# Previously...

- Identification/Decision trees
- Random forests
- Law of Large Number
- Asymptotic equipartition (AEP) and typical sequences

# This time

- Joint typical sequences
- Covering and Packing Lemmas
- Channel coding setup
- Channel coding rate
- Channel capacity
- Channel Coding Theorem

# Jointly typical sequences

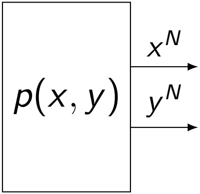
For a pair of sequences $x^N$ and $y^N$, we say that they are jointly typical if

$$2^{-N(H(X,Y)+\epsilon)} \leq p(x^N, y^N) \leq 2^{-N(H(X,Y)-\epsilon)}$$

and $x^N$ and $y^N$ themselves are typical

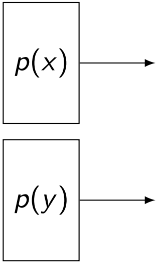
As in the single sequence case,

- Any sequence pair drawing from a joint source $p(x,y)$ is essentially jointly typical
- There are $\sim 2^{NH(X,Y)}$ jointly typical sequences



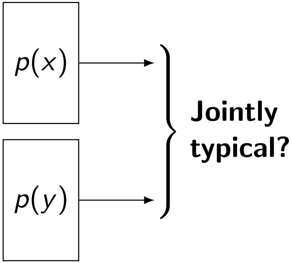
# Joint typicality of independent seqences

- Given sequences $X^N$ and $Y^N$ independently drawn from discrete memoryless sources $p(x)$ and $p(y)$



# Joint typicality of independent seqences

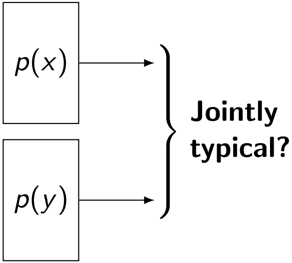
- Given sequences $X^N$ and $Y^N$ independently drawn from discrete memoryless sources $p(x)$ and $p(y)$
- What is the probability that $X^N$ and $Y^N$ are jointly typical?



# Joint typicality of independent seqences

- Given sequences $X^N$ and $Y^N$ independently drawn from discrete memoryless sources $p(x)$ and $p(y)$
- What is the probability that $X^N$ and $Y^N$ are jointly typical?

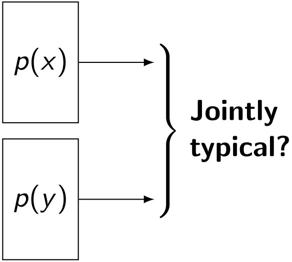
$$Pr((X^N, Y^N) \in \mathcal{A}_\epsilon^{(N)})$$
$$= \sum_{\{(x^N,y^N)|(x^N,y^N)\in\mathcal{A}_\epsilon^{(N)}\}} p(x^N, y^N)$$



# Joint typicality of independent seqences

- Given sequences $X^N$ and $Y^N$ independently drawn from discrete memoryless sources $p(x)$ and $p(y)$
- What is the probability that $X^N$ and $Y^N$ are jointly typical?

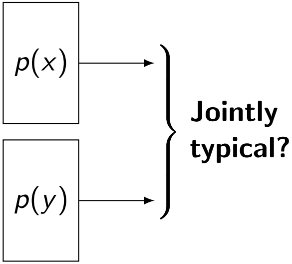
$$Pr((X^N, Y^N) \in \mathcal{A}_\epsilon^{(N)})$$
$$= \sum_{\{(x^N,y^N)|(x^N,y^N)\in\mathcal{A}_\epsilon^{(N)}\}} p(x^N, y^N)$$
$$= \sum_{\{(x^N,y^N)|(x^N,y^N)\in\mathcal{A}_\epsilon^{(N)}\}} p(x^N)p(y^N)$$



# Joint typicality of independent seqences

- Given sequences $X^N$ and $Y^N$ independently drawn from discrete memoryless sources $p(x)$ and $p(y)$
- What is the probability that $X^N$ and $Y^N$ are jointly typical?

