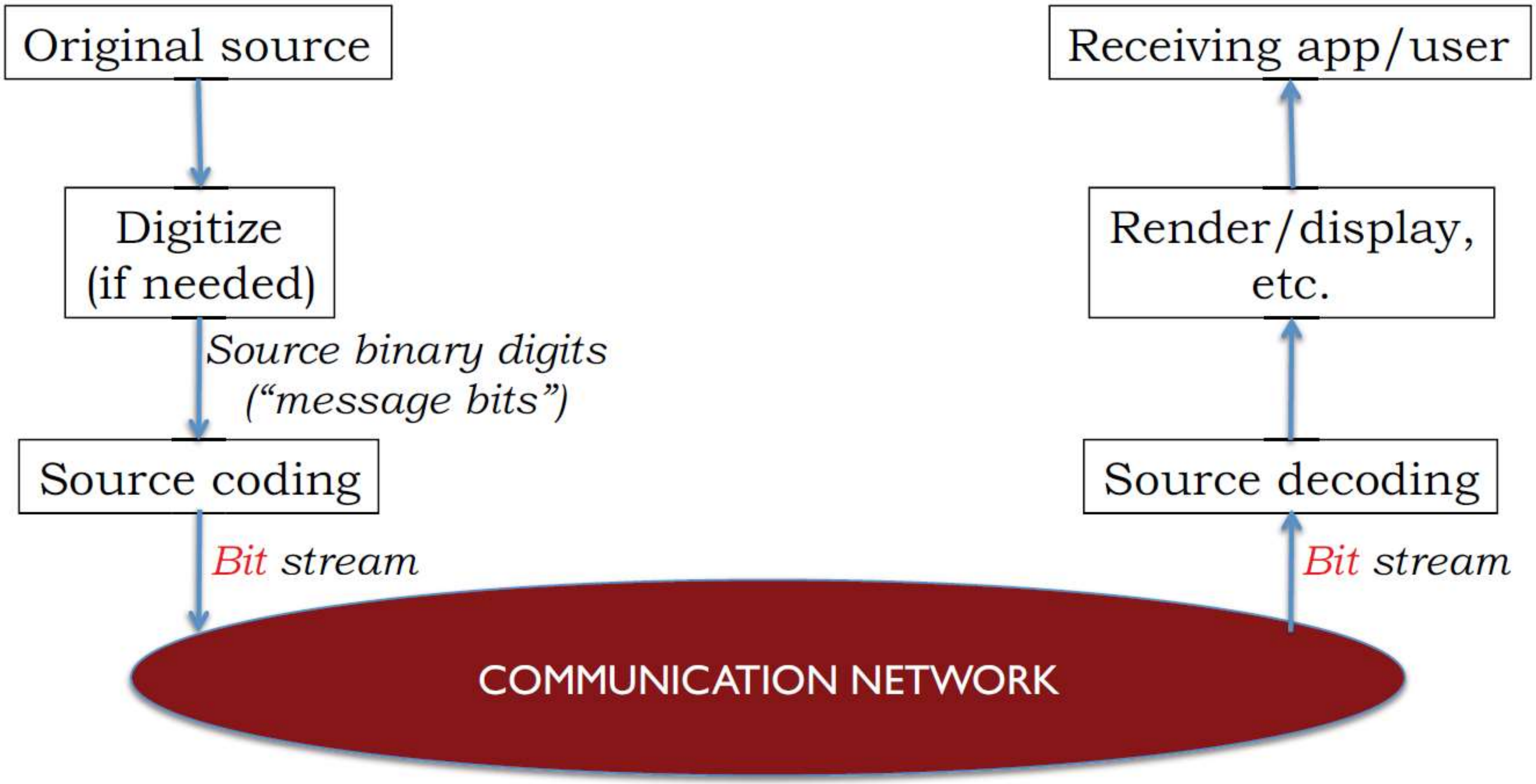
FUNDAMENTALS OF INFORMATION SCIENCE:

PART 3: CODING TECHNOLOGIES

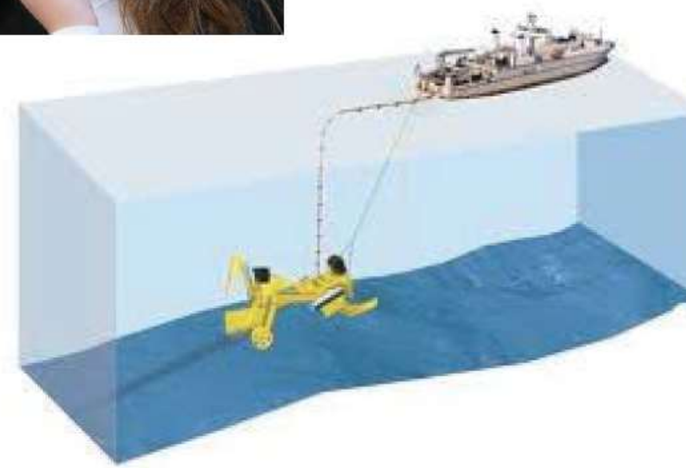
Shandong University
2024 Spring

Lecture 2.1: Communication Channels

Communication System



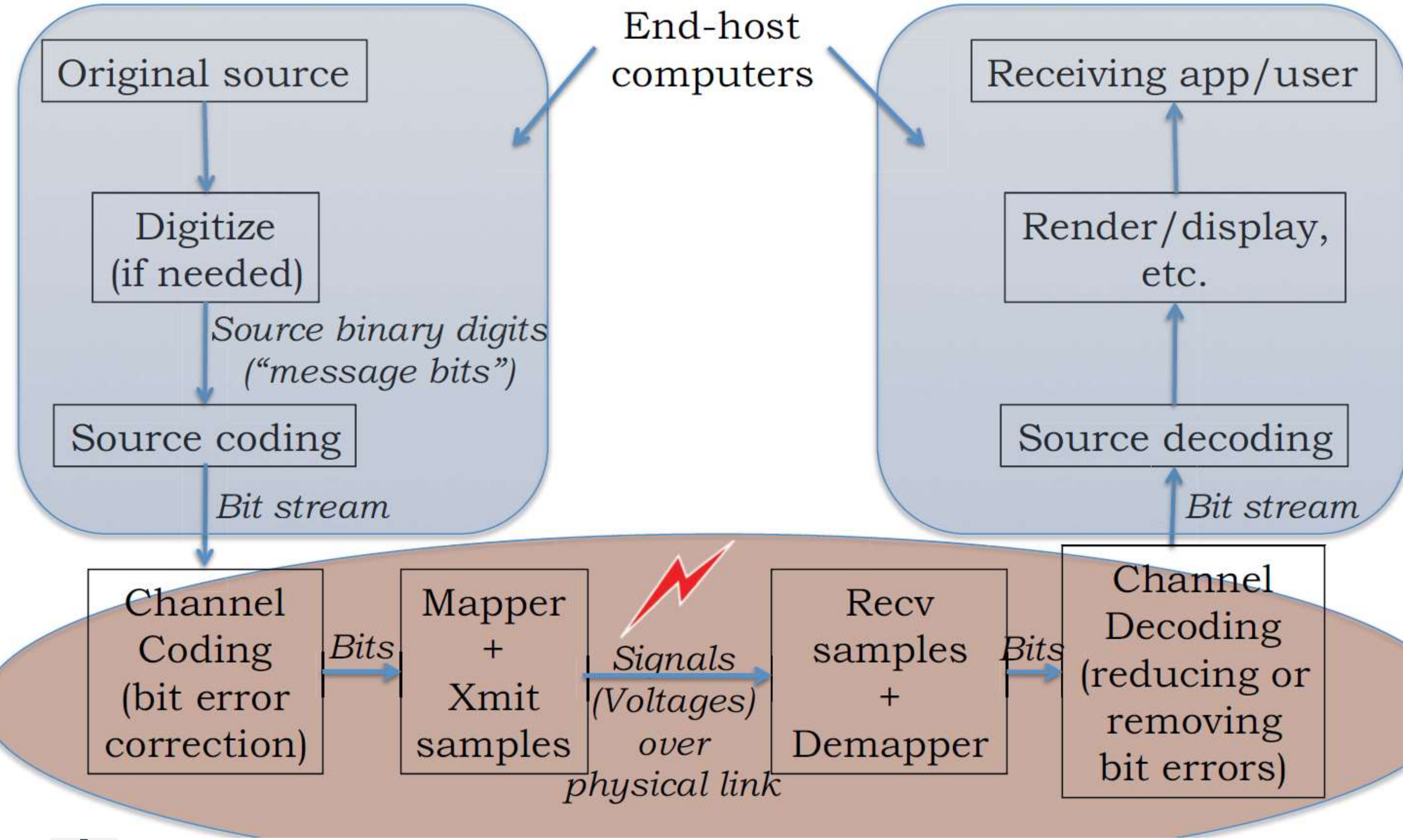
Physical Communication Links are Inherently Analog



