# I232 Information Theory
# Chapter 8: Channel Coding Theorem

Brian Kurkoski

Japan Advanced Institute of Science and Technology

2023 May

# Main Idea for This Lecture

- Section 7.1 introduced error-correcting codes with rate $R$. For example, the repeat code in Section 7.2.
- Section 7.3 introduced channel capacity $C$

In this lecture, we connect these two ideas:

**Channel Coding Theorem**

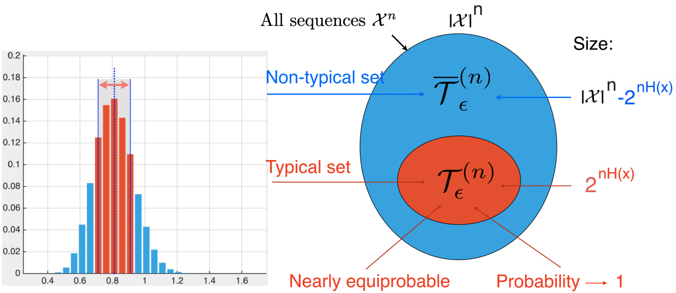
We can have reliable communications if and only if $R < C$.

# Recall Single Variable Typical Sets from Lecture 5

## Definition

The *typical set* $\mathcal{T}_\epsilon^{(n)}$ is the set of sequences $\mathbf{x} \in \mathcal{X}^n$ with sample entropy $\epsilon$-close to the true entropies:

$$\mathcal{T}_\epsilon^{(n)} = \left\{ \mathbf{x} \in \mathcal{X}^n : \left| -\frac{1}{n} \log p_{\mathbf{X}}(\mathbf{x}) - H(\mathsf{X}) \right| < \epsilon \right\}$$



# Outline

8.1 Joint Typicality and Joint AEP

8.2 Shannon's Channel Coding Theorem

8.3 Proof of Channel Coding Theorem – Direct Part

8.4 Proof of Channel Coding Theorem – Converse Part

## 8.1 Joint Typicality and Joint AEP

Joint typicality extends the ideas of typicality from a sequence of one variable **X** to two variables **X**, **Y**.

Let **X**, **Y** be jointly distributed vectors with $p_{\mathbf{XY}}(\mathbf{x}, \mathbf{y})$:

$$\begin{aligned}\mathbf{x} &= (x_1, x_2, \ldots, x_n)\\ \mathbf{y} &= (y_1, y_2, \ldots, y_n)\\ (\mathbf{x}, \mathbf{y}) &= \Big((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\Big)\end{aligned}$$

The $(\mathsf{X}_i, \mathsf{Y}_i)$ are pairwise i.i.d.:

$$p_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{n} p_{\mathsf{XY}}(x_i, y_i) = \prod_{i=1}^{n} p_{\mathsf{Y}|\mathsf{X}}(y|x) p_{\mathsf{X}}(x).$$

as the input and output of a DMC.

# Joint Typicality

The set $\mathcal{T}_\epsilon^{(n)}$ of *jointly typical* sequences $(\mathbf{x}, \mathbf{y})$ satisfy:

$$\begin{aligned}\mathcal{T} = \Big\{(\mathbf{x}, \mathbf{y}) \in \mathcal{X}^n \times \mathcal{Y}^n :\\ \Big| -\frac{1}{n}\log p_{\mathsf{X}}(\mathbf{x}) - H(\mathsf{X})\Big| < \epsilon,\\ \Big| -\frac{1}{n}\log p_{\mathsf{Y}}(\mathbf{y}) - H(\mathsf{Y})\Big| < \epsilon,\\ \Big| -\frac{1}{n}\log p_{\mathsf{XY}}(\mathbf{x}, \mathbf{y}) - H(\mathsf{X}, \mathsf{Y})\Big| < \epsilon\Big\}\end{aligned}$$

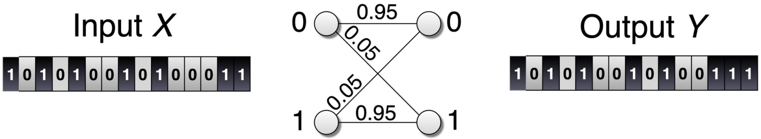
# Definition of Jointly Typical Sequences

