Information in Processes I

Reading for this lecture:

EIT, Secs. 5.1-5.6 and 7.1-7.7.

Interactive Labs:
Processes, Word Distributions, and Models

Previously: Entropy motivated as a measure of surprise.

Today:

How to compress a process:

Can't do better than H(X)

(Shannon's First Theorem)

How to communicate a process's data:

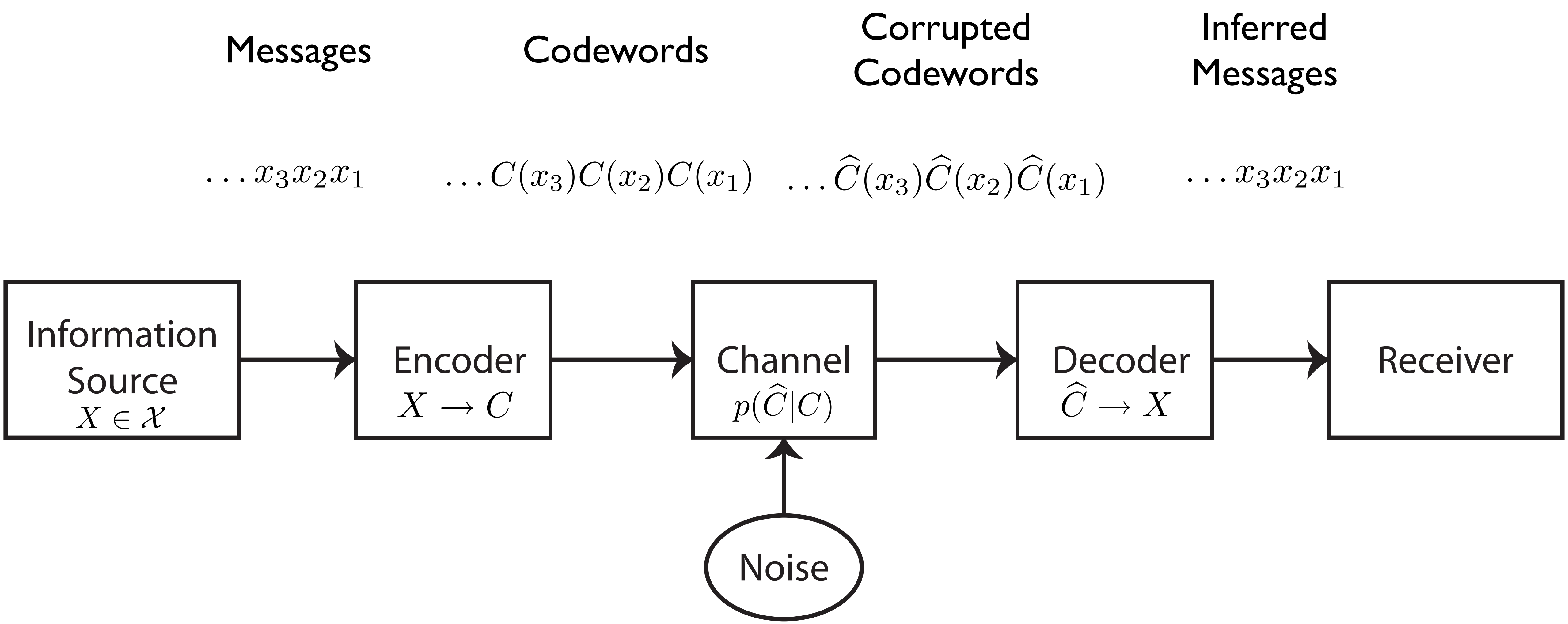
Can transmit error-free at rates up to channel capacity

(Shannon's Second Theorem)

Both results give *operational meaning* to entropy.

Information in Processes ...

Communication channel:



Information in Processes ...

Information source:

Random variable: $X \sim p(x)$

Message: $x \in \mathcal{X}$

Code:

Encoding alphabet:

$$\mathcal{D} = \{0, 1\} \qquad \mathcal{D}^+ = \bigcup_{L=1}^{\infty} \mathcal{D}^L$$

Codeword:

$$C(x) \in \mathcal{D}^+$$

Codeword length:

$$l(x) = ||C(x)||$$

Codebook:

A codebook maps messages into codewords:

$$C : \mathcal{X} \to \mathcal{D}^+$$

Expected length:

$$\langle l(x) \rangle = \sum_{x \in \mathcal{X}} p(x) l(x)$$

Code rate: $R(C) = \langle l(x) \rangle$ bits per message

Information in Processes ...

Codebook ...

