# 4F5: Advanced Communications and Coding

## Handout 4: The Channel Coding Theorem

Ramji Venkataramanan

Signal Processing and Communications Lab
Department of Engineering
ramji.v@eng.cam.ac.uk

Michaelmas Term 2015



# Definition of a Channel Code


We use the channel $n$ times to transmit a message $W \in \{1, \dots, M\}$.

An *$(M, n)$ channel code* for the channel $(\mathcal{X}, \mathcal{Y}, P_{Y|X})$ consists of:

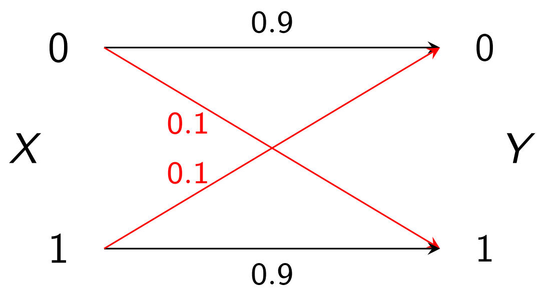
1. A set of messages $\{1, \dots, M\}$
2. An encoding function $X^n : \{1, \dots, M\} \to \mathcal{X}^n$ that assigns a *codeword* to each message. The set of codewords $\{X^n(1), \dots, X^n(M)\}$ is called the *codebook*
3. A decoding function $g : \mathcal{Y}^n \to \{1, \dots, M\}$, which produces a guess of the transmitted message for each received vector

*$M$ messages $\longleftrightarrow$ $\log M$ bits*

The *rate* $R$ of an $(M, n)$ code is

$$R = \frac{\log M}{n} \text{ bits/transmission}$$

# Preview of the Channel Coding Theorem




For intuition, let us start with the BSC(0.1):

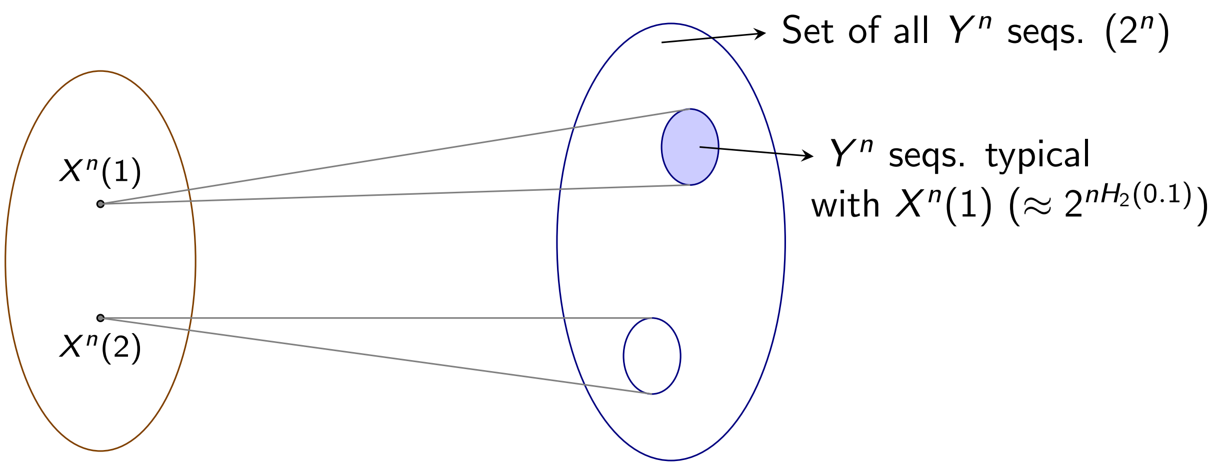
- For input sequence $X^n$, the output $Y^n$ is generated as

$$Y_i = X_i \oplus E_i \qquad \text{for } i = 1, \dots, n$$

- $E_1, \dots, E_n$ i.i.d $\sim$ Bernoulli(0.1) is the sequence of errors introduced by the channel ($\oplus$ denotes modulo-two addition)
- For large $n$, the number of ones in $(E_1, \dots, E_n) \approx 0.1n$ (AEP)

How big is the set of $Y^n$ sequences "typical" with any given $X^n$?
Ans: $\approx 2^{nH_2(0.1)}$






