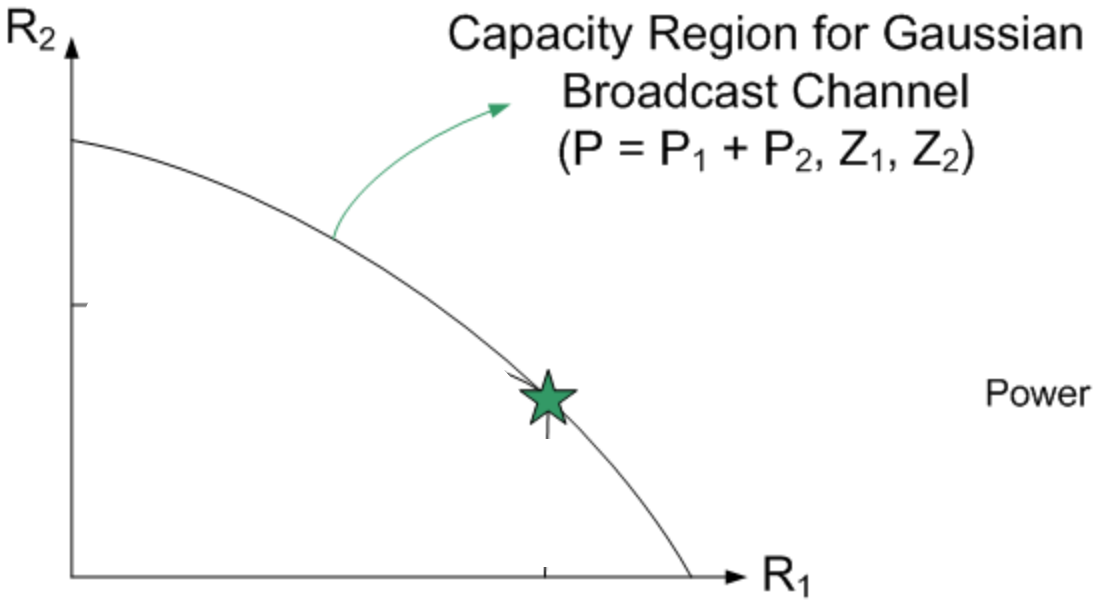
Multiple Access Channel (MAC)