The set $\mathcal{T}_\epsilon^{(n)}$ of *jointly typical* sequences $(\mathbf{x}, \mathbf{y})$ satisfy:

$$
\begin{aligned}
& \mathcal{T} = \Big\{ (\mathbf{x}, \mathbf{y}) \in \mathcal{X}^n \times \mathcal{Y}^n : \\
\boxed{\mathbf{x}\text{ sample entropy is }\epsilon\text{-close to }H(\mathsf{X})} \longrightarrow \quad & \Big| -\frac{1}{n} \log p_{\mathsf{X}}(\mathbf{x}) - H(\mathsf{X}) \Big| < \epsilon, \\
\boxed{\mathbf{y}\text{ sample entropy is }\epsilon\text{-close to }H(\mathsf{Y})} \longrightarrow \quad & \Big| -\frac{1}{n} \log p_{\mathsf{Y}}(\mathbf{y}) - H(\mathsf{Y}) \Big| < \epsilon, \\
\boxed{(\mathbf{x}, \mathbf{y})\text{ sample entropy is }\epsilon\text{-close to }H(\mathsf{X}, \mathsf{Y})} \longrightarrow \quad & \Big| -\frac{1}{n} \log p_{\mathsf{XY}}(\mathbf{x}, \mathbf{y}) - H(\mathsf{X}, \mathsf{Y}) \Big| < \epsilon \Big\}
\end{aligned}
$$

# Intuition of Joint Typicality

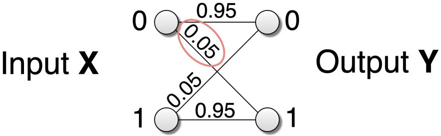
Let $\mathbf{x}$ be the input to the DMC and let $\mathbf{y}$ be the output.

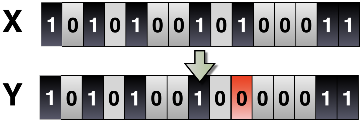


Intuition: Given arbitrary $(\mathbf{x}, \mathbf{y})$, can we say this output $\mathbf{y}$ is a good "explanation" of the input $\mathbf{x}$?

# Joint Typicality – Is $(\mathbf{x}, \mathbf{y})$ Generated by the System?

Is $(\mathbf{x}, \mathbf{y})$ generated by the system? $\Leftrightarrow$ Are $\mathbf{x}$ and $\mathbf{y}$ jointly typical?

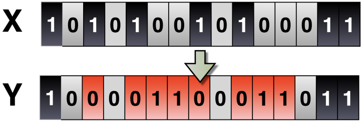




X: 1 0 1 0 1 0 0 1 0 1 0 0 0 1 1

Y: 1 0 1 0 1 0 0 1 0 0 0 0 0 1 1

Yes, with high probability



X: 1 0 1 0 1 0 0 1 0 1 0 0 0 1 1

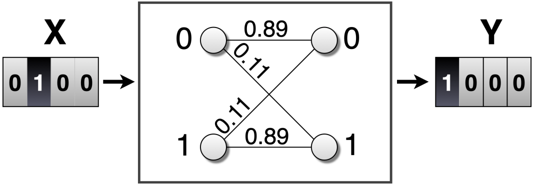
Y: 1 0 0 0 0 1 1 0 0 0 1 1 0 1 1

No, with high probability

# Numerical Example: Jointly Typical Sequences

Consider the following:

- BSC with channel error $p = 0.11$
- input distribution $p(x) = [0.75, 0.25]$
- With $n = 4$, input $\mathbf{x} = 0100$ and output $\mathbf{y} = 1000$.



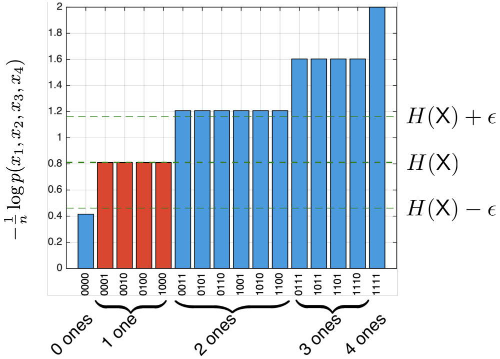
**Question 1:** For $\epsilon = 0.35$, is $(\mathbf{x}, \mathbf{y}) = (0100, 1000)$ in the typical set $\mathcal{T}_{0.35}^{(4)}$?