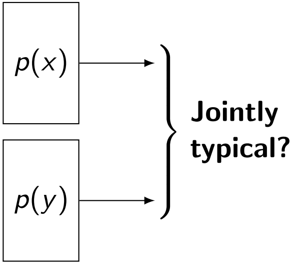
$$
\begin{aligned}
&Pr((X^N, Y^N) \in \mathcal{A}_\epsilon^{(N)}) \\
&= \sum_{\{(x^N,y^N)|(x^N,y^N)\in\mathcal{A}_\epsilon^{(N)}\}} p(x^N, y^N) \\
&= \sum_{\{(x^N,y^N)|(x^N,y^N)\in\mathcal{A}_\epsilon^{(N)}\}} p(x^N)p(y^N) \\
&\leq \sum_{\{(x^N,y^N)|(x^N,y^N)\in\mathcal{A}_\epsilon^{(N)}\}} 2^{-N(H(X)-\epsilon)}2^{-N(H(Y)-\epsilon)}
\end{aligned}
$$



# Joint typicality of independent seqences

- Given sequences $X^N$ and $Y^N$ independently drawn from discrete memoryless sources $p(x)$ and $p(y)$
- What is the probability that $X^N$ and $Y^N$ are jointly typical?

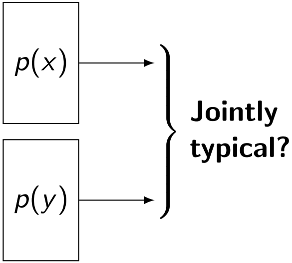
$$
\begin{aligned}
&Pr((X^N, Y^N) \in \mathcal{A}_\epsilon^{(N)}) \\
&= \sum_{\{(x^N,y^N)|(x^N,y^N)\in\mathcal{A}_\epsilon^{(N)}\}} p(x^N, y^N) \\
&= \sum_{\{(x^N,y^N)|(x^N,y^N)\in\mathcal{A}_\epsilon^{(N)}\}} p(x^N)p(y^N) \\
&\leq \sum_{\{(x^N,y^N)|(x^N,y^N)\in\mathcal{A}_\epsilon^{(N)}\}} 2^{-N(H(X)-\epsilon)}2^{-N(H(Y)-\epsilon)} \\
&\leq 2^{-N(I(X;Y)-3\epsilon)}
\end{aligned}
$$



# Joint typicality of independent seqences

- Given sequences $X^N$ and $Y^N$ independently drawn from discrete memoryless sources $p(x)$ and $p(y)$
- What is the probability that $X^N$ and $Y^N$ are jointly typical?

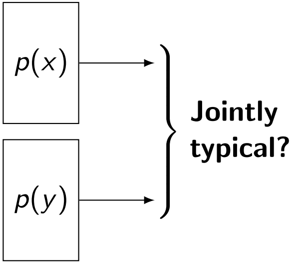
$$Pr((X^N, Y^N) \in \mathcal{A}_\epsilon^{(N)})$$
$$= \sum_{\{(x^N,y^N)|(x^N,y^N)\in\mathcal{A}_\epsilon^{(N)}\}} p(x^N, y^N)$$
$$= \sum_{\{(x^N,y^N)|(x^N,y^N)\in\mathcal{A}_\epsilon^{(N)}\}} p(x^N)p(y^N)$$



# Joint typicality of independent seqences

- Given sequences $X^N$ and $Y^N$ independently drawn from discrete memoryless sources $p(x)$ and $p(y)$
- What is the probability that $X^N$ and $Y^N$ are jointly typical?

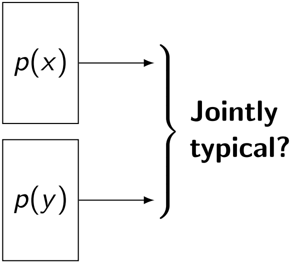
$$Pr((X^N, Y^N) \in \mathcal{A}_\epsilon^{(N)})$$
$$= \sum_{\{(x^N,y^N)|(x^N,y^N)\in\mathcal{A}_\epsilon^{(N)}\}} p(x^N, y^N)$$
$$= \sum_{\{(x^N,y^N)|(x^N,y^N)\in\mathcal{A}_\epsilon^{(N)}\}} p(x^N)p(y^N)$$
$$\geq \sum_{\{(x^N,y^N)|(x^N,y^N)\in\mathcal{A}_\epsilon^{(N)}\}} 2^{-N(H(X)+\epsilon)}2^{-N(H(Y)+\epsilon)}$$