Analog = continuous-valued, continuous-time

Voltage waveform on a cable
Light on a fiber, or in free space
Radio (EM) waves through the atmosphere
Acoustic waves in air or water
Indentations on vinyl or plastic
Magnetization of a disc or tape

Single Link Communication



Digital Signaling: Map Bits to Signals

Key Idea: "Code" or map or modulate the desired bit sequence onto a (continuous-time) analog signal, communicating at some bit rate (in bits/sec).

To help us extract the intended bit sequence from the noisy received signals, we'll map bits to signals using a fixed set of discrete values. For example, in a *bi-level signaling* (or *bi-level mapping*) scheme we use two "voltages":

V0 is the binary value "0"
V1 is the binary value "1"

If V0 = -V1 (and often even otherwise) we refer to this as ***bipolar*** signaling.

At the receiver, process and sample to get a "voltage"

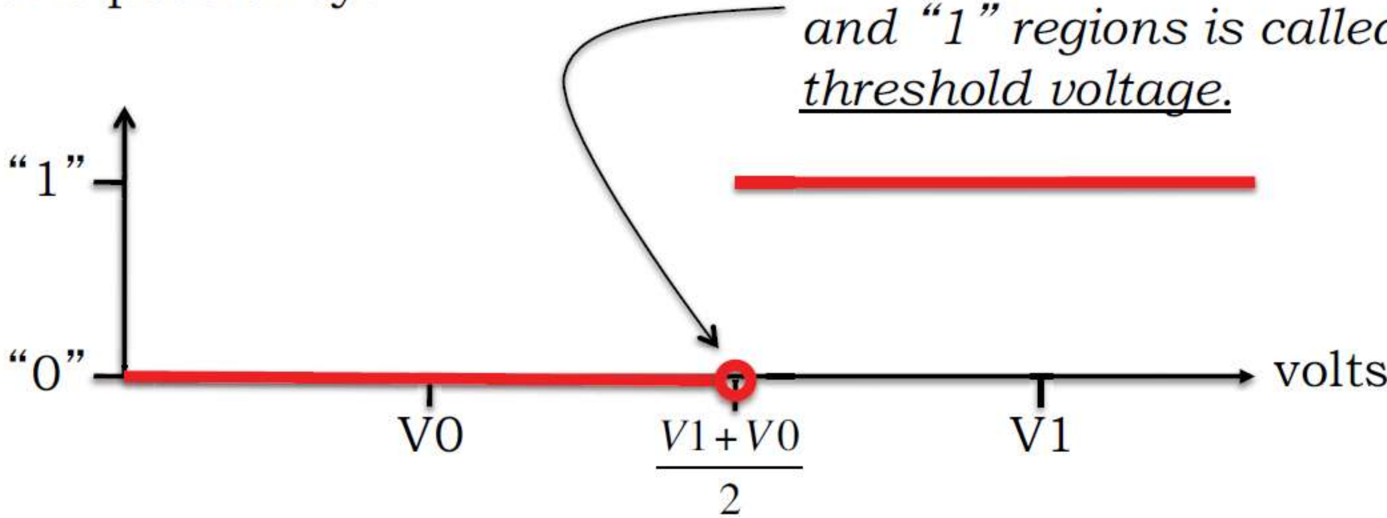
- Voltages near V0 would be interpreted as representing "0"
- Voltages near V1 would be interpreted as representing "1"
- If we space V0 and V1 far enough apart, we can tolerate some degree of noise --- but there will be occasional errors!

Digital Signaling: Receiving Signals

We can specify the behavior of the receiver with a graph that shows how incoming voltages are mapped to "0" and "1".

One possibility:

The boundary between "0" and "1" regions is called the <u>threshold voltage</u>.

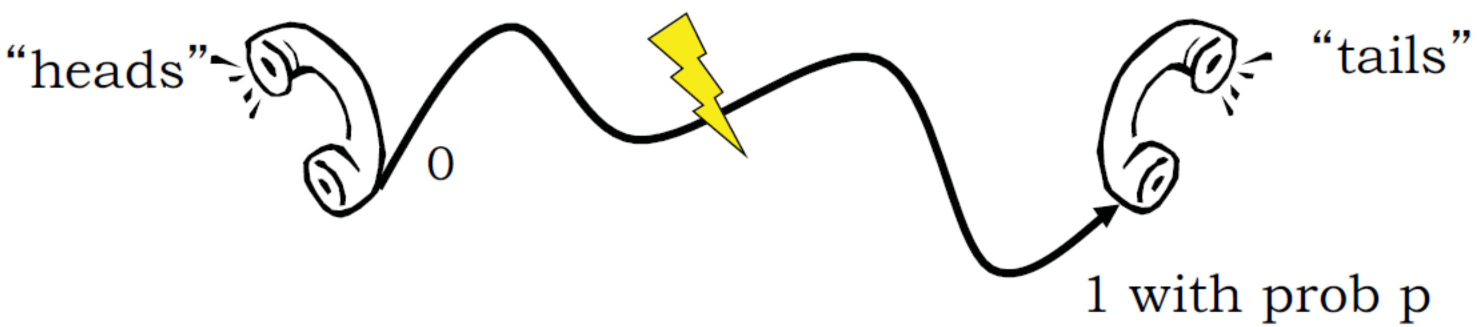


If received voltage between $V0$ & $\frac{V1+V0}{2}$ → "0", else "1"

Bit-In Bit-Out Model: Binary Symmetric Channel

Suppose that during transmission a "0" is turned into a "1" or a "1" is turned into a "0" with probability p, independently of transmissions at other times

This is a *binary symmetric channel* (BSC) --- a useful and widely used abstraction