**Question 2:** What are all the typical sequences in the set $\mathcal{T}_{0.35}^{(4)}$? ★1

# Typical Sequences in X Only

X $\sim p(x) = [0.75, 0.25]$, $H(\mathsf{X}) = 0.8113$, $n = 4$ and $\epsilon = 0.35$

| $x_1, x_2, x_3, x_4$ | $-\frac{1}{n}\log p(x_1, x_2, x_3, x_4)$ |
|---|---|
| 0 0 0 0 | 0.4150 |
| 1 0 0 0 | 0.8113 |
| 0 1 0 0 | 0.8113 |
| 0 0 1 0 | 0.8113 |
| 0 0 0 1 | 0.8113 |
| 1 1 0 0 | 1.2075 |
| 1 0 1 0 | 1.2075 |
| 0 1 1 0 | 1.2075 |
| 1 0 0 1 | 1.2075 |
| 0 1 0 1 | 1.2075 |
| 0 0 1 1 | 1.2075 |
| 1 1 1 0 | 1.6038 |
| 1 1 0 1 | 1.6038 |
| 1 0 1 1 | 1.6038 |
| 0 1 1 1 | 1.6038 |
| 1 1 1 1 | 2.0000 |

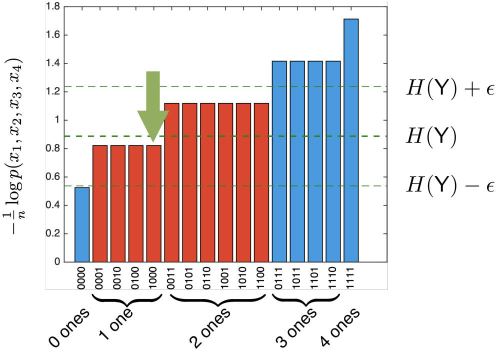


$\Rightarrow$ Typical $\mathbf{x}$ are $\{0001, 0010, 0100, 1000\}$

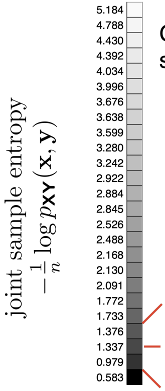
# Typical Sequences in Y Only

$\mathsf{Y} \sim p_{\mathsf{Y}}(y) = [0.695, 0.305]$, $H(\mathsf{X}) = 0.887$, $n = 4$ and $\epsilon = 0.35$

- Typical $\mathbf{y}$ are $\{0001, 0010, 0100, 1000, 1100, 1010, 1001, 0110, 0101, 0011\}$
- Is $\mathbf{y} = 1000$ typical? $\Rightarrow$ Yes.



# Joint Sample Entropy in X and Y



Gray represents joint sample entropy value

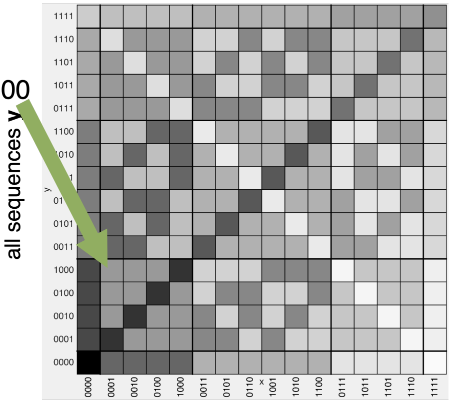
$H(\mathsf{X},\mathsf{Y})+\epsilon \quad = 1.66$

$H(\mathsf{X},\mathsf{Y}) \quad = 1.31$

$H(\mathsf{X},\mathsf{Y})-\epsilon \quad = 0.96$

**[0.9612, HXY=1.3112, 1.6612]**

0100,1000



all sequences **x**

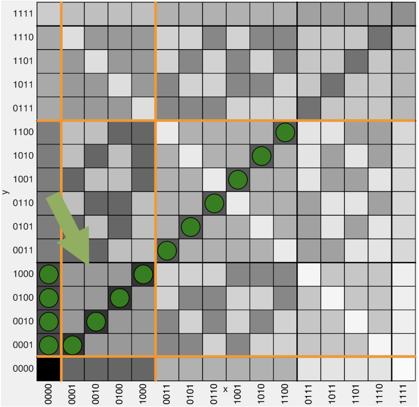
★2

# Joint Sample Entropy in X and Y

- Green dots show $(\mathbf{x}, \mathbf{y})$ typical sequences:

$$
\begin{aligned}
\{&(0000, 0001),\\
&(0000, 0010),\\
&(0000, 0100),\\
&(0000, 1000),\\
&(0001, 0001),\\
&\vdots,\\
&(1100, 1100)\}
\end{aligned}
$$

- Is $(\mathbf{x}, \mathbf{y}) = (0100, 1000)$ typical? $\Rightarrow$ No, is not one of the sequences.