# Joint typicality of independent seqences

- Given sequences $X^N$ and $Y^N$ independently drawn from discrete memoryless sources $p(x)$ and $p(y)$
- What is the probability that $X^N$ and $Y^N$ are jointly typical?

$$
\begin{aligned}
&Pr((X^N, Y^N) \in \mathcal{A}_\epsilon^{(N)}) \\
&= \sum_{\{(x^N,y^N)|(x^N,y^N)\in\mathcal{A}_\epsilon^{(N)}\}} p(x^N, y^N) \\
&= \sum_{\{(x^N,y^N)|(x^N,y^N)\in\mathcal{A}_\epsilon^{(N)}\}} p(x^N)p(y^N) \\
&\geq \sum_{\{(x^N,y^N)|(x^N,y^N)\in\mathcal{A}_\epsilon^{(N)}\}} 2^{-N(H(X)+\epsilon)}2^{-N(H(Y)+\epsilon)} \\
&\geq (1-\delta)2^{-N(I(X;Y)+3\epsilon)}
\end{aligned}
$$



# Packing lemma

- Instead of drawing one $Y^N$ sequences, let us draw $M$ of them

# Packing lemma

- Instead of drawing one $Y^N$ sequences, let us draw $M$ of them
- The probability of any of the sequence to be jointly typical with $X^N$ is bounded by

$$Pr(\text{Any one of } M \ Y^N \text{ jointly typical with } X^N)$$

# Packing lemma

- Instead of drawing one $Y^N$ sequences, let us draw $M$ of them
- The probability of any of the sequence to be jointly typical with $X^N$ is bounded by

$$\begin{aligned} & Pr(\text{Any one of } M \ Y^N \text{ jointly typical with } X^N) \\ \leq & M Pr((X^N, Y^N) \in \mathcal{A}_\epsilon^N(X, Y)) \\ \leq & M 2^{-N(I(X;Y)-3\epsilon)} \end{aligned}$$

# Packing lemma

- Instead of drawing one $Y^N$ sequences, let us draw $M$ of them
- The probability of any of the sequence to be jointly typical with $X^N$ is bounded by

$$
\begin{aligned}
&Pr(\text{Any one of } M \; Y^N \text{ jointly typical with } X^N) \\
\leq &MPr((X^N, Y^N) \in \mathcal{A}_\epsilon^N(X, Y)) \\
\leq &M2^{-N(I(X;Y)-3\epsilon)} \\
\leq &2^{-N(I(X;Y)-R-3\epsilon)}
\end{aligned}
$$

where $2^{NR} = M$

# Packing lemma

- Instead of drawing one $Y^N$ sequences, let us draw $M$ of them
- The probability of any of the sequence to be jointly typical with $X^N$ is bounded by

$$
\begin{aligned}
& Pr(\text{Any one of } M\ Y^N \text{ jointly typical with } X^N) \\
\leq & M Pr((X^N, Y^N) \in \mathcal{A}_\epsilon^N(X, Y)) \\
\leq & M 2^{-N(I(X;Y)-3\epsilon)} \\
\leq & 2^{-N(I(X;Y)-R-3\epsilon)} \to 0 \text{ as } N \to \infty \text{ and } I(X;Y) - 3\epsilon > R,
\end{aligned}
$$

where $2^{NR} = M$

# Packing lemma

- Instead of drawing one $Y^N$ sequences, let us draw $M$ of them
- The probability of any of the sequence to be jointly typical with $X^N$ is bounded by

$$
\begin{aligned}
&Pr(\text{Any one of } M\ Y^N \text{ jointly typical with } X^N) \\
\leq & M Pr((X^N, Y^N) \in \mathcal{A}_\epsilon^N(X, Y)) \\
\leq & M 2^{-N(I(X;Y)-3\epsilon)} \\
\leq & 2^{-N(I(X;Y)-R-3\epsilon)} \rightarrow 0 \text{ as } N \rightarrow \infty \text{ and } I(X;Y) - 3\epsilon > R,
\end{aligned}
$$

where $2^{NR} = M$

Since $\epsilon$ can be made arbitrarily small as $N$ increases, as long as $I(X;Y) > R$, we can find a sufficiently large $N$ so that we can "pack" the $M$ $Y^N$ with $X^N$ and none of the $Y^N$ will be jointly typical with $X^N$

# Covering lemma

- Again, draw $M(=2^{NR})$ $Y^N$ sequences
- Under what condition that *at least one* $Y^N$ jointly typical with $X^N$?