Result: An encoding of an information source

$$\ldots x_3 x_2 x_1 \quad \longrightarrow \quad \ldots C(x_3) C(x_2) C(x_1)$$

Compression, if

$$R(C) < \log_2 |\mathcal{X}|$$

Information in Processes ...

Example: $\mathcal{X} = \{a, b, c, d\}$

$X \sim p(x)$

Distribution:
$$\begin{aligned} p(a) &= \tfrac{1}{2} \\ p(b) &= \tfrac{1}{4} \\ p(c) &= \tfrac{1}{8} \\ p(d) &= \tfrac{1}{8} \end{aligned}$$

$$H(X) = 1.75 \text{ bits}$$

Codebook:
$$\begin{aligned} C(a) &= 0 \\ C(b) &= 10 \\ C(c) &= 110 \\ C(d) &= 111 \end{aligned}$$

$$R(C) = 1.75 \text{ bits}$$

Compression! $R(C) < \log_2 |\{a, b, c, d\}| = \log_2 4 = 2 \text{ bits}$

Information in Processes ...

Kinds of codes: How to decode?

Nonsingular code: $x_i \neq x_j \Rightarrow C(x_i) \neq C(x_j)$

Extension of a code: C^+

Source word: $x^L = x_1 x_2 \ldots x_L$

$$C(x^L) = C(x_1)C(x_2) \ldots C(x_L)$$

Uniquely decodable code: $C^+ \text{ nonsingular } \forall x^L \in \mathcal{X}^L$

Prefix code:

No $C(x_i)$ prefix of $C(x_j), \ i \neq j$

Determine codewords directly, no long look-ahead

Uniquely decodable

Kinds of codes ...

Example (continued):

Codebook:
$$\begin{aligned} C(a) &= 0 \\ C(b) &= 10 \\ C(c) &= 110 \\ C(d) &= 111 \end{aligned}$$

Encoding:

$$acdbac \rightarrow 0110111100110$$

Decoding:

$$0110111100110 \rightarrow \overbrace{0}^{a}\ \overbrace{110}^{c}\ \overbrace{111}^{d}\ \overbrace{10}^{b}\ \overbrace{0}^{a}\ \overbrace{110}^{c}$$