- The high-probability set of typical $Y^n$ sequences for a given $X^n(1)$ is much smaller than $2^n$
- Pick $X^n(2)$ "far enough" away from $X^n(1)$.
- Then the typical set of $Y^n$'s for $X^n(2)$ is *non-intersecting* with the typical set for $X^n(1)$.
- Number of distinct messages we can transmit = max. number of non-intersecting sets. (similar to noisy keyboard channel)

$$M \approx \frac{2^n}{2^{nH_2(0.1)}} \quad \Rightarrow \quad \text{Rate } R \approx 1 - H_2(0.1)$$

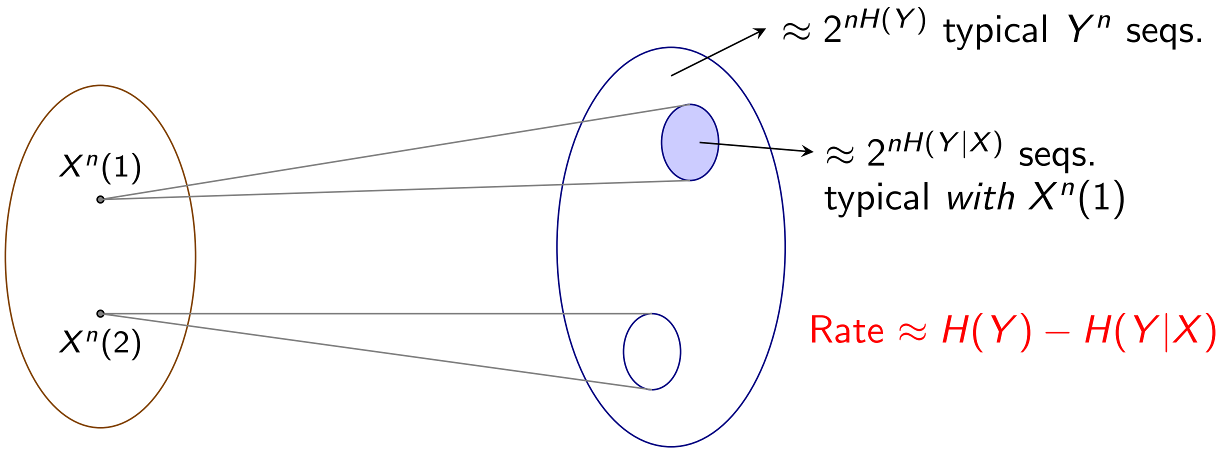
## The idea for a general DMC ...

Fix an input pmf $P_X$. Together with the channel $P_{Y|X}$, this gives

$$P_{XY} = P_X P_{Y|X}, \qquad P_Y = \sum_x P_X P_{Y|X}.$$

If $X^n(1), X(2), \ldots, X^n(M)$ are generated i.i.d $\sim P_X$, then:

- When $X^n(k)$ is transmitted, the set of highly likely $Y^n$'s has approximately $2^{nH(Y|X)}$ sequences for *each* $k \in \{1, \ldots, M\}$
- These sets are non-intersecting with high probability






## Joint typicality – A Motivating Example

Let $(X^n, Y^n)$ be drawn i.i.d. according to the following joint pmf:

| | | $Y$ | |
|---|---|---|---|
| | $P_{XY}$ | 0 | 1 |
| $X$ | 0 | 0.4 | 0.1 |
| | 1 | 0.1 | 0.4 |

$$Pr(X^n = x^n, Y^n = y^n) = \prod_{i=1}^{n} P_{XY}(x_i, y_i), \qquad \text{for all } (x^n, y^n).$$

Note that $P_X(0) = P_X(1) = \frac{1}{2}$, $P_Y(0) = P_Y(1) = \frac{1}{2}$.

For large $n$, what can we say about the sequences $(X^n, Y^n)$?

$$\text{E.g.} \quad X^n = 0\ 0\ 1\ 0\ 1\ 1\ 1\ 0\ 1\ 0\ 1\ 0\ 0\ 1\ 0$$
$$Y^n = 0\ 1\ 1\ 0\ 1\ 1\ 0\ 0\ 1\ 0\ 1\ 1\ 0\ 1\ 0$$