# Covering lemma

- Again, draw $M(=2^{NR})$ $Y^N$ sequences
- Under what condition that *at least one* $Y^N$ jointly typical with $X^N$?

$$Pr((X^N(m), Y^N) \notin \mathcal{A}_\epsilon^{(N)}(X, Y) \text{ for all } m)$$

# Covering lemma

- Again, draw $M(=2^{NR})$ $Y^N$ sequences
- Under what condition that *at least one* $Y^N$ jointly typical with $X^N$?

$$Pr((X^N(m), Y^N) \notin \mathcal{A}_\epsilon^{(N)}(X, Y) \text{ for all } m)$$
$$= \prod_{m=1}^{M} Pr((X^N(m), Y^N) \notin \mathcal{A}_\epsilon^{(N)}(Y, X))$$

# Covering lemma

- Again, draw $M(=2^{NR})$ $Y^N$ sequences
- Under what condition that *at least one* $Y^N$ jointly typical with $X^N$?

$$Pr((X^N(m), Y^N) \notin \mathcal{A}_\epsilon^{(N)}(X, Y) \text{ for all } m)$$
$$= \prod_{m=1}^{M} Pr((X^N(m), Y^N) \notin \mathcal{A}_\epsilon^{(N)}(Y, X))$$
$$= \prod_{m=1}^{M} \left[1 - Pr((X^N(m), Y^N) \in \mathcal{A}_\epsilon^{(N)}(Y, X))\right]$$

# Covering lemma

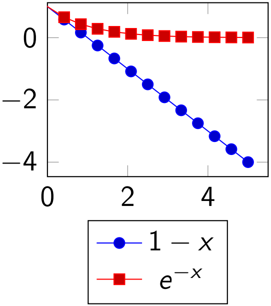
- Again, draw $M(=2^{NR})$ $Y^N$ sequences
- Under what condition that *at least one* $Y^N$ jointly typical with $X^N$?

$$
\begin{aligned}
&Pr((X^N(m), Y^N) \notin \mathcal{A}_\epsilon^{(N)}(X, Y) \text{ for all } m) \\
=&\prod_{m=1}^{M} Pr((X^N(m), Y^N) \notin \mathcal{A}_\epsilon^{(N)}(Y, X)) \\
=&\prod_{m=1}^{M} \left[1 - Pr((X^N(m), Y^N) \in \mathcal{A}_\epsilon^{(N)}(Y, X))\right] \\
\leq&(1-(1-\delta)2^{-N(I(Y;X)+3\epsilon)})^M
\end{aligned}
$$

# Covering lemma

- Again, draw $M(=2^{NR})$ $Y^N$ sequences
- Under what condition that *at least one* $Y^N$ jointly typical with $X^N$?

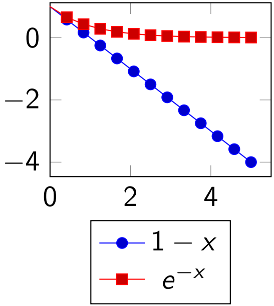
$$
\begin{aligned}
& Pr((X^N(m), Y^N) \notin \mathcal{A}_\epsilon^{(N)}(X,Y) \text{ for all } m) \\
=& \prod_{m=1}^{M} Pr((X^N(m), Y^N) \notin \mathcal{A}_\epsilon^{(N)}(Y,X)) \\
=& \prod_{m=1}^{M} \left[1 - Pr((X^N(m), Y^N) \in \mathcal{A}_\epsilon^{(N)}(Y,X))\right] \\
\le& (1-(1-\delta)2^{-N(I(Y;X)+3\epsilon)})^M \\
\le& \exp(-M(1-\delta)2^{-N(I(Y;X)+3\epsilon)})
\end{aligned}
$$