# Joint Sample Entropy in X and Y

**Question 1:** For $\epsilon = 0.35$, is $(\mathbf{x}, \mathbf{y}) = (0100, 1000)$ in the typical set $\mathcal{T}_{0.35}^{(4)}$?

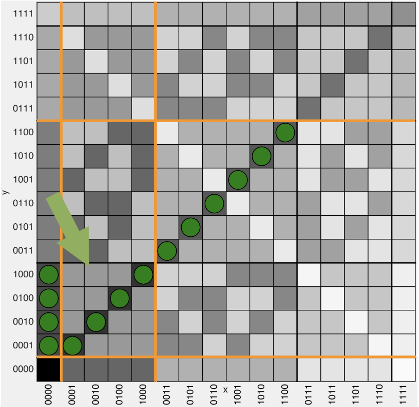
$(0100, 1000)$ must satsify 3 conditions to be jointly typical:

✓ $\mathbf{x}$ is $\epsilon$-close to $H(\mathsf{X})$

✓ $\mathbf{y}$ is $\epsilon$-close to $H(\mathsf{Y})$

× $(\mathbf{x}, \mathbf{y})$ is not $\epsilon$-close to $H(\mathsf{X}, \mathsf{Y})$

$\Rightarrow$ $(\mathbf{x}, \mathbf{y}) = (0100, 1000)$ is not jointly typical, and it is not in the typical set $\mathcal{T}_{0.35}^{(4)}$.
If $\mathbf{x} = 0100$ is the channel input, then $\mathbf{y} = 1000$ is not a good "explanation" of the output.



# Joint Sample Entropy in $X$ and $Y$

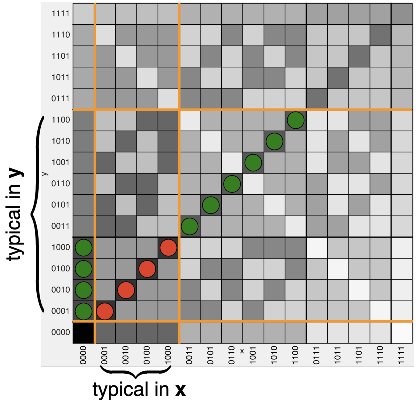
**Question 2:** What are all the typical sequences in the set $\mathcal{T}_{0.35}^{(4)}$?

Jointly typical sequences are:

- typical in $\mathbf{x}$
- typical in $\mathbf{y}$
- typical in $\mathbf{x}, \mathbf{y}$

which are:

$$\begin{aligned}\mathcal{T}_{0.35}^{(4)} = \{&(0001, 0001),\\ &(0010, 0010),\\ &(0100, 0100),\\ &(1000, 1000)\}\end{aligned}$$



shown with red dots. Why are these a good "explanation" of the system?

# Joint Asymptotic Equipartition Property

As $n \to \infty$, the probability that a randomly drawn sequence is jointly typical tends to 1.

## Proposition (Proposition 8.1)

*Joint Asymptotic Equipartition Property*. Let $(\mathbf{x}, \mathbf{y})$ be sequences of length $n$ drawn i.i.d. from $p_{\mathbf{XY}}(\mathbf{x}, \mathbf{y})$. Then:

$$\lim_{n \to \infty} \Pr\left[(\mathbf{X}, \mathbf{Y}) \in \mathcal{T}_\epsilon^{(n)}\right] = 1.$$

- This is the two-variable analog of the one-variable AEP Proposition 5.2
- Proved using the weak law of large numbers.

# Most Sequences are Typical

### Proposition (Proposition 8.2)

*Most sequences are typical*

$$\Pr\left[(\mathbf{X},\mathbf{Y}) \in \mathcal{T}_\epsilon^{(n)}\right] \geq 1-\epsilon,$$

for sufficiently large $n$. For the complementary event

$$\Pr\left[(\mathbf{X},\mathbf{Y}) \notin \mathcal{T}_\epsilon^{(n)}\right] < \epsilon.$$

- That is, a sequence drawn randomly from $p_{\mathsf{XY}}(x,y)$ is typical with high probability.
- Proved using definition of joint typicality: $\log p_{\mathsf{XY}}(x,y) \leq -n(H(\mathsf{X},\mathsf{Y})-\epsilon)$.