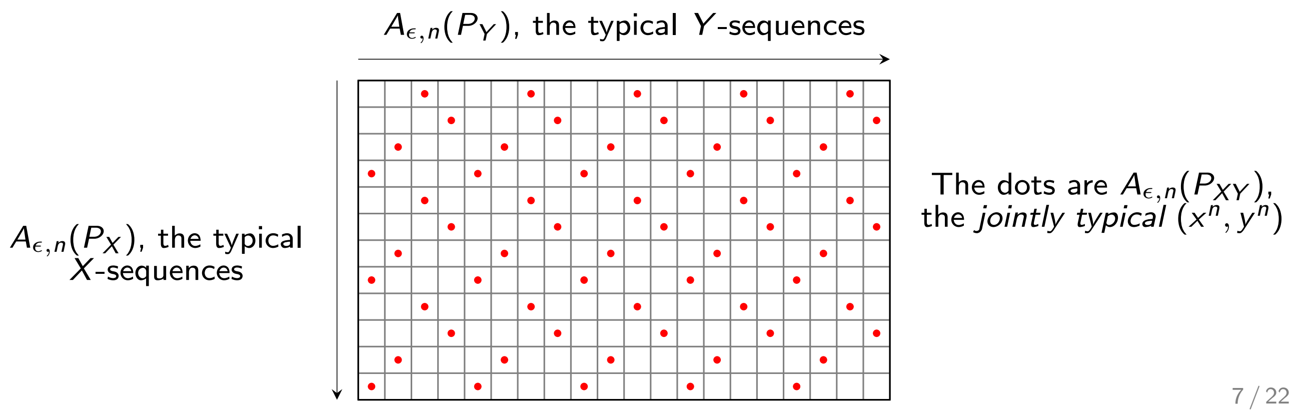
- $X^n$ and $Y^n$ will each have approximately 50% ones.
- The number of $(X_i, Y_i)$ pairs that are $(0,0), (0,1)$,(1,0)$, (1,1)$ will be close to $.4n$, $.1n$, $.1n$, $.4n$, respectively.

# Joint Typical Set

The set $A_{\epsilon,n}$ of *jointly typical* sequences $\{(x^n, y^n)\}$ with respect to a joint pmf $P_{XY}$ is defined as

$$A_{\epsilon,n} = \Big\{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n \text{ such that}$$
$$\left|-\tfrac{1}{n}\log P_X(x^n) - H(X)\right| < \epsilon,$$
$$\left|-\tfrac{1}{n}\log P_Y(y^n) - H(Y)\right| < \epsilon,$$
$$\left|-\tfrac{1}{n}\log P_{XY}(x^n, y^n) - H(X,Y)\right| < \epsilon\Big\}$$

where $P_{XY}(x^n, y^n) = \prod_{i=1}^n P_{XY}(x_i, y_i)$.






# The Joint AEP

Let $(X^n, Y^n)$ be a pair of sequences drawn i.i.d. according to $P_{XY}$, i.e.,

$$Pr(X^n = x^n, Y^n = y^n) = \prod_{i=1}^n P_{XY}(x_i, y_i), \quad \text{for all } (x^n, y^n)$$

Then for any $\epsilon > 0$:

1. $Pr((X^n, Y^n) \in A_{\epsilon,n}) \to 1$ as $n \to \infty$
2. $|A_{\epsilon,n}| \le 2^{n(H(X,Y)+\epsilon)}$
3. If $(\tilde{X}^n, \tilde{Y}^n)$ are a pair of sequences drawn i.i.d. according to $P_X P_Y$ [i.e., $\tilde{X}^n$ and $\tilde{Y}^n$ are independent with the same *marginals* as $P_{XY}$], then
$$Pr((\tilde{X}^n, \tilde{Y}^n) \in A_{\epsilon,n}) \le 2^{-n(I(X;Y)-3\epsilon)}$$

# Proof of Joint AEP

*Claim* 1:
When $(X^n, Y^n)$ are generated i.i.d. according to $P_{XY}$,

$$Pr\left(\left|-\frac{1}{n}\log P_X(x^n) - H(X)\right| < \epsilon\right) \to 1 \text{ as } n \to \infty,$$
$$Pr\left(\left|-\frac{1}{n}\log P_Y(y^n) - H(Y)\right| < \epsilon\right) \to 1 \text{ as } n \to \infty,$$
$$Pr\left(\left|-\frac{1}{n}\log P_{XY}(x^n, y^n) - H(X, Y)\right| < \epsilon\right) \to 1 \text{ as } n \to \infty.$$

The proof of the above is very similar to that of the AEP in Handout 1 and follows from the WLLN. Thus
$Pr((X^n, Y^n) \in A_{\epsilon,n}) \to 1$ as $n \to \infty$. $\square$