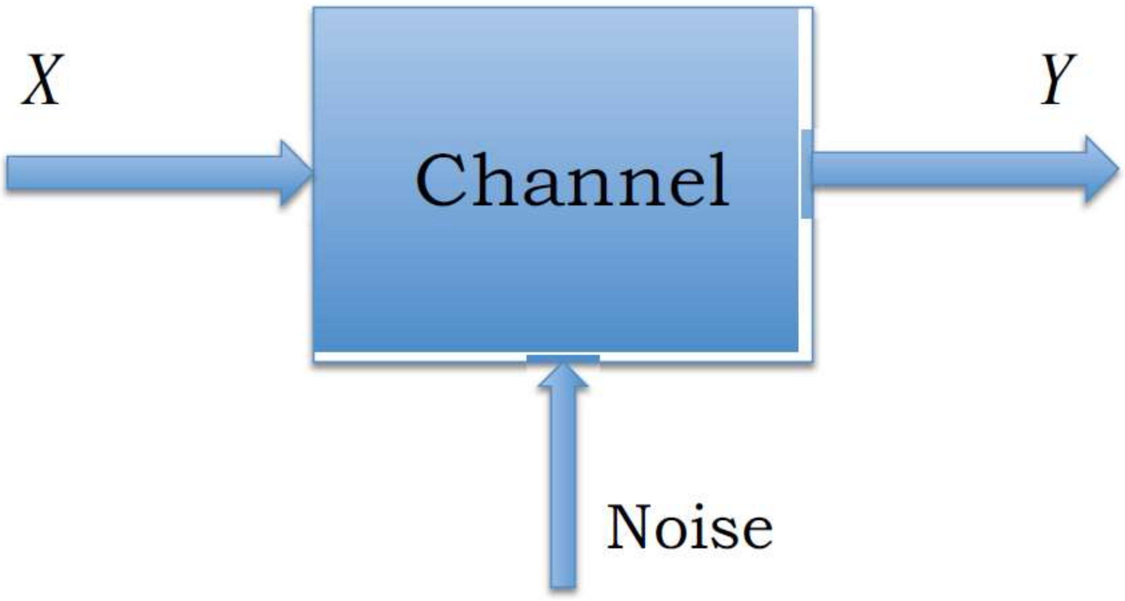


Mutual Information

$$I(X;Y) = H(X) - H(X \mid Y)$$

How much is our uncertainty about X reduced by knowing Y?

Evidently a central question in communication or, more generally, inference. Thank you, Shannon!



Conditional Entropy and Mutual Information

To compute conditional entropy:

$$H(X \mid Y = y_j) = \sum_{i=1}^{m} p(x_i \mid y_j) \log_2 \left(\frac{1}{p(x_i \mid y_j)} \right)$$

$$H(X \mid Y) = \sum_{i=1}^{m} H(X \mid Y = y_j) p(y_j)$$

$$H(X,Y) = H(X) + H(Y \mid X)$$
$$= H(Y) + H(X \mid Y)$$

because

$$p(x_i, y_j) = p(x_i) p(y_j \mid x_i)$$
$$= p(y_j) p(x_i \mid y_j)$$

$\boxed{I(X;Y) = I(Y;X)}$ $\Rightarrow$ mutual information is symmetric

Mutual Information of Binary Symmetric Channel (BSC)

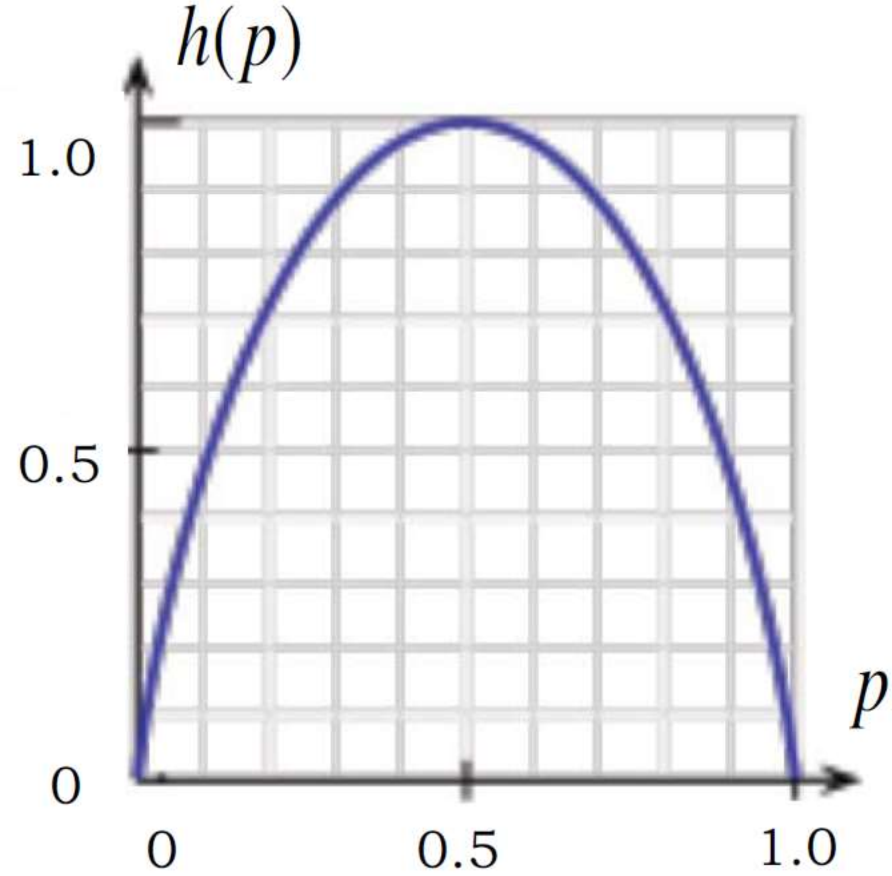
With probability p the input binary digit gets flipped before being presented at the output.

$$I(X;Y) = I(Y;X) = H(Y) - H(Y \mid X)$$
$$= 1 - H(Y \mid X=0)p_X(0) - H(Y \mid X=1)p_X(1)$$
$$= 1 - h(p)$$

Binary Entropy Function h(p)

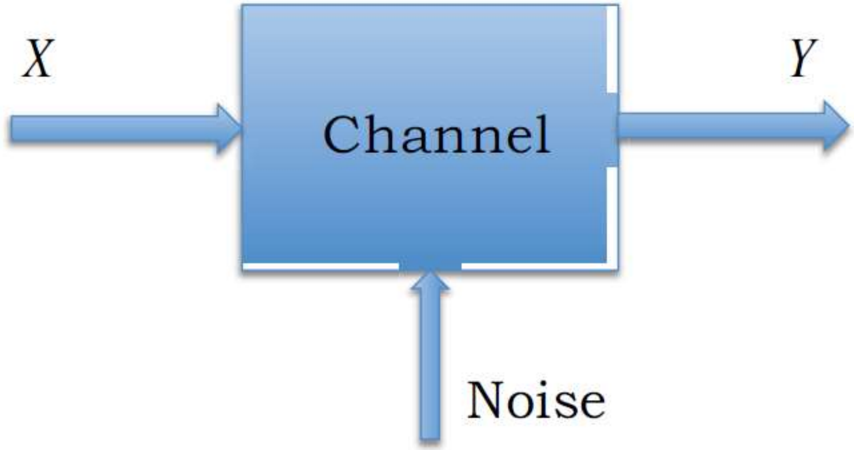
Heads (or C=1) with probability p

Tails (or C=0) with probability $1-p$



$$H(C) = -p\log_2 p - (1-p)\log_2(1-p) = h(p)$$

Channel Capacity



To characterize the *channel*, rather than the input and output, define

$$C = \max I(X;Y) = \max\{H(X) - H(X|Y)\}$$

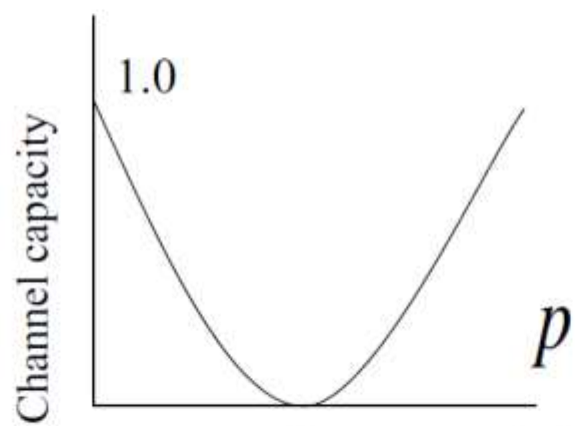
where the maximization is over all possible distributions of X.