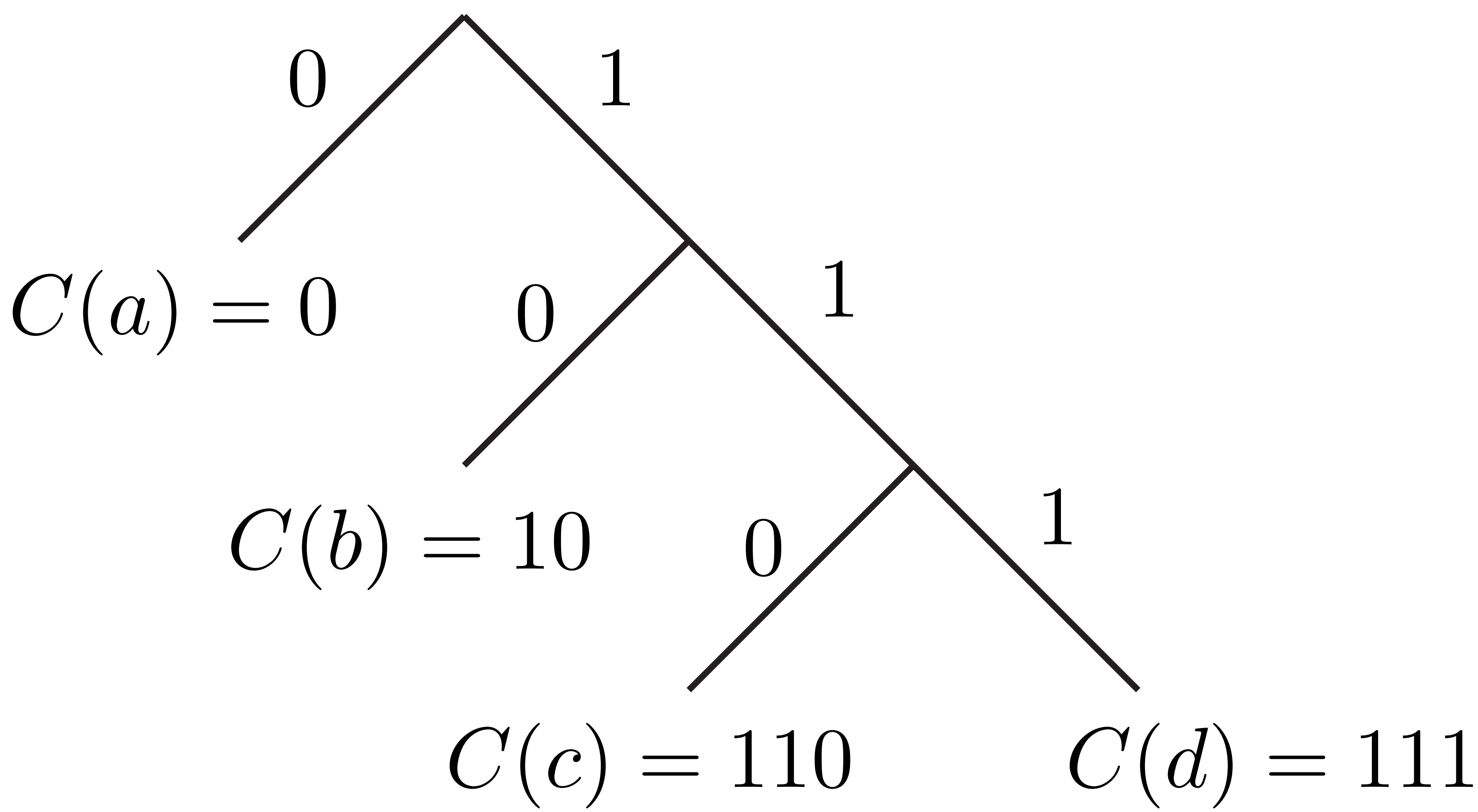
$$0110111100110 \rightarrow acdbac$$

A prefix code.

Information in Processes ...

Example (continued):

Prefix code corresponds to a tree of codewords.



$$
\begin{aligned}
C: \sum &= 2^{-l(a)} + 2^{-l(b)} + 2^{-l(c)} + 2^{-l(d)} \\
&= 2^{-1} + 2^{-2} + 2^{-3} + 2^{-3} \\
&= 1
\end{aligned}
$$

Information in Processes ...

Given a source: How to construct a prefix code?

Kraft inequality:

Codeword lengths in prefix codebook satisfy

$$\sum_{x \in \mathcal{X}} 2^{-l(x)} \leq 1$$

Optimal codes:

How are codeword lengths related to message probabilities?

Given an information source, find codebook such that

1. Minimize expected code length:

$$R = \langle l(x) \rangle = \sum_{x \in \mathcal{X}} p(x) l(x)$$

2. Subject to constraint of decodability:

$$\sum_{x \in \mathcal{X}} 2^{-l(x)} \leq 1$$

Information in Processes ...

Optimal codes ...

Satisfy constraint using Lagrange multiplier λ:

$$J = \sum_{x \in C} p(x) l(x) + \lambda \left(\sum_{x \in C} 2^{-l(x)} \right)$$

For each $x \in \mathcal{X}$:

$$\frac{\partial J}{\partial l} = 0 = p(x) - \lambda 2^{-l(x)} \log_e 2 \quad \text{Or} \quad 2^{-l(x)} = \frac{p(x)}{\lambda \log_e 2}$$

Constraint: $\lambda = 1/\log_e 2$

Thus, optimal codebook has lengths:

$$l(x) = -\log_2 p(x)$$

And, average length:

$$\langle l(x) \rangle = H(X)$$

Information in Processes ...

Optimal codes ...

Not implementable:

$-\log p(x_i)$ is not an integer length!

Any prefix code: $\langle l(x_i) \rangle \geq H(X)$

$l(x_i) = \lceil -\log_2 p(x_i) \rceil$ Shannon coding

Use (say) Shannon-Fano or Huffman code, then have

$$H(X) \leq \langle l(x) \rangle \leq H(X) + 1$$

Can be large cost!

Information in Processes ...

Source $\sim P(x)$, but you or your codebook uses $Q(x)$.

Use incorrect probability model for code construction: $Q \neq P$

$$\langle l(x) \rangle = H(X) + \mathcal{D}(P||Q)$$

Data Compression Theorem (Shannon's First Theorem):

$$R(C) \geq H(X)$$

Cannot compress source below its entropy rate.

Operational meaning of entropy:

Fundamental limit on compression.

How much can you compress?