*Claim* 2:
We have

$$\begin{aligned} 1 &= \sum_{x^n, y^n} P_{XY}(x^n, y^n) \\ &\geq \sum_{(x^n,y^n)\in A_{\epsilon,n}} P_{XY}(x^n, y^n) \\ &\overset{(a)}{\geq} \sum_{(x^n,y^n)\in A_{\epsilon,n}} 2^{-n(H(X,Y)+\epsilon)} \\ &= 2^{-n(H(X,Y)+\epsilon)}\,|A_{\epsilon,n}| \end{aligned}$$

Hence $|A_{n,\epsilon}| \leq 2^{n(H(X,Y)+\epsilon)}$. (Inequality $(a)$ follows from the definition of the typical set.) $\square$

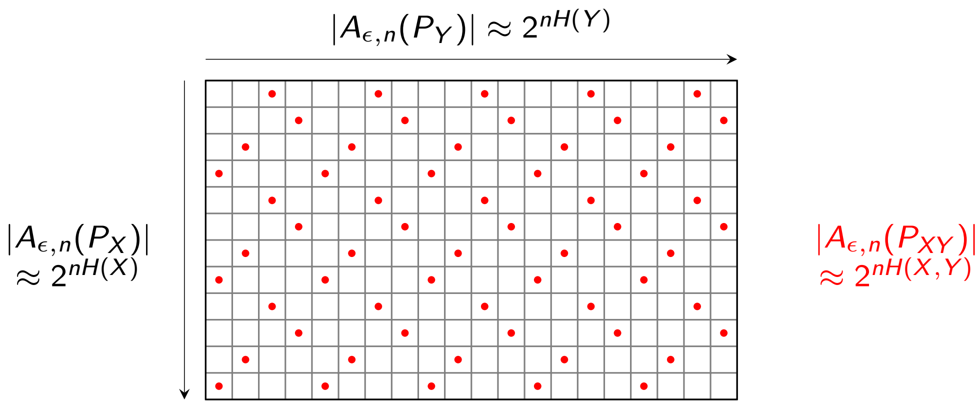
# Proof of Joint AEP contd.

*Claim* 3:

If $\tilde{X}^n$ and $\tilde{Y}^n$ are independent but have the same *marginals* as $X^n$ and $Y^n$, respectively, then

$$\begin{aligned} Pr((\tilde{X}^n, \tilde{Y}^n) \in A_{\epsilon,n}) &= \sum_{(x^n,y^n)\in A_{\epsilon,n}} P_X(x^n)P_Y(y^n) \\ &\leq 2^{n(H(X,Y)+\epsilon)} \cdot 2^{-n(H(X)-\epsilon)} \cdot 2^{-n(H(Y)-\epsilon)} \\ &= 2^{-n(I(X;Y)-3\epsilon)}. \end{aligned}$$

□






# The Probability of Error of a Code


**Rate $R$ code**

Recall that a $(2^{nR}, n)$ code for the channel $(\mathcal{X}, \mathcal{Y}, P_{Y|X})$ consists of a set of messages $\{1, \ldots, 2^{nR}\}$, an encoding function, and a decoding function.

The *maximal* probability of error of a $(2^{nR}, n)$ code is defined as

$$\max_{k\in\{1,\ldots,2^{nR}\}} Pr(\hat{W} \neq k \mid W = k)$$

The *average* probability of error of a $(2^{nR}, n)$ code is

$$\frac{1}{2^{nR}} \sum_{k=1}^{2^{nR}} Pr(\hat{W} \neq k \mid W = k)$$

$W$ and $\hat{W}$ denote the transmitted, and decoded messages respectively.

# The Channel Coding Theorem

$$W \longrightarrow \boxed{\text{Encoder}} \xrightarrow{X^n} \boxed{\text{DMC } P_{Y|X}} \xrightarrow{Y^n} \boxed{\text{Decoder}} \longrightarrow \hat{W}$$

**Theorem**

*"For a DMC with capacity $\mathcal{C}$, all rates less than $\mathcal{C}$ are achievable."*

*Specifically,*

1. *Fix $R < \mathcal{C}$ and pick any $\epsilon > 0$. Then, for all sufficiently large $n$ there exists an $(2^{nR}, n)$ code with maximal probability of error less than $\epsilon$.*
2. *Conversely, any sequence of $(2^{nR}, n)$ codes with maximal probability of error $P_e^{(n)} \to 0$ as $n \to \infty$ must have $R \leq \mathcal{C}$.*



# Proof of the Coding Theorem

We will first prove *achievability* of all rates $R < \mathcal{C}$ (the first part).