This is the most we can expect to reduce our uncertainty about X through knowledge of Y, and so must be *the most information we can expect to send through the channel on average, per use of the channel*. Thank you, Shannon!

Capacity of Binary Symmetric Channel (BSC)

Easiest to compute as $C = \max\left\{H(Y) - H(Y \mid X)\right\}$, still over all possible probability distributions for X. The second term doesn't depend on this distribution, and the first term is maximized when 0 and 1 are equally likely at the input. So invoking our mutual information example earlier:

➔ $\boxed{C = 1 - h(p)}$



What channel capacity tells us about how fast and how accurately we can communicate

…

Why Channel Capacity

- Look at communication systems:
  Landline Phone, Radio $\rightarrow$ TV, Cellphone $\rightarrow$ Smartphone, WiFi
- Communication is very tied to specific source
- To break this tie, Shannon propose to focus on information, then computation
- First ask the question: what is the fundamental limit
- Then ask how to achieve this limit (took 60 years to get there! but huge success)
- All communication system are designed based on the principle of IT

Shannon’ s Secret of Success

- Start with simple model, then complicated

“Stylized” Models

- Let the code length goes to infinity, then back
- Study random coding, prove the feasibility

“Asymptotic is the first term in Taylor series expansion, and theory is the first term in the Taylor series of practice.”

- Tom Cover, 1990

Shannon’s Secret of Success

- Start with simple model, then complicated

“Stylized” Models

- Let the code length goes to infinity, then back

- Study random coding, prove the feasibility

“Asymptotic is the first term in Taylor series expansion, and theory is the first term in the Taylor series of practice.”

- Tom Cover, 1990

Channel Capacity: Intuition

$$C = \log \#\{ \text{ of identifiable inputs by passing through the channel with low error}\}$$

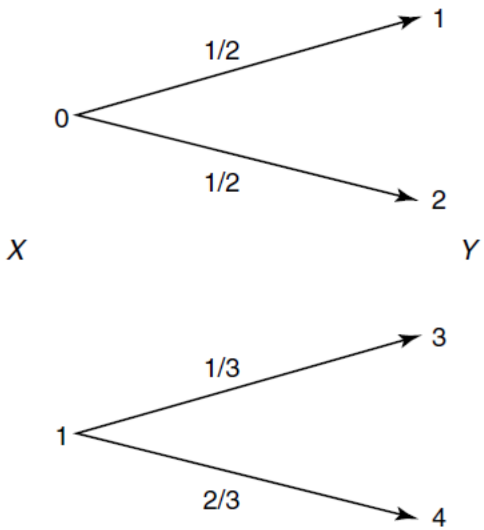
Shannon's second theorem:

$$\text{“information” channel capacity} = \text{“operational” channel capacity}$$

Binary Noiseless Channel

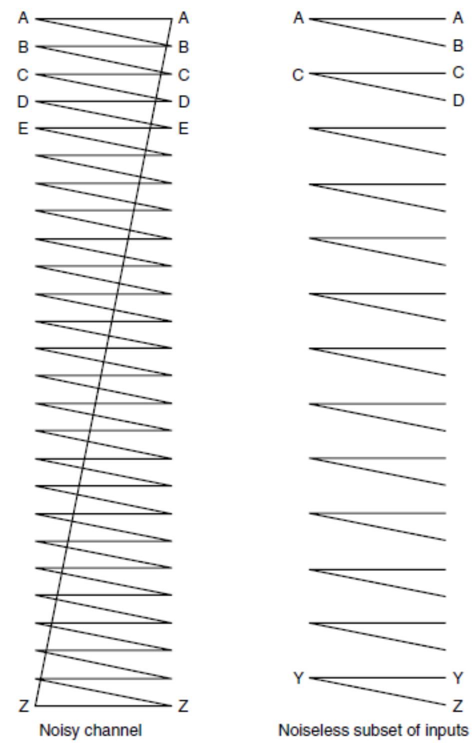
$C = \log 2 = 1$ bit

Noisy Channel with Non-Overlapping Outputs



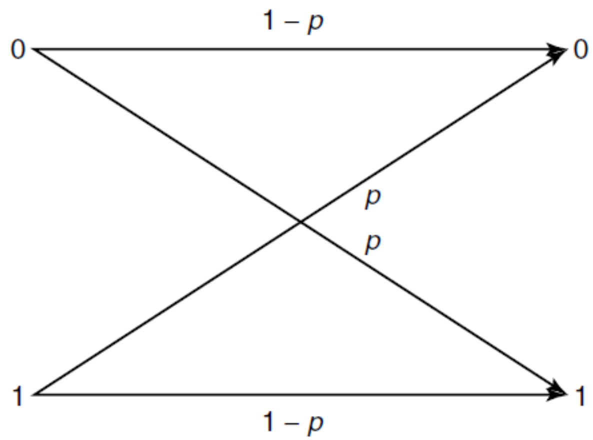
$C = \log 2 = 1$ bit

Noisy Typewriter



$$C = \log 13 \text{ bits}$$

Binary Symmetric Channel



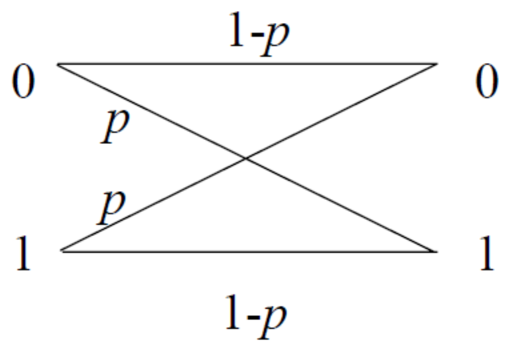
$C = 1 - H(p)$ bits.

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

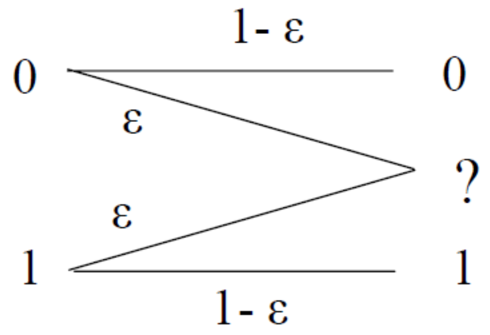
CD-ROM read channel

More Channels

- Binary symmetric channel BSC(p)