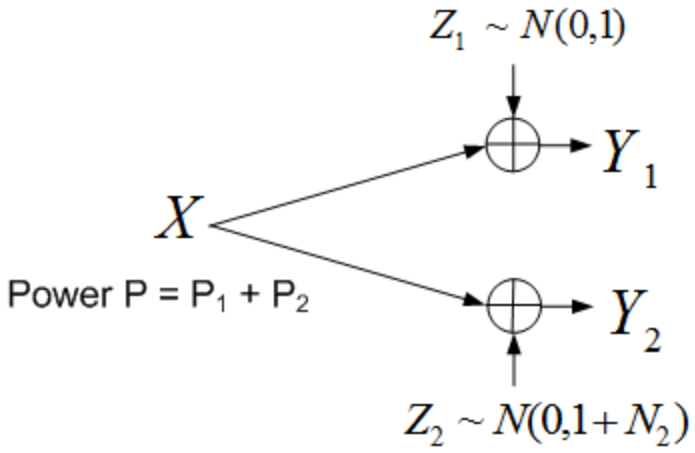






Gaussian Broadcast Channel





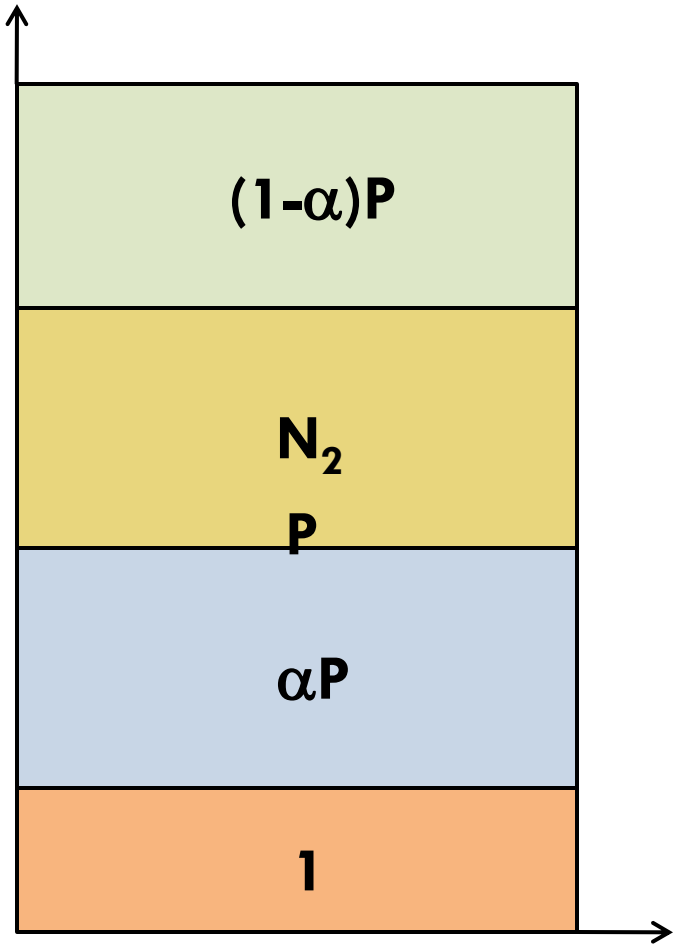
$$C_{BC}\{R_1, R_2\}:$$

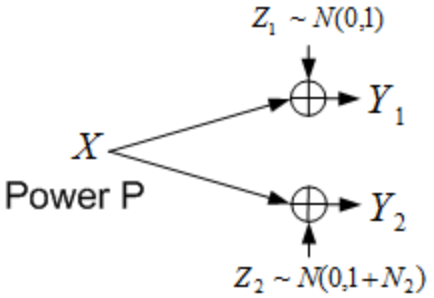
$$0 \le \alpha \le 1$$

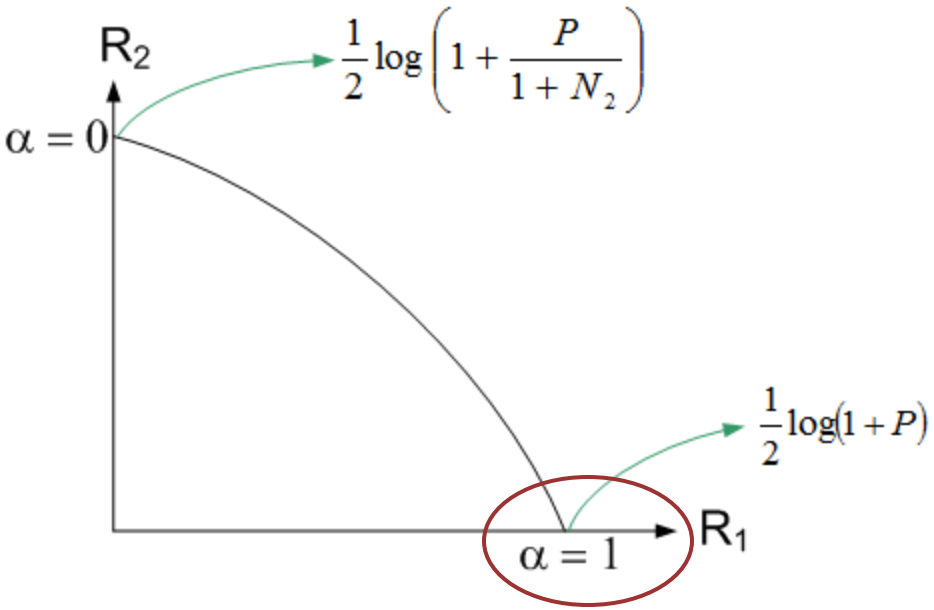
$$0 \le R_1 \le \frac{1}{2}\log(1 + \alpha P)$$

$$0 \le R_2 \le \frac{1}{2}\log\left(1 + \frac{(1-\alpha)P}{1 + N_2 + \alpha P}\right)$$