*Codebook Generation:*

- Fix rate $R < \mathcal{C}$ and input pmf $P_X$. We generate each of the $2^{nR}$ codewords independently according to the distribution

$$Pr(X^n(k) = (x_1, \ldots, x_n)) = \prod_{i=1}^{n} P_X(x_i) \quad \text{for } k = 1, \ldots, 2^{nR}.$$

- We can think of the codebook $\mathcal{B}$ as a $2^{nR} \times n$ matrix:

$$\mathcal{B} = \begin{bmatrix} X_1(1) & X_2(1) & \ldots & X_n(1) \\ X_1(2) & X_2(2) & \ldots & X_n(2) \\ \vdots & \vdots & \ddots & \vdots \\ X_1(2^{nR}) & X_2(2^{nR}) & \ldots & X_n(2^{nR}) \end{bmatrix}$$

- Each entry in the matrix is chosen i.i.d according to $P_X$. The probability that we generate a particular codebook $\{x^n(1), \ldots, x^n(2^{nR})\}$ is $\prod_{w=1}^{2^{nR}} \prod_{i=1}^{n} P_X(x_i(w))$

## Encoding

$$W \longrightarrow \boxed{\text{Encoder}} \xrightarrow{X^n} \boxed{\text{DMC } P_{Y|X}} \xrightarrow{Y^n} \boxed{\text{Decoder}} \longrightarrow \hat{W}$$

1. A code $\mathcal{B}$ is generated as described previously. The code is revealed to both sender and receiver, who also know the channel transition matrix $P_{Y|X}$.
2. To transmit message $W$, the encoder sends $X^n(W)$ over the channel.
3. The receiver receives a sequence $Y^n$ generated according to
$$\prod_{i=1}^{n} P_{Y|X}(Y_i \mid X_i(W)) \tag{1}$$
4. From $Y^n$, the receiver has to guess which message was sent. How?

- Assuming a uniform prior on the messages, the optimal decoding rule is *max-likelihood* decoding: decode the message $\hat{W}$ that maximises (1).
- But we'll use *joint typical* decoding, which is easier to analyse



## Joint Typicality Decoder

The decoder declares that the message $\hat{W}$ was sent if *both* the following conditions are satisfied:

- $(X^n(\hat{W}), Y^n)$ is jointly typical with respect to $P_X P_{Y|X}$.
- There exists no other message $W' \neq \hat{W}$ such that $(X^n(W'), Y^n)$ is jointly typical.

If no such $\hat{W}$ is found or there is more than one such, an error is declared.

## Analysing the probability of error

*Averages, and more averages ...*

- The average probability of error for a *given* codebook $\mathcal{B}$ is
$$\frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} Pr(\hat{W} \neq w \mid \mathcal{B}, W = w) \tag{2}$$
- Analysing this for a *specific* codebook is hard. So we will calculate the average of (2) over all codebooks, i.e.,
$$\bar{P}_e = \frac{1}{2^{nR}} \sum_{\mathcal{B}} \sum_{w=1}^{2^{nR}} Pr(\hat{W} \neq w \mid \mathcal{B}, W = w) Pr(\mathcal{B}). \tag{3}$$
$$\bar{P}_e = \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_{\mathcal{B}} Pr(\hat{W} \neq w \mid \mathcal{B}, W = w) Pr(\mathcal{B}). \tag{4}$$
- Recall that $Pr(\mathcal{B})$ is the probability corresponding to picking each symbol of $\mathcal{B}$ i.i.d. $\sim P_X$.
- Since all the messages are equally likely, we can assume that the first message is the transmitted one. Thus
$$\bar{P}_e = \sum_{\mathcal{B}} Pr(\hat{W} \neq 1 \mid \mathcal{B}, W = 1) Pr(\mathcal{B}). \tag{5}$$



## Error Analysis

Assuming $W = 1$ was transmitted, there are two sources of error:

1. $X^n(1)$ is not jointly typical with the output $Y^n$
2. $X^n(w)$ is jointly typical with $Y^n$ for some $w \neq 1$.