Redundancy of a random variable: (identity codebook)

$$\mathcal{R} = \log_2 |\mathcal{X}| - H(X)$$

Example (continued):

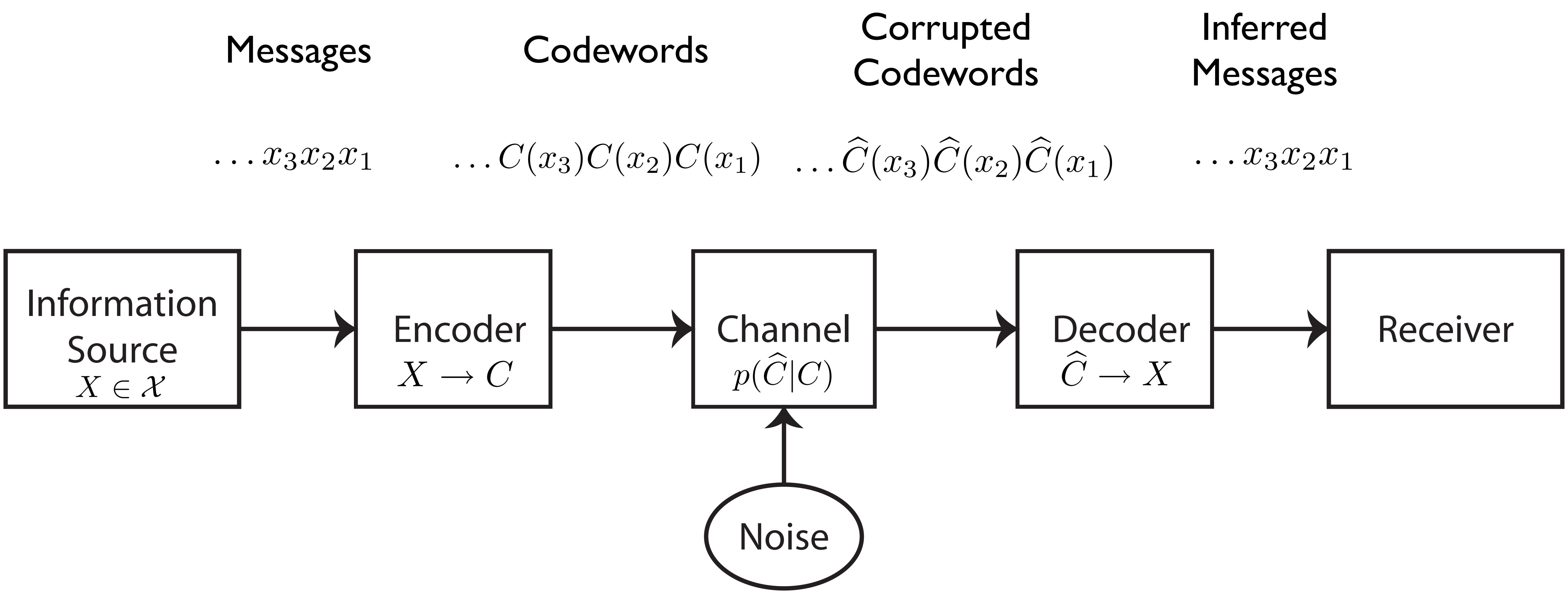
$$\mathcal{R} = \log_2 4 - 1.75 = 0.25 \text{ bits}$$

Messages are redundant,
but encoding is not.
Code saturates Shannon First Theorem bound:

$$R(C) = H(X)$$

Information in Processes ...

Communication channel:



Reliable transmission through noisy channel: Possible?

How to code in presence of distorted codewords?

Information in Processes ...

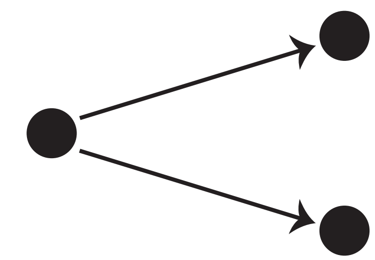
Coding for Communication Channels:

Kinds of channel:

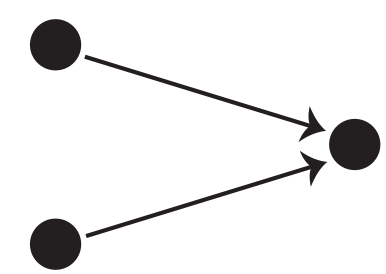
- Phone line, ftp/http transfer, monologue, ...
- Dynamical system at time t and t+1
- Spin system at one site and another
- Measuring instrument
- Learning channel

Channel coding problem is to overcome errors:

Equivocation:
Same input sequence leads to different outputs



Ambiguity:
Two different inputs lead to same output



Strategy:
Find channel inputs that are *least ambiguous* given distortion properties.
Codebook: Map information source onto those inputs.

Information in Processes ...

Coding for Communication Channels ...

Discrete channel:

Input: $X \sim p(x)$

Output: $Y \sim p(y)$

Channel: $p(y|x)$

Memoryless channel:

$$p(y_t|x_t x_{t-1} \cdots) = p(y_t|x_t)$$

Channel Capacity:

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

Highest rate one can transmit over channel.

Information in Processes ...

Coding for Communication Channels ...

Channel Capacity ...