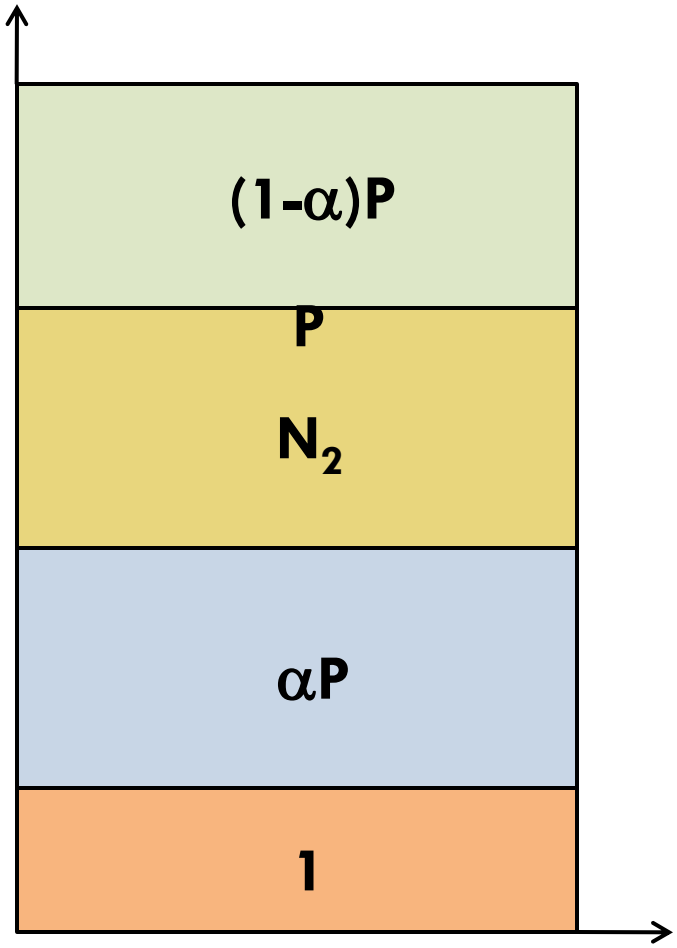
Superposition coding

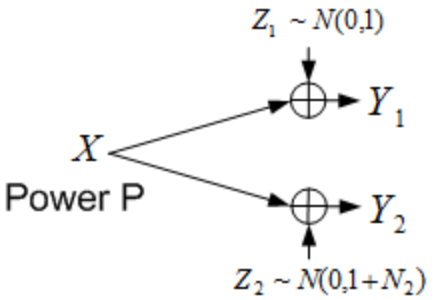


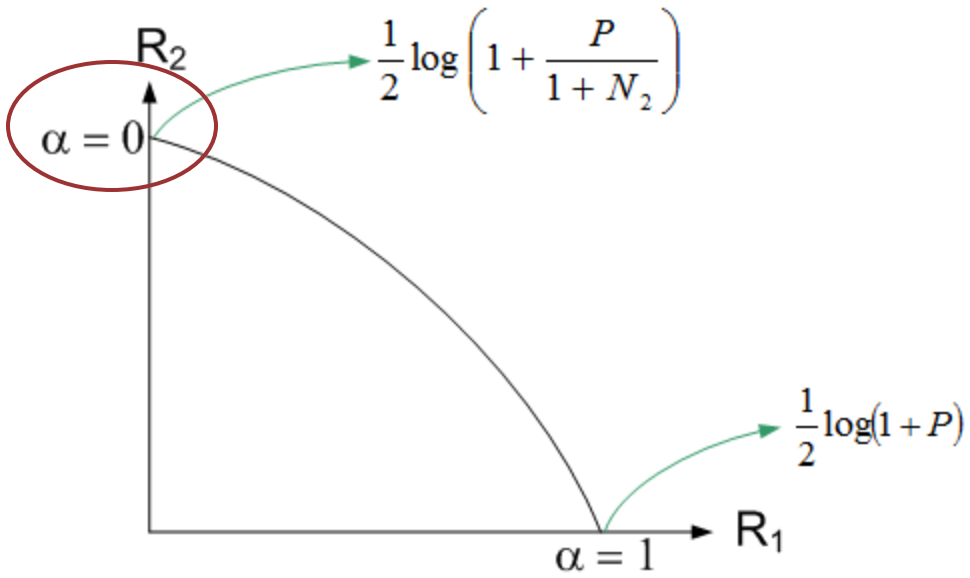




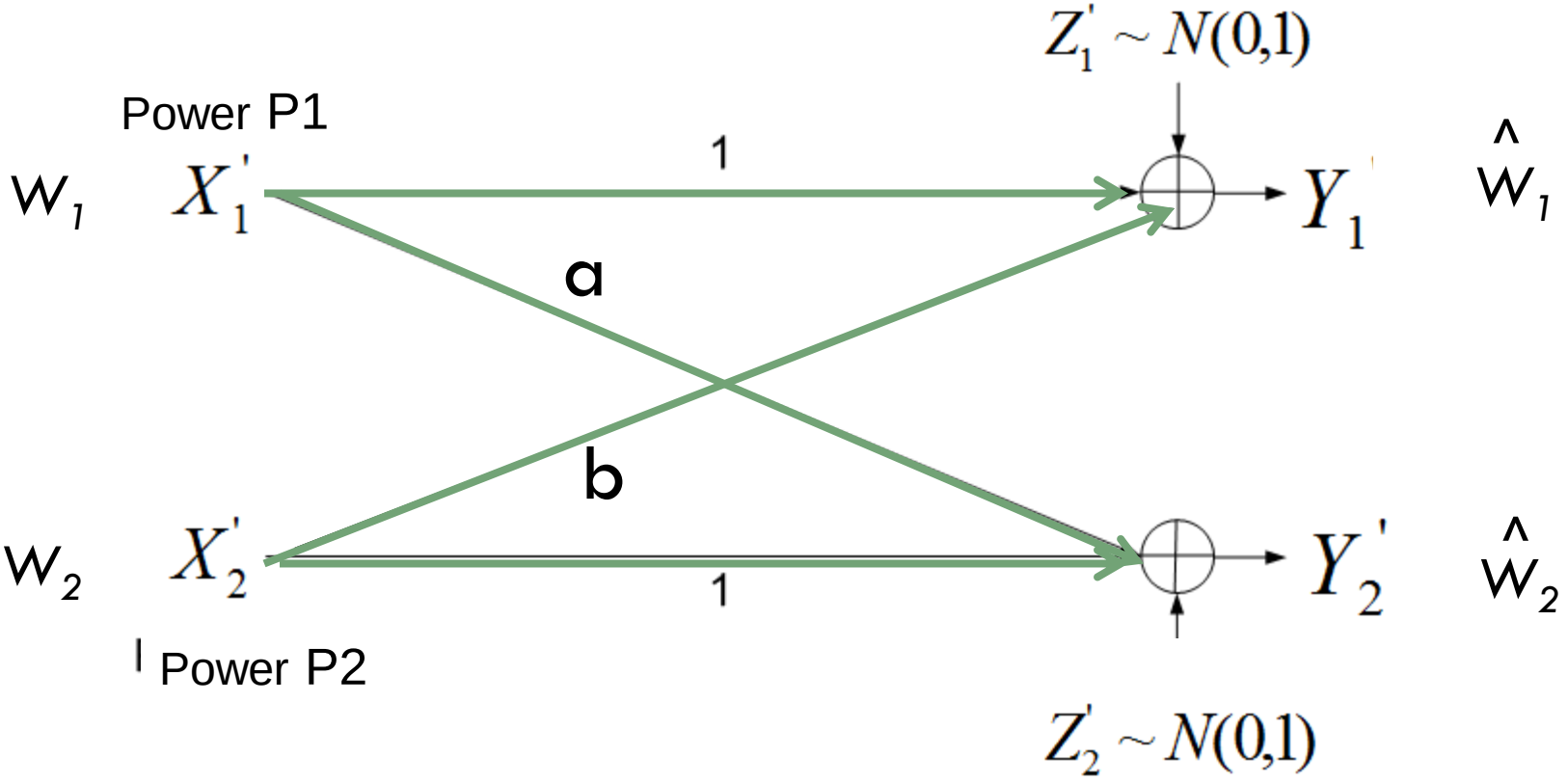
Superposition coding



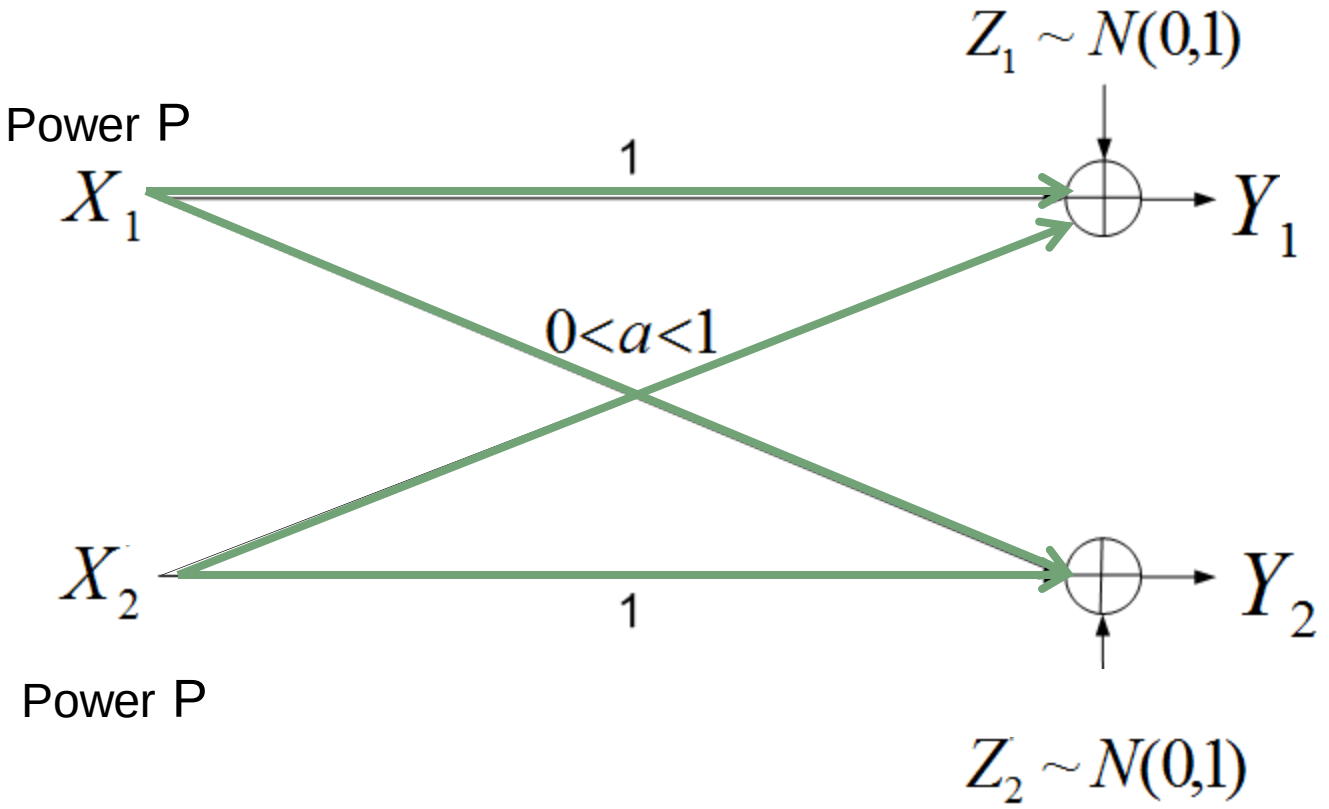




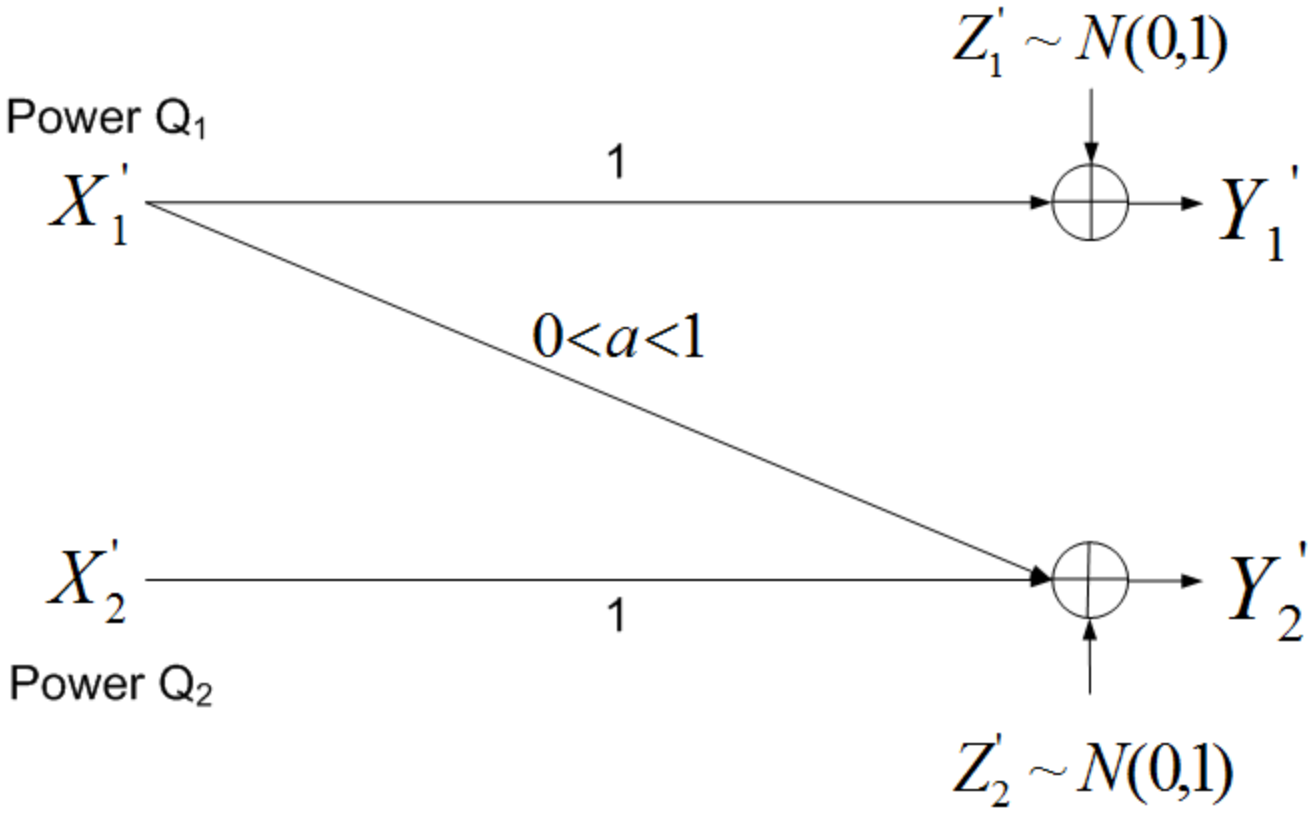
Standard Gaussian Interference Channel



Symmetric Gaussian Interference Channel



Z-Gaussian Interference Channel



Interference Channel: Strategies

Things that we can do with interference:

1. Ignore (take interference as noise (IAN)
2. Avoid (divide the signal space (TDM/FDM))
3. Partially decode both interfering signals
4. Partially decode one, fully decode the other
5. Fully decode both (only good for strong interference, $a \geq 1$)

Interference Channel: Brief history

- Carleial (1975): Very strong interference does not reduce capacity ($a^2 \geq 1+P$)
- Sato (1981), Han and kobayashi (1981): Strong interference ($a^2 \geq 1$) : IFC behaves like 2 MACs
- Motahari, Khandani (2007), Shang, Kramer and Chen (2007), Annapureddy, Veeravalli (2007):

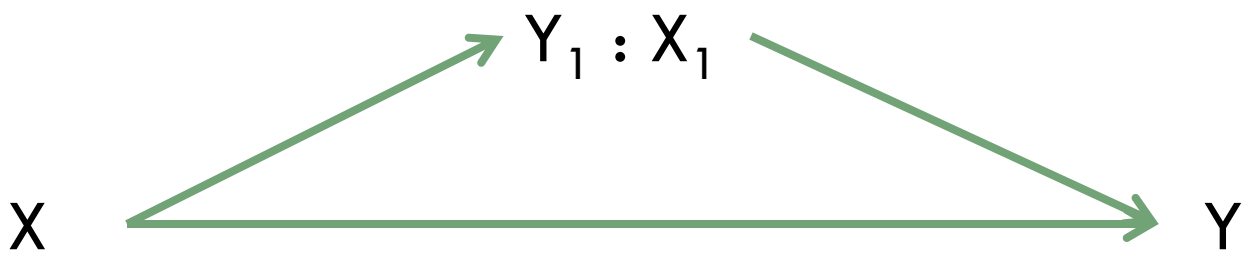
  Very weak interference ($2a(1+a^2P) \leq 1$) :

  Treat interference as noise – (IAN)

Interference Ch.: History (continued)

- Sason (2004): Symmetrical superposition to beat TDM – found part of optimal choice for α
- Etkin, Tse, Wang (2008): capacity to within 1 bit, good heuristical choice of $\alpha P = 1/a^2$
- C (2011): Noisebergs to compute Gaussian H+K region for Z IFCs
- C, Nair (2012,2013): Some progress on achievable region of symmetric Gaussian IFCs

Relay Channel



- The relay channel is said to be physically degraded if $p(y,y_1 \mid x,x_1)=p(y_1 \mid x,x_1)\, p(y \mid y_1,x_1)$.
- So Y is a degradation of the relay signal Y_1 .
- Theorem: $C = \sup_{p(x,x_1)} \min \{ I(X,X_1;Y_1), I(X;Y_1 \mid X_1)\}$

Wrap Up

Many new fronts:

Joint source and channel coding

Coding for channels with side information

Distributed source coding

Network strategies

Merging of Network Coding and Multi User IT

- Many thanks!

max@fee.unicamp.br