(Note: The joint typicality is with respect to $P_{XY}$)

- Let $E_k$ be the event that $X^n(k)$ and $Y^n$ are jointly typical.
- Then:
$$\begin{aligned} \bar{P}_e &= P(E_1^c \cup E_2 \cup \ldots \cup E_{2^{nR}}) \\ &\leq P(E_1^c) + P(E_2) + \ldots + P(E_{2^{nR}}) \end{aligned}$$

### 1) Showing $P(E_1^c)$ is small:

- Recall $X^n(1)$ i.i.d. $\sim P_X$.
- The channel generates $Y^n$ symbol by symbol according to $P_{Y|X}(Y_i \mid X_i(1))$ for $i = 1, \ldots, n$.
- Therefore $(X^n(1), Y^n)$ is generated i.i.d $\sim P_X P_{Y|X}$
- Joint AEP implies that $P(E_1^c) \leq \epsilon$ for sufficiently large $n$

## 2) Showing $P(E_2) + \ldots + P(E_{2^{nR}})$ is small:

For $k \neq 1$:

- $X^n(k)$ was generated *independently* from $X^n(1)$, and $Y^n$ is obtained by passing $X^n(1)$ through the channel.
- Hence $X^n(k)$ and $Y^n$ are independent for $k \neq 1$.
- Further, $X^n(k)$ is i.i.d. $\sim P_X$, and $Y^n$ is i.i.d. $\sim P_Y$.
- From the *Joint AEP*, the probability that $X^n(k)$ and $Y^n$ are jointly typical according to $P_{XY}$ is $\leq 2^{-n(I(X;Y)-3\epsilon)}$.

$$\Rightarrow \quad P(E_2) + \ldots + P(E_{2^{nR}}) \leq (2^{nR} - 1)\, 2^{-n(I(X;Y)-3\epsilon)}$$

Putting the two parts together:

$$\begin{aligned}\bar{P}_e &\leq P(E_1^c) + P(E_2) + \ldots + P(E_{2^{nR}}) \\ &\leq \epsilon + 2^{nR} 2^{-n(I(X;Y)-3\epsilon)} \\ &\stackrel{(a)}{\leq} \epsilon + \epsilon\end{aligned}$$

$(a)$ is true when $R < I(X;Y) - 3\epsilon$ and $n$ is sufficiently large.



*So far, we have shown that*:

For any $\epsilon > 0$, when $R < I(X;Y) - 3\epsilon$, the probability of error averaged over all messages *and* all codebooks is small, i.e.,

$$\bar{P}_e = \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_{\mathcal{B}} Pr(\hat{W} \neq w \mid \mathcal{B}, W = w) Pr(\mathcal{B}) \leq 2\epsilon$$

**Final Steps**:

1. Choose $P_X$ to be one that maximises $I(X;Y)$.
2. Get rid of average over codebooks: As $\bar{P}_e \leq 2\epsilon$, there exists at least one codebook $\mathcal{B}^*$ with

$$P_e(\mathcal{B}^*) = \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} Pr(\hat{W} \neq w \mid \mathcal{B}^*, W = w) \leq 2\epsilon$$

3. Throw away the worst half of the codewords in $\mathcal{B}^*$: For $\mathcal{B}^*$, the probability of error *averaged* over all messages is $\leq 2\epsilon$. Thus the probability of error must be $\leq 4\epsilon$ for *at least* half the messages.

## The Final Code

- The number of codewords in this improved version of $\mathcal{B}^*$ is $2^{nR}/2$. Its rate is

$$\frac{\log(2^{nR}/2)}{n} = R - \frac{1}{n}.$$

- Since $R$ is any rate less than $\mathcal{C} - 3\epsilon$, we have shown the existence of a code with rate

$$\mathcal{C} - 4\epsilon - \frac{1}{n}$$

  whose *maximal* probability of error satisfies

$$\max_{w} Pr(\hat{W} \neq w \mid W = w) \leq 4\epsilon$$

- Since $\epsilon > 0$ is an arbitrary constant, we have shown that for *any* $R < \mathcal{C}$, for sufficiently large $n$ there exists a code with arbitrarily small maximal error probability.

This proves the first part of the channel coding theorem.



## Summary

**Key ideas** in the proof of achievability:

- Allow an arbitrarily small but non-zero probability of error
- Use the channel many times in succession, so that the law of large numbers comes into effect (large $n$)
- *Random Coding*: Calculate the average probability of error over a random choice of codebooks, which can then be used to show the existence of at least one good code

**Proof of converse (next handout)**

To show that we cannot achieve rates $> \mathcal{C}$, need two new tools:

- Data Processing Inequality
- Fano's Inequality