# Independent Sequences Are Not Jointly Typical

Consider that $\mathsf{X}, \mathsf{Y}$ are jointly distributed. Let $\widetilde{\mathsf{X}}$ and $\widetilde{\mathsf{Y}}$ be *independent* random variables with the same distribution as $\mathsf{X}$ and $\mathsf{Y}$.

## Proposition (Proposition 8.4)

*Independent sequences are not jointly typical* If $\widetilde{\mathsf{X}}$ and $\widetilde{\mathsf{Y}}$ are independent random variables with the same distribution as $\mathsf{X}$ and $\mathsf{Y}$, then:

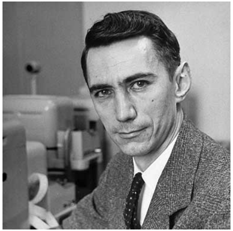
$$\Pr\left((\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}}) \in \mathcal{T}_\epsilon^{(n)}\right) \leq 2^{-n(I(\mathsf{X};\mathsf{Y})-3\epsilon)}$$

- Independent $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$ will be jointly typical with low probability.
- If $\mathbf{x}$ does not "explain" $\mathbf{y}$, then the probability they are jointly typical is small.
- Proved using $p_{\mathsf{XY}}(x, y) = p_{\mathsf{X}}(x)p_{\mathsf{Y}}(y)$ for the independent variables.
- Since $I(\mathsf{X}, \mathsf{Y}) \geq 0$, we have $2^{-nI(\mathsf{X};\mathsf{Y})}$ small for large $n$.

★poll

# 8.2 Shannon's Channel Coding Theorem

- Shannon's channel coding theorem is perhaps the most celebrated result in information theory.
- States that it is possible to have reliable communications over unreliable channels.
- Gives the maximum rate of communications over a given channel.



- C. E. Shannon, "A Mathematical Theory of Communications," *Bell System Technical Journal*, July–October 1948

# Shannon's Channel Coding Theorem

*Informal* version of the channel coding theorem:

We can have reliable communications if and only if $R < C$.

*Formal* version of the channel coding theorem:

## Proposition (Proposition 8.5, Channel Coding Theorem)

- **Direct Part or Achievability:** For every rate $R \leq C$, there exists a $(2^{nR}, n)$ code with
$$\lim_{n\to\infty} \lambda^{(n)} = 0.$$
- **Converse Part:** Any $(2^{nR}, n)$ codes with $\lim_{n\to\infty} \lambda^{(n)} = 0$ must have
$$R \leq C.$$

Recall that $R$ is the code rate, $C$ is the channel capacity and $\lambda^{(n)}$ is the maximum probability of error.

# Example: Achievable Rates for BSC $\alpha = 0.1$

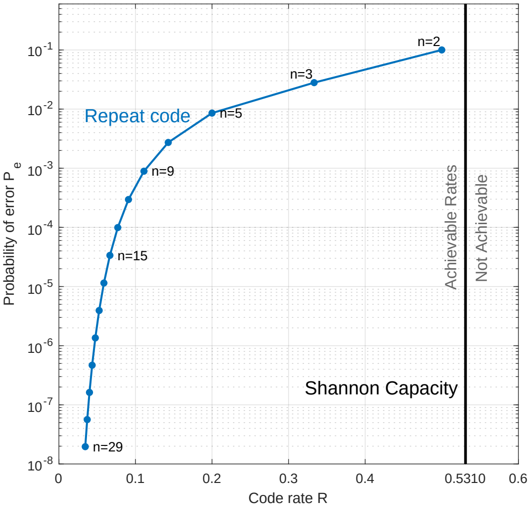
Consider a BSC with $\alpha = 0.1$. The capacity is $C = 1 - h(\alpha) \approx 0.531$.

By the Channel Coding Theorem, only codes with rates $R < 0.531$ can achieve low probability of error.

For the repeat code, $P_e$ can be computed:

- The repeat code has $P_e \to 0$ as $n \to \infty$.

# Example: Achievable Rates for BSC $\alpha = 0.1$

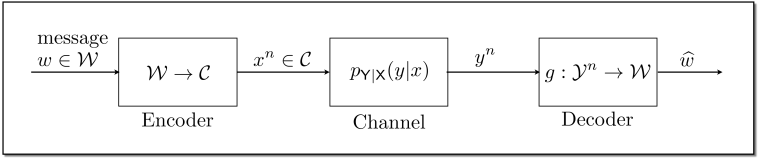


The repeat code is not good asymptotically:

- For the repeat code, $P_e \to 0$, as $n \to \infty$.
- But also $R \to 0$. The gap between the repeat code and capacity increases
- The repeat code is not a very good code

However, good codes do exist, such as polar codes and LDPC codes.