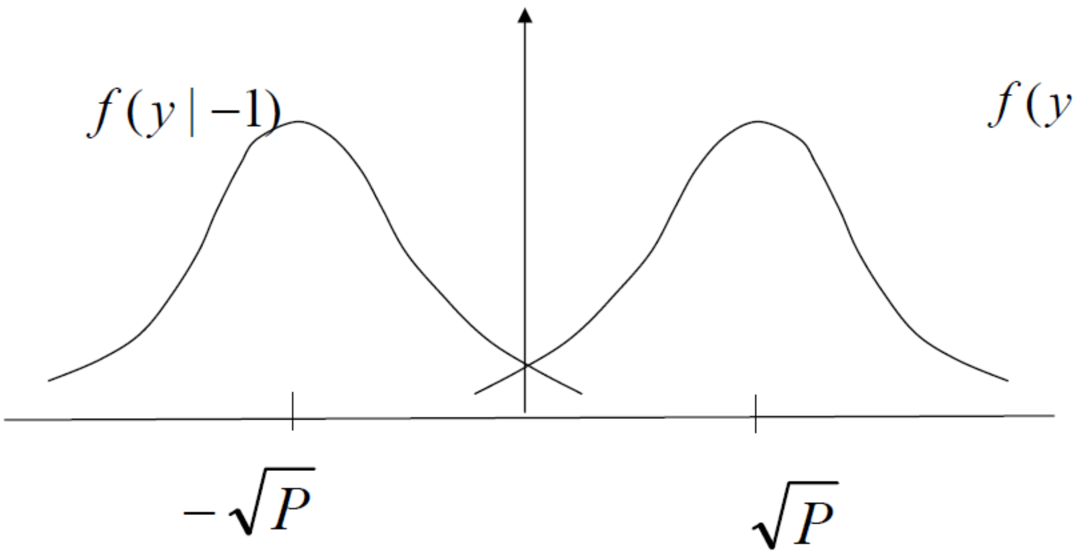


- Binary erasure channel BEC(ε)



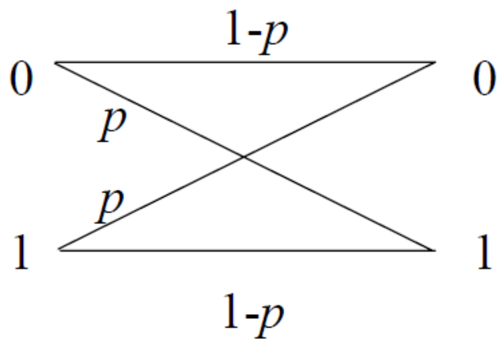
More Channels

- Additive white Gaussian noise channel AWGN

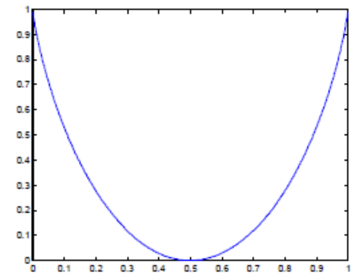


Channels and Capacities

- Binary symmetric channel BSC(p)

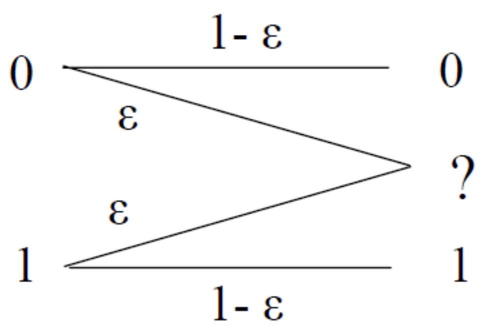


$$C = 1 - H_2(p)$$

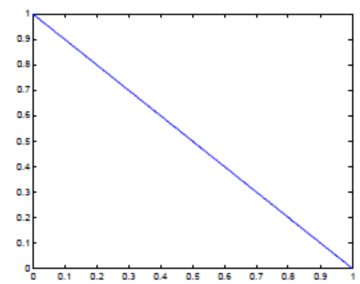


$$H_2(p) = -p\log_2 p - (1-p)\log_2(1-p)$$

- Binary erasure channel BEC(ε)

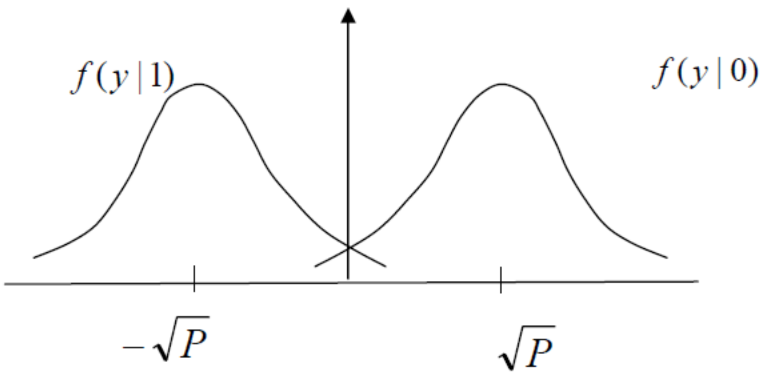


$$C = 1 - \varepsilon$$

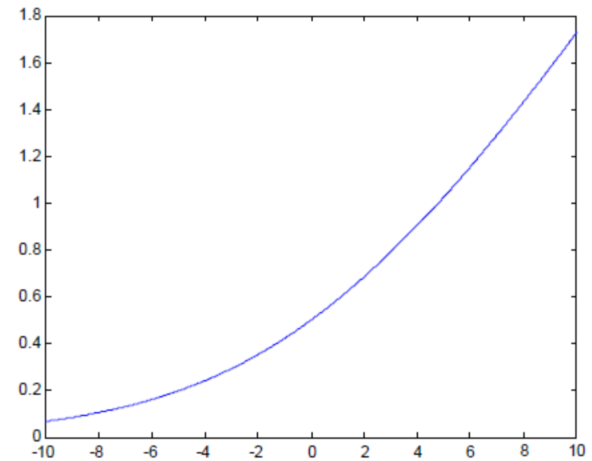


Channels and Capacities

- Additive white Gaussian noise channel AWGN



$$C = \tfrac{1}{2}\log_2\left(1+\frac{P}{\sigma^2}\right)$$

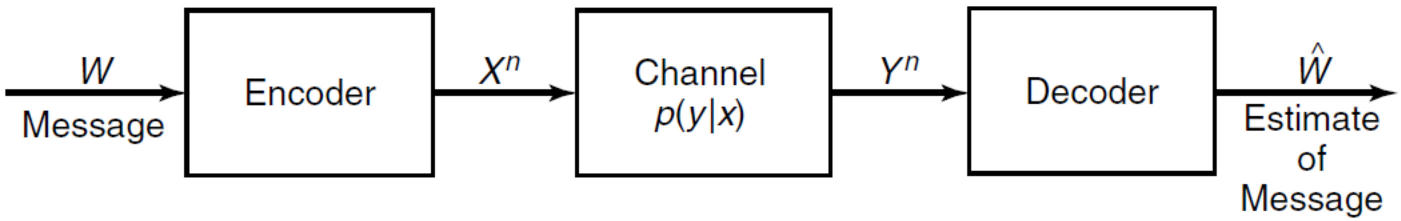


Discrete Memoryless Channel (DMC)

- Discrete channel:
  - input alphabet: $\mathcal{X}$
  - output alphabet: $\mathcal{Y}$
  - probability transition matrix $p(y|x)$

- Memoryless channel:
  the probability distribution of the output depends only on the inputs at that time

Communication System Model



- $X^n = [X_1, \ldots, X_n]$
- $Y^n = [Y_1, \ldots, Y_n]$
- channel: $p(y|x)$: probability of observing y given input symbol x

Communication System Model

- Symbols from some finite alphabet are mapped into some sequence of the channel symbols
- Output sequence is random but has a distribution that depends on the input sequences
- From output sequence, we try to recover the transmitted message
- Each possible input sequences induces several possible outputs, and hence inputs are confusable
- Can we choose a “non-confusable” subset of input sequences?