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

Extremes of no communication:

No info to send: $H(X) = 0$

$$I(X;Y) = H(X) - H(X|Y) = 0 - 0 = 0$$

Complete distortion:

Output independent of input: $X \perp Y$

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

Information in Processes ...

Coding for Communication Channels ...

Duality:

Compression removes redundancy to give smallest description.

Encoding adds redundancy to compensate channel errors.

Information in Processes ...

Coding for Communication Channels ...

Example: Noiseless Binary Channel

$x \in \{0, 1\} \quad y = \{a, b\}$

$$p(y|x) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

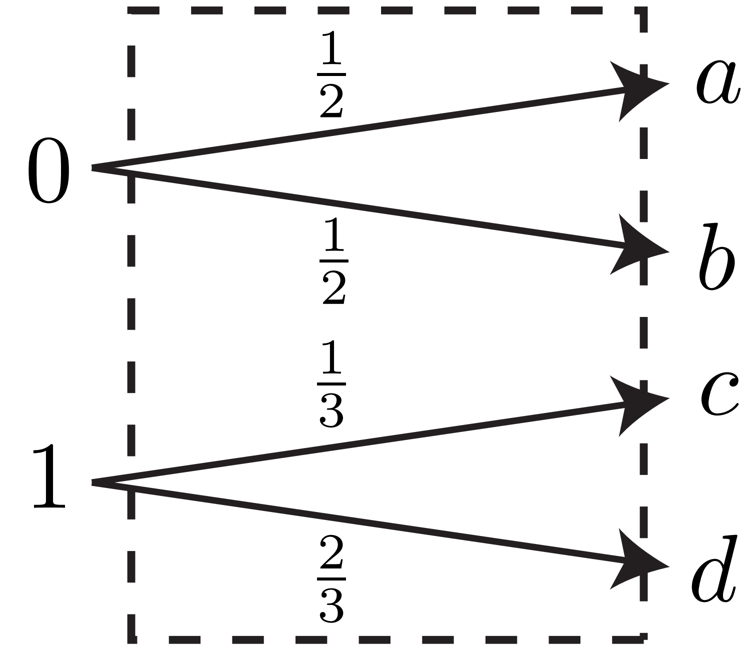
Capacity:
$$\begin{aligned} \mathcal{C} &= \max_{p(x)} I(X;Y) \\ &= \max_{p(x)} (H(X) - H(X|Y)) \\ &= \max_{p(x)} (H(X) - 0) \\ &= 1 \text{ bit} \end{aligned}$$

Achieved when source is: $p(x) = (\frac{1}{2}, \frac{1}{2})$

Example: Noisy Channel with Nonoverlapping Output

$x \in \{0, 1\}$ $\quad y \in \{a, b, c, d\}$

$$p(y|x) = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{3} & \frac{2}{3} \end{pmatrix}$$



Capacity:
$$\begin{aligned} \mathcal{C} &= \max_{p(x)} I(X;Y) \\ &= \max_{p(x)} (H(X) - H(X|Y)) \\ &= \max_{p(x)} (H(X) - 0) \\ &= 1 \text{ bit} \end{aligned}$$

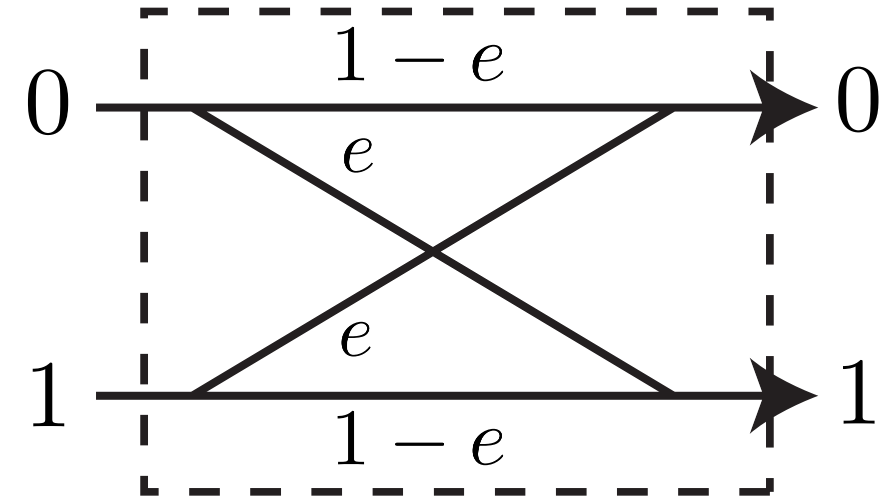
Achieved when source is: $p(x) = (\frac{1}{2}, \frac{1}{2})$

Information in Processes ...

Coding for Communication Channels ...