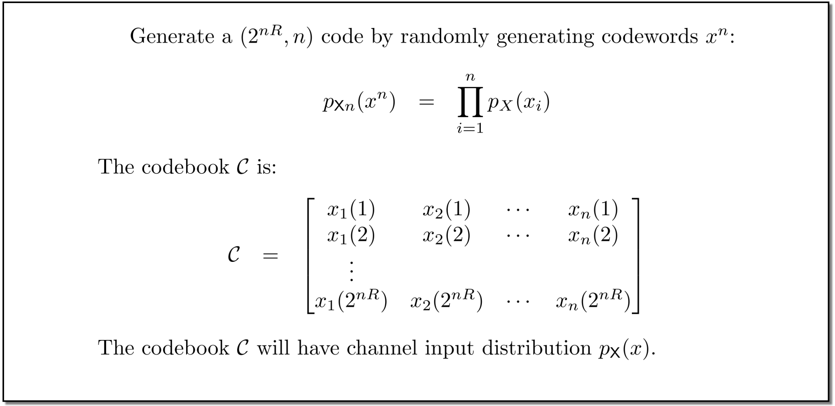
# Random Coding Scheme


message
$w \in \mathcal{W}$
$\mathcal{W} \to \mathcal{C}$
Encoder
$x^n \in \mathcal{C}$
$p_{\mathsf{Y}|\mathsf{X}}(y|x)$
Channel
$y^n$
$g : \mathcal{Y}^n \to \mathcal{W}$
Decoder
$\widehat{w}$

The proof will use the following communication system:

- **Encoder:** Randomly generates a codebook
- **Channel:** Discrete memoryless channel
- **Decoder:** Uses jointly-typical decoding

# Random Codebook Construction



Generate a $(2^{nR}, n)$ code by randomly generating codewords $x^n$:

$$p_{\mathsf{X}_n}(x^n) \quad = \quad \prod_{i=1}^{n} p_X(x_i)$$

The codebook $\mathcal{C}$ is:

$$\mathcal{C} \quad = \quad \begin{bmatrix} x_1(1) & x_2(1) & \cdots & x_n(1) \\ x_1(2) & x_2(2) & \cdots & x_n(2) \\ \vdots & & & \\ x_1(2^{nR}) & x_2(2^{nR}) & \cdots & x_n(2^{nR}) \end{bmatrix}$$

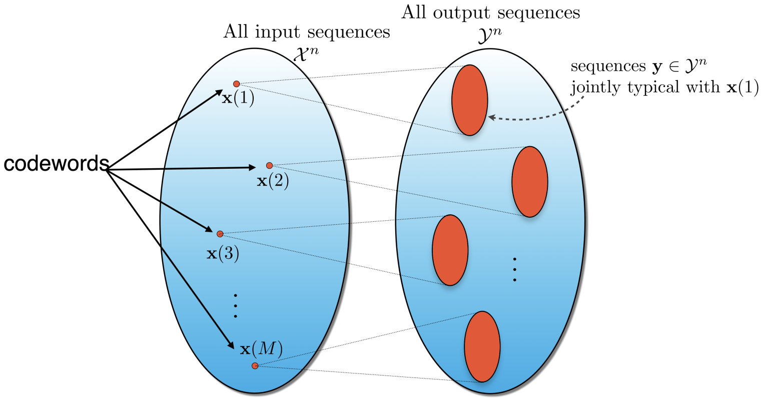
The codebook $\mathcal{C}$ will have channel input distribution $p_{\mathsf{X}}(x)$.

# Notes about Random Codebook Construction

1. Codebook and channel input use the same alphabet $\mathcal{X}$
2. Codebook distribution $p_{\mathsf{X}}(x)$ is equal to the channel input distribution.
    - Later, we choose $p_{\mathsf{X}}(x)$ to be the capacity-achieving input distribution $p^*_{\mathsf{X}}(x)$
3. Codebook is random. But, but the encoder and decoder agree on the codebook.

# Jointly Typical Sequences

Visualize the set of jointly typical sequences



# Jointly Typical Decoding

Jointly typical decoding is performed.