Duality

- Data compression: we remove all the redundancy in the data to form the most compressed version possible

- Data transmission: we add redundancy in a controlled manner to combat errors in the channel

Summary

- Channel capacity:

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

$$\text{intuition: } C = \log\{\#\text{of distinguishable inputs}\}$$

- DMC (discrete memoryless channel)

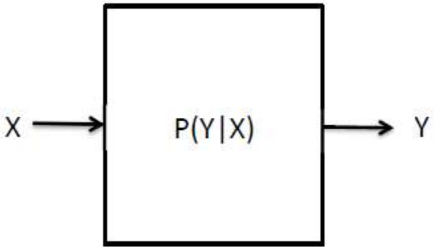
Lecture 2.2: Channel Coding Theorem

Information Channel Capacity

For discrete memoryless channel (DMC)

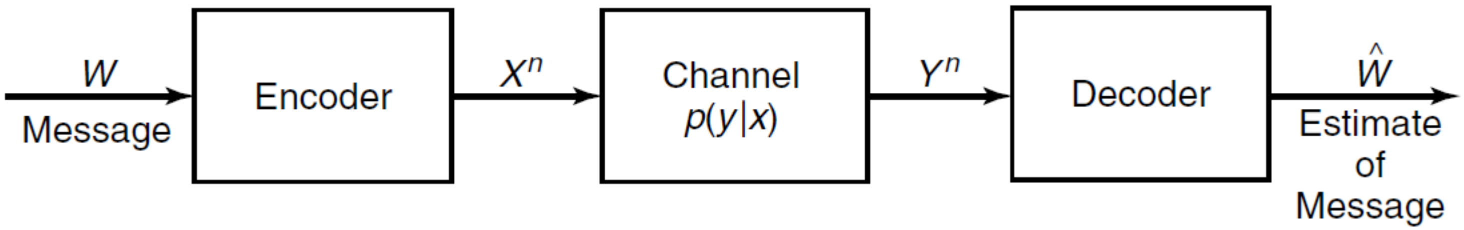
$$C = \max_{p(x)} I(X;Y)$$

- $C \geq 0$ since $I(X;Y) \geq 0$, $C \leq \log|\mathcal{X}|$, $C \leq \log|\mathcal{Y}|$



Discrete: $\mathcal{X}$, $\mathcal{Y}$ discrete
Memoryless: $p(Y^n|X^n) = \prod_{i=1}^{n} p(y_i|x_i)$

Communication System Model



- $W \in \{1, 2, \ldots, M\}$: source message
- X^n: sequence of channel symbols
- Y^n: output sequence, $Y^n \sim p(y^n|x^n)$
- $\hat{W}$: recovered message, according to decoding function $\hat{W} = g(Y^n)$

Fundamental Question

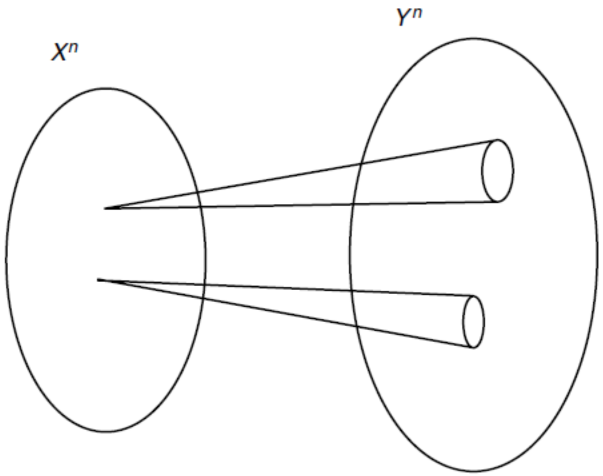
- How fast can we transmit information over a communication channel?
- suppose a source sends r messages per second, and the entropy of a message is H bits per message, information rate is $R = rH$ bits/second
- intuition: as R increases, error will increase
- surprisingly, error can be nearly zero, as long as

$$R < \underbrace{R_{\max}}_{\text{“operational channel capacity”}}$$

- Shannon showed $R_{\max} = C$

Basic Idea

- For large block length, every channel looks like the noisy type writer channel
- Channel has a subset of inputs that produce “disjoint” sequences at the output



Code Rate

- Rate of an (M, n) code is

$$R = \frac{\log M}{n} \text{ bit per transmission}$$

- On the other hand, we usually write

$$M = \lceil 2^{nR} \rceil$$

Assumption about the Channel

- Transmit large block length: n over n transmissions
- DMC

$$p(y^n|x^n) = \prod_{i=1}^{n} p(y_i|x_i)$$

- channel without feedback:

$$p(y_k|x^k, y^{k-1}) = p(y_k|x_k), k = 1, \ldots, n$$

Error Probability

- Conditional probability of error

$$\lambda_i = P\{g(Y^n) \neq i | X^n = x^n(i)\}$$

- Maximal probability of error

$$\lambda^{(n)} = \max_{i=1}^{m} \lambda_i$$

- Average probability of error

$$P_e^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i$$

- $P_e^{(n)} \leq \lambda^{(n)}$
- If W uniform distributed,

$$P_e^{(n)} = P\{W \neq g(Y^n))\}$$

Achievable Rate

A rate R is achievable:

if exists a sequence of $\lceil 2^{nR}, n \rceil$ codes such that $\lambda^{(n)} \to 0$ as $n \to 0$.

Channel Coding Theorem

Theorem. *(Shannon, 1948)*
For a DMC

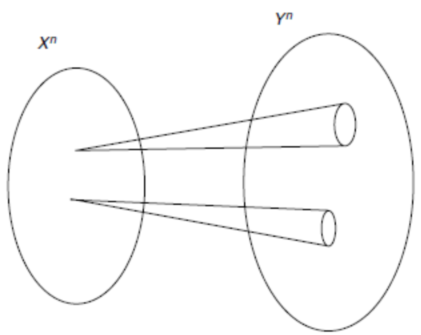
1. *all rates below capacity* $R < C$ *are achievable.*

2. *Converse: any sequence of* $(2^{nR}, n)$ *codes with* $\lambda^{(n)} \to 0$ *must have* $R \leq C$.

Reliable communication over noisy channel is possible!

Proof Idea

- for each (typical) X^n, there are $\approx 2^{nH(Y|X)}$ possible Y^n
- Total number of (typical) Y^n is $2^{nH(Y)}$
- Total number of disjoint inputs should be $2^{n(H(Y)-H(Y|X))} = 2^{nI(X;Y)}$
- To formalize these ideas, we need "joint typical sequences"