Example: Binary Symmetric Channel with error probability e

$$p(y|x) = \begin{pmatrix} 1-e & e \\ e & 1-e \end{pmatrix}$$



Capacity:

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

$$\mathcal{C} = \max_{p(x)} I(X;Y) = 1 - H(e)$$

Achieved when source is:

$$p(x) = (\tfrac{1}{2}, \tfrac{1}{2})$$

Complete distortion:

$$e = \tfrac{1}{2} \Rightarrow \mathcal{C} = 0$$

Information in Processes ...

Properties of Channel Capacity:

(1) Positive: $\mathcal{C} \geq 0$
(2) Bounded: $\mathcal{C} \leq \log_2 |\mathcal{X}|$ & $\mathcal{C} \leq \log_2 |\mathcal{Y}|$
(3) Continuity: $I(X;Y)$ continuous in $p(x)$
(4) Concavity: $I(X;Y)$ concave function of $p(x)$

Global maximum exists: $\mathcal{C} = \max_{p(x)} \; I(X;Y)$

No closed-form solution, in general,
but use nonlinear optimization methods to find.

Channel Coding Theorem (Shannon's Second Theorem):

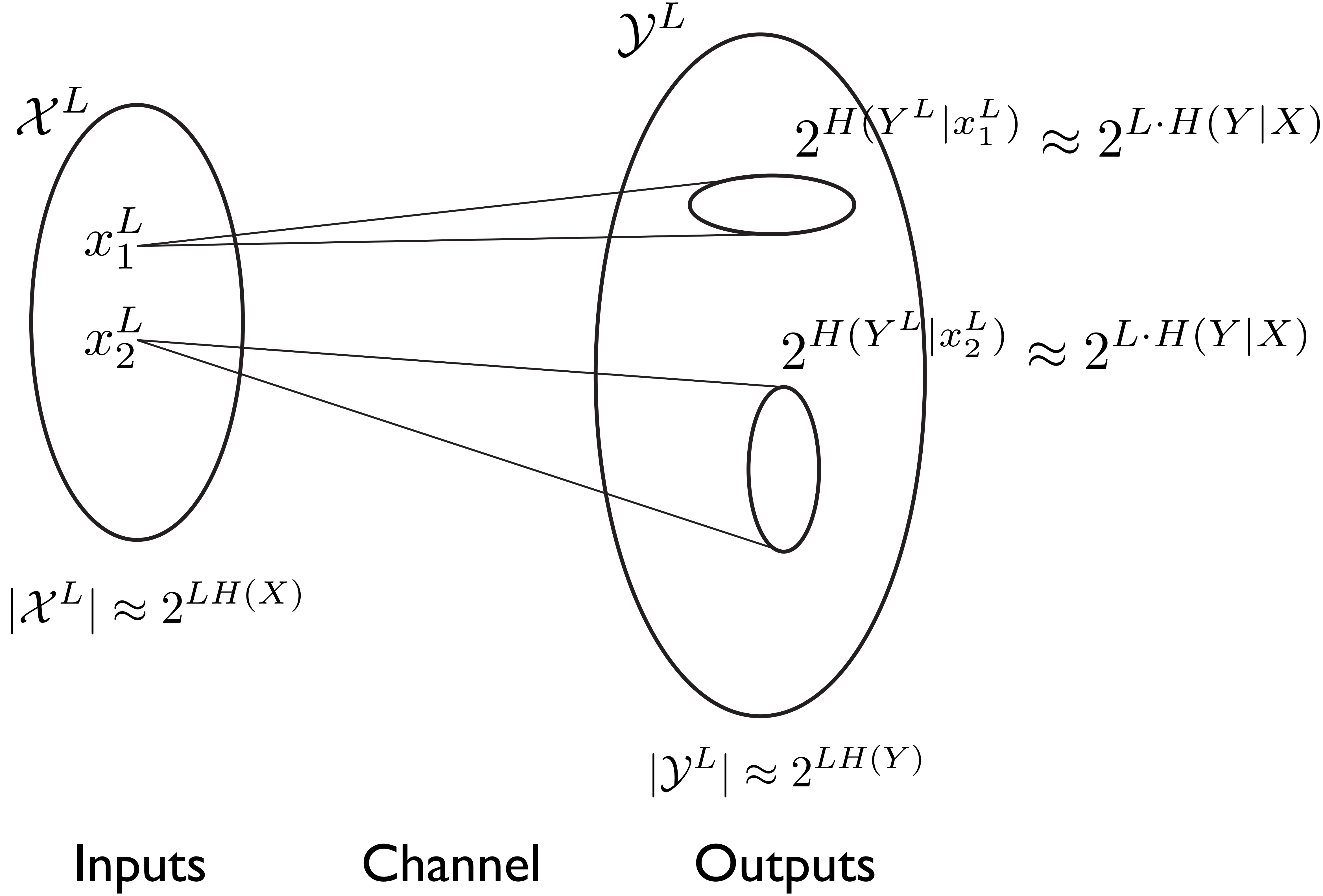
(1) Capacity is the maximum reliable transmission rate.
(2) Error-free codes exist if $R < \mathcal{C}$.