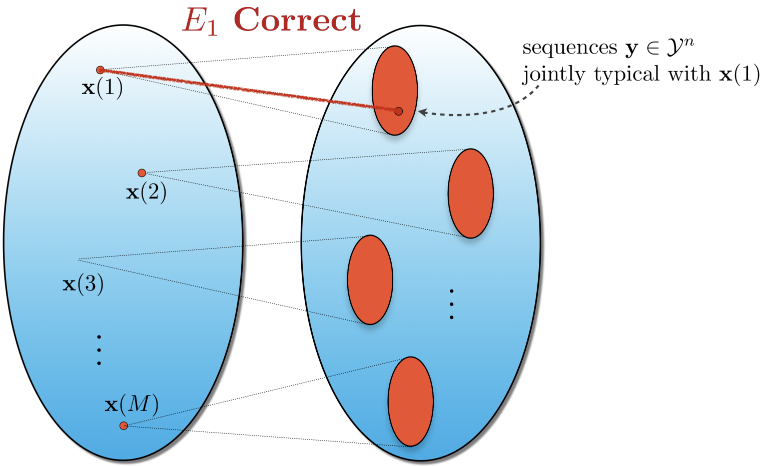
Given decoder input $\mathbf{y}$, the decoder estimates $\hat{w}$ if:

- $(\mathbf{x}(\hat{w}), \mathbf{y})$ are jointly typical, and
- There is no other index $w' \neq \hat{w}$ such that $(\mathbf{x}(w'), \mathbf{y})$ are jointly typical.

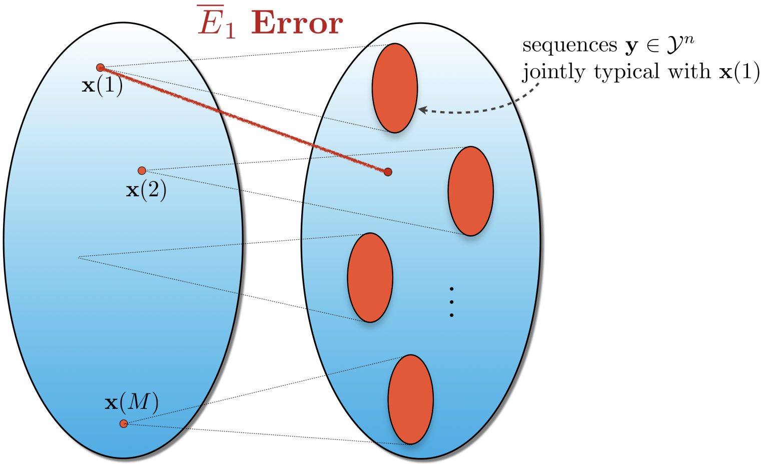
Define the event $\mathsf{E}_w$ as:

$$\mathsf{E}_w = \text{Event that } (\mathbf{x}(w), \mathbf{y}) \text{ are jointly typical.}$$

# Jointly Typical Decoding



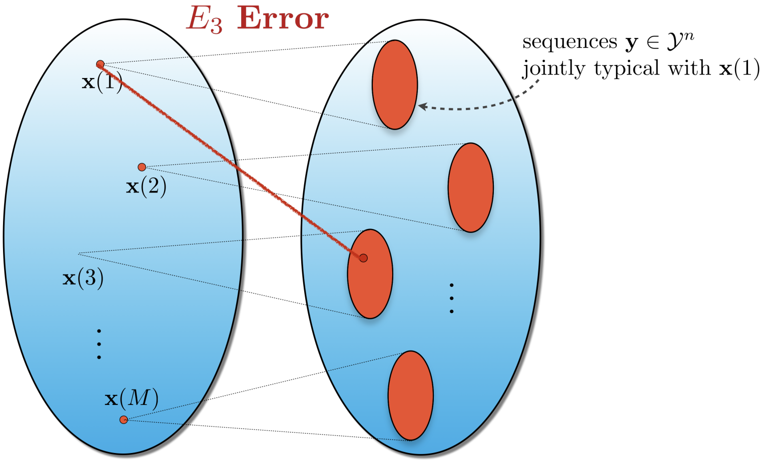
# Jointly Typical Decoding



$\overline{E}_1$ **Error**

sequences $\mathbf{y} \in \mathcal{Y}^n$
jointly typical with $\mathbf{x}(1)$

# Jointly Typical Decoding



Error: $\mathbf{x}(1)$ was transmitted but $(\mathbf{x}(3), \mathbf{y})$ are jointly typical.