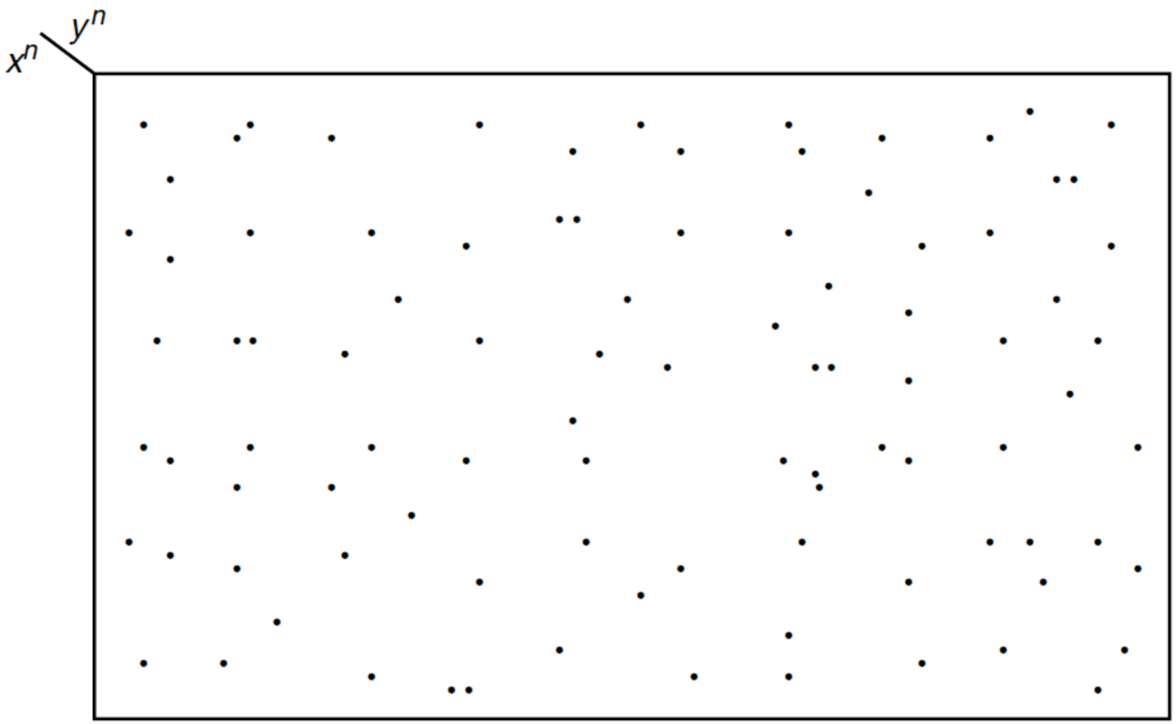


Joint Typical Sequence

- Associate a "fan" with each codeword X^n
- We decode Y^n as the ith index if the codeword $X^n(i)$ is "joint typical" with Y^n
- Set $A_\epsilon^{(n)}$ of jointly typical sequences $\{(x^n, y^n)\}$ is

$$
\begin{aligned}
A_\epsilon^{(n)} = &\{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \\
&|-\frac{1}{n}\log p(x^n) - H(X)| < \epsilon \\
&|-\frac{1}{n}\log p(y^n) - H(Y)| < \epsilon \\
&|-\frac{1}{n}\log p(x^n, y^n) - H(X,Y)| < \epsilon\}
\end{aligned}
$$

Joint Typical Sequence



$2^{nH(X)}$ typical X^n, $2^{nH(Y)}$ typical Y, not all pairs of typical X^n and Y^n are also jointly typical. Any randomly chosen pair is jointly typical is $2^{-nI(X;Y)}$.

Joint AEP

- Let (X^n, Y^n) be sequences of length n drawn i.i.d. according to $p(x^n, y^n) = \prod_{i=1}^n p(x_i, y_i)$. Then

1. $P((X^n, Y^n) \in A_\epsilon^{(n)}) \to 1$ as $n \to \infty$
2. $|A_\epsilon^{(n)}| \le 2^{n(H(X,Y)+\epsilon)}$
3. If $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)$, then

$$P\{(\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\} \le 2^{-n(I(X;Y)-3\epsilon)}$$

For sufficient large n,

$$(1-\epsilon)2^{n(H(X,Y)-\epsilon)} \le |A_\epsilon^{(n)}|$$

$$P\{(\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\} \ge (1-\epsilon)2^{-n(I(X;Y)+3\epsilon)}$$

Channel Coding Theorem

Theorem. *(Shannon, 1948)*
For a DMC

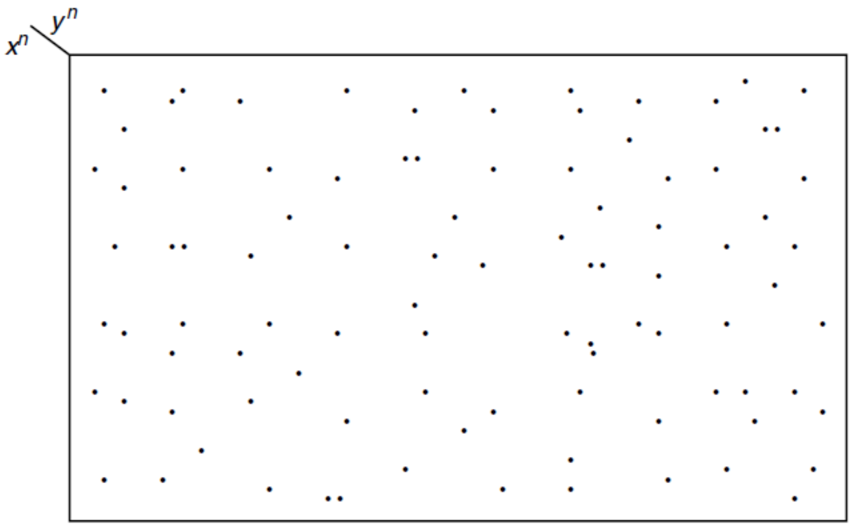
1. *all rates below capacity* $R < C$ *are achievable.*

2. *Converse: any sequence of* $(2^{nR}, n)$ *codes with* $\lambda^{(n)} \to 0$ *must have* $R \leq C$.

Reliable communication over noisy channel is possible!

Joint Typical Decoding

- Decoder find $\hat{W}$ if $(X^n(\hat{W}), Y^n)$ is jointly typical
- No confusion: no more than $X^n(\hat{W})$ jointly typical with Y^n



Proof for Achievability

- calculate the probability of error averaged over all codes randomly generated according to $p(x)$

- Average P_e does not depend on which index was sent

- For typical X^n, two type of errors

(a) (X^n, Y^n) not jointly typical
(b) $(\tilde{X}^n, Y^n)$ is typical, but $\tilde{X}^n \neq X^n$

- Use AEP to bound (a) and (b)

- Conditional probability of error

$$\lambda_i = P\{g(Y^n) \neq i | X^n = x^n(i)\}$$

Proof for Achievability

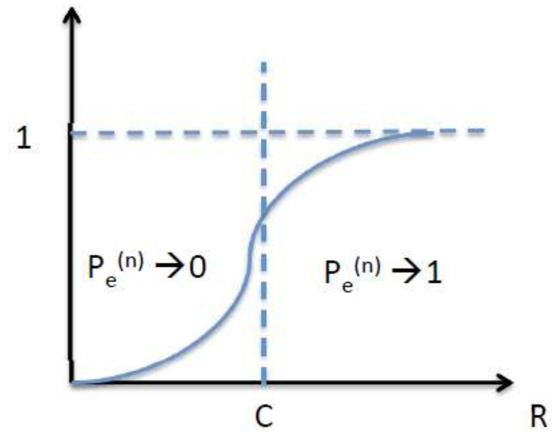
Define the following events:

$$E_i = \{ (X^n(i), Y^n) \text{ is in } A_\epsilon^{(n)}\}, \quad i \in \{1, 2, \ldots, 2^{nR}\},$$

$$\begin{aligned}
\Pr(\mathcal{E}|W = 1) &= P\left(E_1^c \cup E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}}|W = 1\right) \\
&\le P(E_1^c|W = 1) + \sum_{i=2}^{2^{nR}} P(E_i|W = 1), \\
&\le \epsilon + \sum_{i=2}^{2^{nR}} 2^{-n(I(X;Y)-3\epsilon)}
\end{aligned}$$

Proof for Converse

- Use Fano's inequality to lower bound P_e



Proof for Converse

Lemma 7.9.1 (*Fano's inequality*) *For a discrete memoryless channel with a codebook $\mathcal{C}$ and the input message W uniformly distributed over 2^{nR}, we have*

$$H(W|\hat{W}) \leq 1 + P_e^{(n)} nR. \tag{7.89}$$

Proof: *Converse to Theorem 7.7.1 (Channel coding theorem).*

$$\begin{aligned}
nR &\overset{\text{(a)}}{=} H(W) \\
&\overset{\text{(b)}}{=} H(W|\hat{W}) + I(W;\hat{W}) \\
&\overset{\text{(c)}}{\leq} 1 + P_e^{(n)} nR + I(W;\hat{W}) \\
&\overset{\text{(d)}}{\leq} 1 + P_e^{(n)} nR + I(X^n;Y^n) \\
&\overset{\text{(e)}}{\leq} 1 + P_e^{(n)} nR + nC,
\end{aligned}$$

Implication of the Theorem

- It shows that there exist good codes with exponentially small probability of error for long block length
- it does not provide a way to construct the best codes
- random code, without structure, very difficult to code (look-up table)
- property of capacity achieving codes
- example of capacity achieving: noisy typewriter
- new capacity achieving code: polar codes (2009)

Asymptotically Error-free at R < C

Shannon showed that one can theoretically transmit information (i.e., message bits) at an average rate $R < C$ per use of the channel, with arbitrarily low error.