Idea:
Model as noisy channel with non-overlapping outputs.

Strategy:
Code long block lengths: $|\mathcal{X}^L| \approx 2^{LH(X)}$
Choose codewords (channel inputs) that
produce non-overlapping outputs.

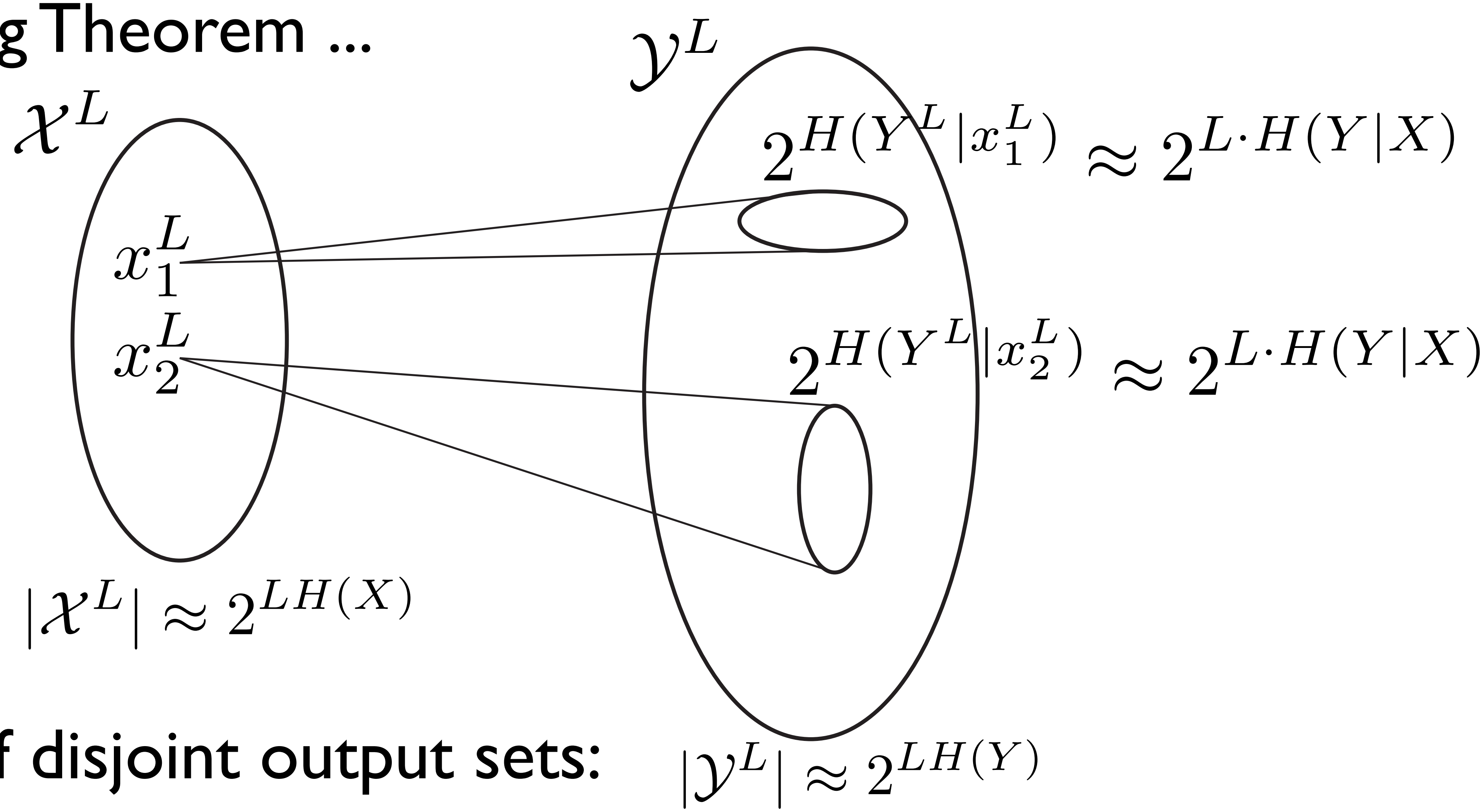
Information in Processes ...

Channel Coding Theorem ... plausibility ...