# What is “Reliable Communication"?

- Recall $\lambda^{(n)}$ is the maximum probability of error.
- **Reliable communications:** we can make $\lambda^{(n)}$ as small as we want. Does not mean it is perfectly reliable, for any finite $n$.
- **Achievability** means probability of error goes to 0 as $n$ gets large

# Comments on the Coding Theorem

1. Long, random codes achieve channel capacity
2. The channel is used $n$ times, and codewords are long sequences of lenght $n$. The law of large numbers is used, and we average over all channel uses.
3. The theorem shows the existence of at least one good code.
4. The theorem does not tell us how to construct practical codes. Random codes and jointly typical decoding used by the proof are not efficient.
5. **Question:** “Given a DMC, is there a code with error probability $\lambda$?"

   **Answer:** yes, if $R < C$, and you allow $n$ to be big enough.

# 8.3 Proof of Channel Coding Theorem – Direct Part

### Proposition

For every rate $R < C$, there exists a $(2^{nR}, n)$ code with $\lambda^{(n)} \to 0$.

**Proof Sketch:**

1. Random codebook construction, jointly typical decoding
2. Error occurs if transmitted $\mathbf{x}$ and received $\mathbf{y}$ are not jointly typical
3. Upper bound the average probability of error
4. Upper bound the maximum probability of error

# Decoding Error

- Error probability for a specific code $\mathcal{C}$:

$$P_e(\mathcal{C}) = \tfrac{1}{M} \sum_{w=1}^{M} \lambda_w(\mathcal{C}).$$

- Error probability averaged over all codes:

$$\Pr(\mathrm{E}) = \sum_{\mathcal{C}} P_e(\mathcal{C}) \Pr(\mathcal{C}).$$

For a fixed code, $P_e(\mathcal{C})$ is very difficult to compute. Instead, we compute $\Pr(\mathrm{E})$, which shows the existence of at least one good code. ★3

# Existence of At Least One Code

Average error probability over all codes is upper bounded by $2\epsilon$.
$\Rightarrow$ At least one code whose maximum error probability is upper bounded by $4\epsilon$.

We showed that:

$$\text{Average probability of error } P_e^{(n)} \leq 2\epsilon.$$

**Expurgate** the worst codewords from the codebook. Then

$$\text{Maximum probability of error } \lambda^{(n)} \leq 4\epsilon,$$

with code rate $R' = R - \frac{1}{n} \rightarrow R$. This proves achievability (direct part) of rates below capacity.

# 8.4 Proof of Channel Coding Theorem – Converse Part

- **Direct part:** Existence of a code $R < C$ for reliable communications.
- **Converse part:** Non-existence of a code $R > C$ for error-free communications.
- Reliable communications is possible **if and only if** $R < C$

# Preliminary Lemmas

Fano's Inequality for channel coding

For a DMC, a codebook $\mathcal{C}$ of length $n$ and rate $R$, and input message $W$:

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

Channel Reuse Does Not Increase Capacity

Let $\mathbf{Y}$ be the result of passing $\mathbf{X}$ through a DMC of capacity $C$. Then:

$$I(\mathbf{X}; \mathbf{Y}) \leq \mathbf{nC} \quad \text{for all } \mathbf{p}(\mathbf{x}).$$

# Converse Part of Shannon's Theorem

Any $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$ must have $R \leq C$.

# Summary

- Shannon's channel coding theorem is a **key result** in information theory
- Channel capacity $C$ is the maximization of $I(X;Y)$
- The rate of a code is $R$
- Shannon's Channel Coding Theorem: **Reliable communications is possible if and only if** $R < C$
- The proof uses **random code** construction, which is not practical
- Practical code constructions: **LDPC** codes and **polar** codes

# Class Info

- Next lecture: Wednesday, May 15. Differential Entropy and the Gaussian Channel. There will be a pop quiz.
- Midterm exam on May 15 at 13:30.
- Homework 7 on LMS (soon)

# Midterm Exam

The exam is closed book. You may use:

- One page of notes, A4-sized paper, double-sided OK.
- Blank scratch paper

You may not use anything else: No printed materials, including books, lecture notes, and slides. No notes (except as above). No internet-connected devices. No calculators ("log 3" is an acceptable answer). You may need to perform a $2 \times 2$ matrix inverse.

Exam Content

- Covers Chapters 1–6
- Study Homework 1–6. Solutions to Homework 1–6 are provided.
- No programming questions.

Practice problems will be provided.