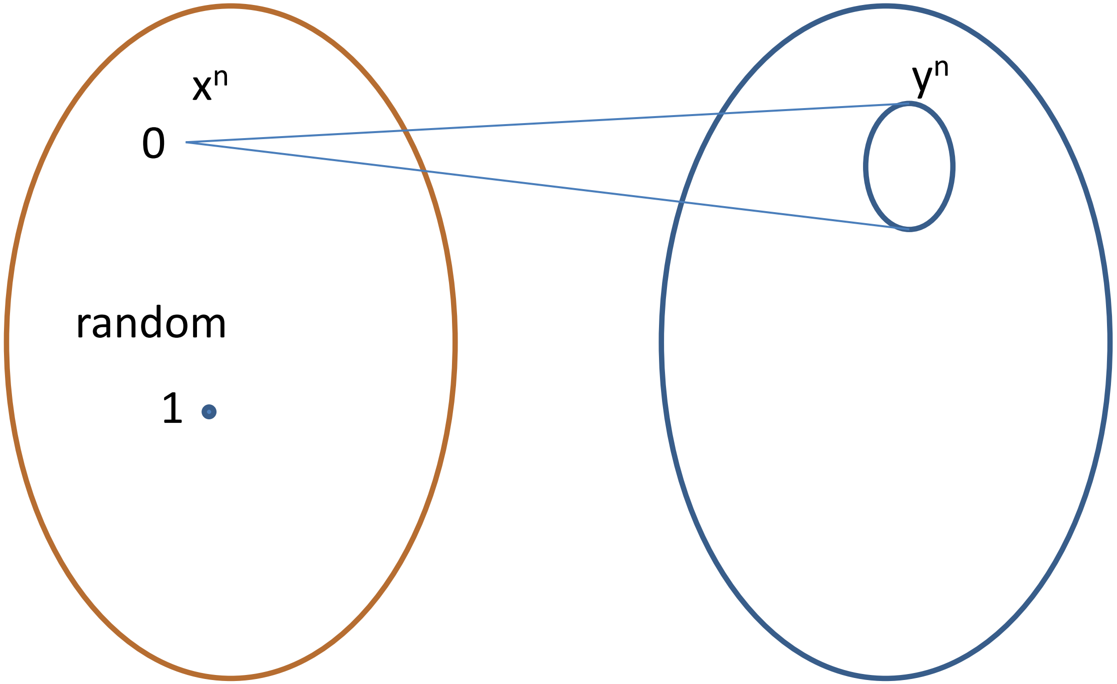
(He also showed the converse, that transmission at an average rate $R \geq C$ incurs an error probability that is lower-bounded by some positive number.)

The secret: Encode blocks of k message bits into n-bit codewords, so $R = k / n$, with k and n very large.

Encoding blocks of k message bits into n-bit codewords to protect against channel errors is an example of

channel coding

Achievability



Error 1: $x^n(0)\ y^n$ Not Typical

$$< \epsilon$$

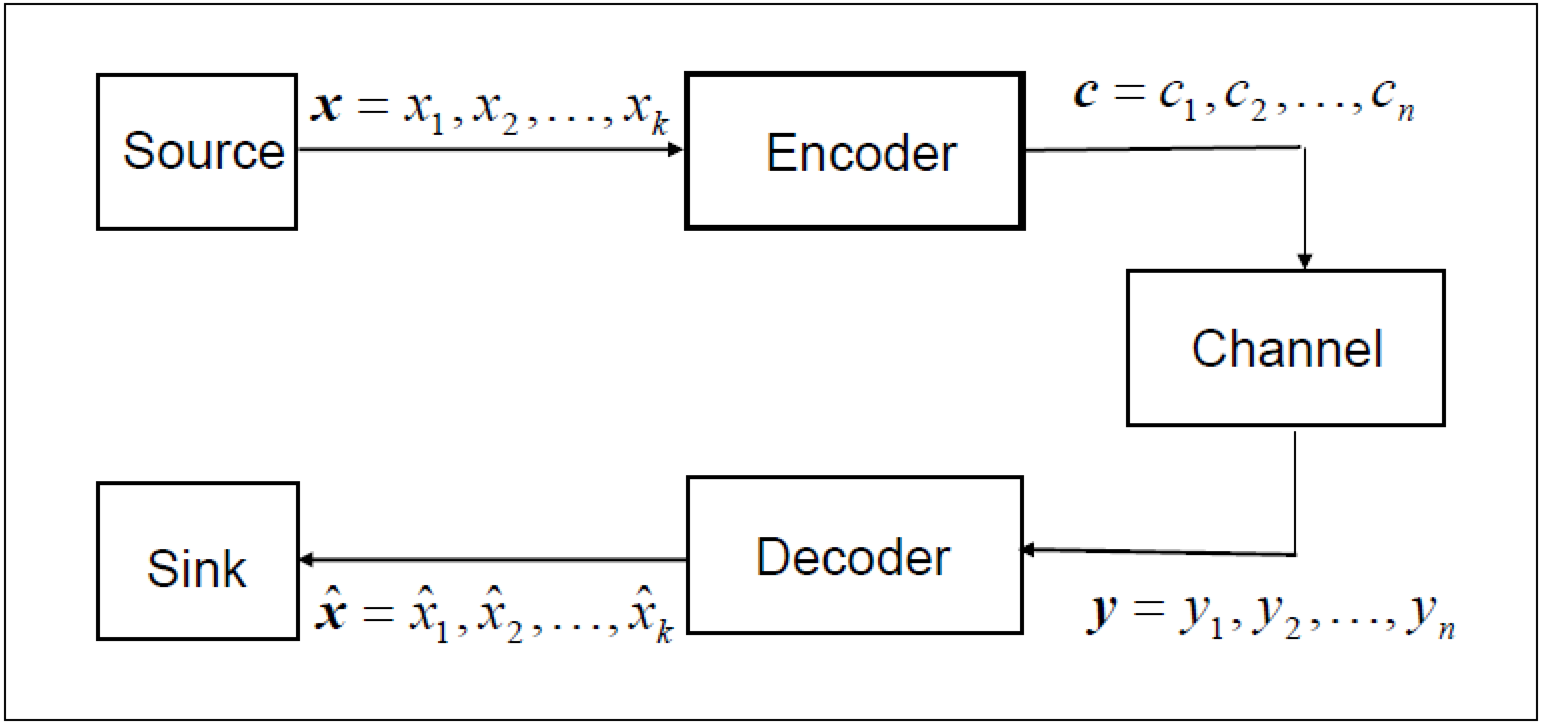
Error 2: $x^n(1)\ y^n$ Typical

$$\leq 2^{-n(I(X;Y)-3\epsilon)}$$

Total Error $\leq \epsilon + \sum_{i=2}^{2^{nR}} 2^{-n(I(X;Y)-3\epsilon)}$

Channel Coding

We use a **code** to communicate over the noisy channel.



Code rate: $R = \dfrac{k}{n}$