Information in Processes ...

Channel Coding Theorem ...



Total number of disjoint output sets:

$$\frac{2^{LH(Y)}}{2^{LH(Y|X)}} = 2^{L(H(Y)-H(Y|X))}$$

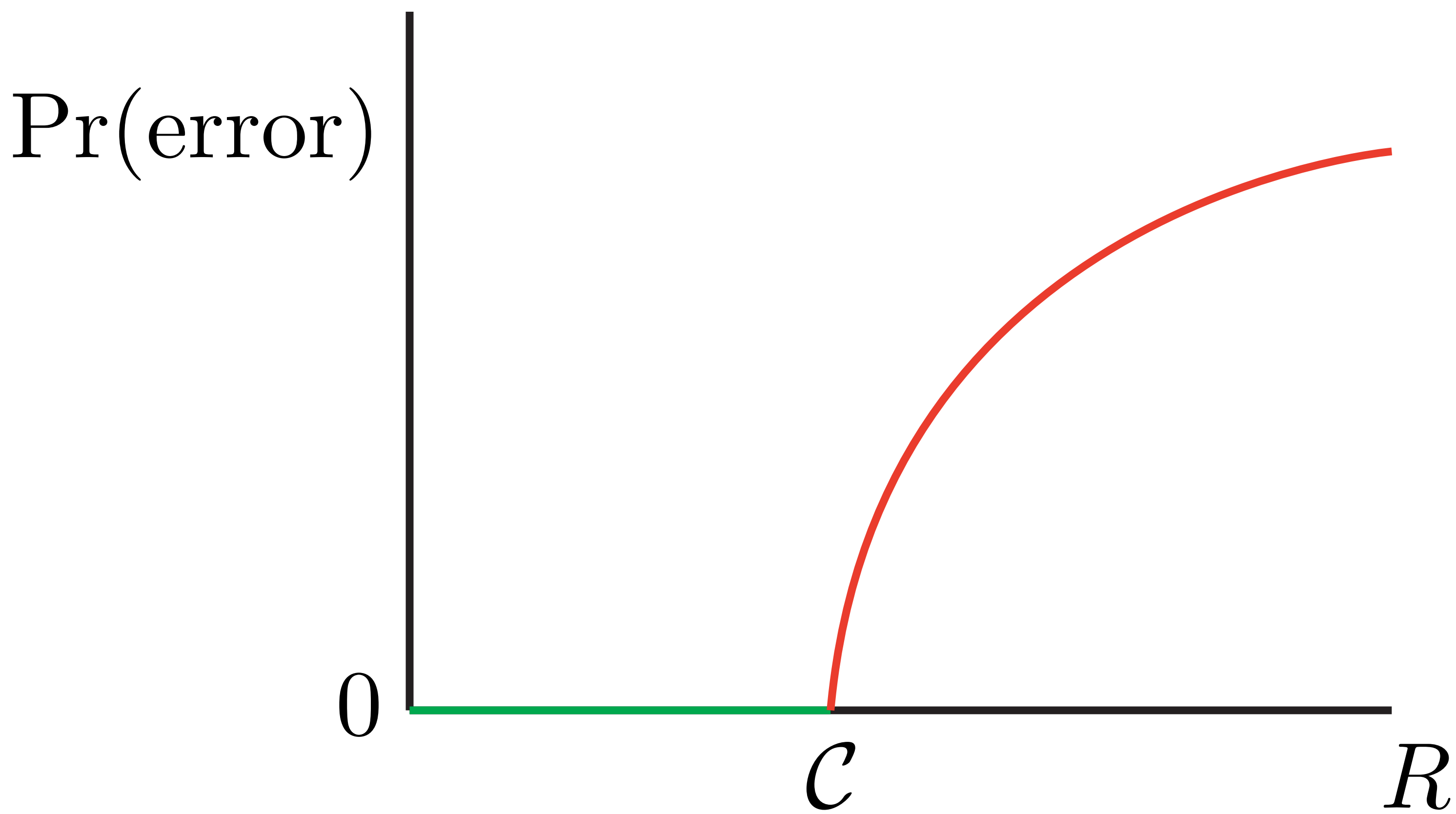
$$= 2^{LI(X;Y)}$$

Can send at most $2^{LI(X;Y)}$ distinguishable sequences of length L.

$\mathcal{C} = \max_{p(x)} \; I(X;Y)$ is the maximum transmission rate.

Information in Processes ...
  Channel Coding Theorem ...

What happens when transmitting above capacity, $R > \mathcal{C}$?



(Typical of measurement systems?)

Information in Processes ...

Reading for next lecture:

EIT, Chapter 4 and CMR article RURO.