# Covering lemma

- Again, draw $M(=2^{NR})$ $Y^N$ sequences
- Under what condition that *at least one* $Y^N$ jointly typical with $X^N$?

$$
\begin{aligned}
& Pr((X^N(m), Y^N) \notin \mathcal{A}_\epsilon^{(N)}(X,Y) \text{ for all } m) \\
=& \prod_{m=1}^{M} Pr((X^N(m), Y^N) \notin \mathcal{A}_\epsilon^{(N)}(Y,X)) \\
=& \prod_{m=1}^{M} \left[1 - Pr((X^N(m), Y^N) \in \mathcal{A}_\epsilon^{(N)}(Y,X))\right] \\
\leq& (1-(1-\delta)2^{-N(I(Y;X)+3\epsilon)})^M \\
\leq& \exp(-M(1-\delta)2^{-N(I(Y;X)+3\epsilon)}) \\
\leq& \exp(-(1-\delta)2^{-N(I(Y;X)-R+3\epsilon)}) \to 0 \text{ as } N \to \infty \text{ and } R > I(X;Y)
\end{aligned}
$$



# Summary of packing lemma and covering lemma

## Packing Lemma

We can "pack" $M = 2^{NR}$ (with $R < I(X;Y)$) $x^N$ together without being jointly typical with $y^N$

## Covering Lemma

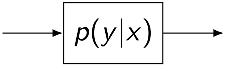
We can "cover" with $M = 2^{NR}$ (with $R > I(X;Y)$) $x^N$ such that at least one $x^N$ being jointly typical with $y^N$

## Remark

- Packing lemma is useful in the proof of channel coding theorem
- Covering lemma is useful in the proof of rate-distortion theorem

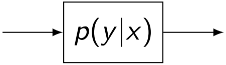
We will look into the above applications later in this course

# Channel coding setup



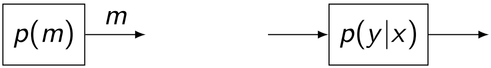

- As the name suggests, the output of a discrete memoryless channel (DMS) only depends on the current input (thus no memoryless). And both its input $X$ and output $Y$ are characterized by the conditional probability $p(y|x)$

# Channel coding setup



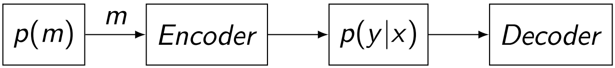

- As the name suggests, the output of a discrete memoryless channel (DMS) only depends on the current input (thus no memoryless). And both its input $X$ and output $Y$ are characterized by the conditional probability $p(y|x)$
- Given an input sequence $x^N = x_1, \cdots, x_N$, the probability of getting an output sequence $y^N = y_1, \cdots, y_N$ is $p(y^N|x^N) = \prod_{i=1}^N p(y_i|x_i)$

# Channel coding setup



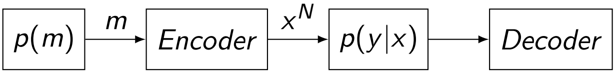

- As the name suggests, the output of a discrete memoryless channel (DMS) only depends on the current input (thus no memoryless). And both its input $X$ and output $Y$ are characterized by the conditional probability $p(y|x)$
- Given an input sequence $x^N = x_1, \cdots, x_N$, the probability of getting an output sequence $y^N = y_1, \cdots, y_N$ is $p(y^N|x^N) = \prod_{i=1}^N p(y_i|x_i)$
- Given a message $m$ (say generated from a distribution $p(m)$)

# Channel coding setup



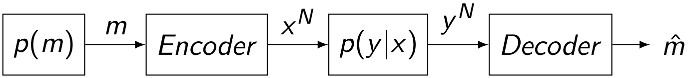

- As the name suggests, the output of a discrete memoryless channel (DMS) only depends on the current input (thus no memoryless). And both its input $X$ and output $Y$ are characterized by the conditional probability $p(y|x)$
- Given an input sequence $x^N = x_1, \cdots, x_N$, the probability of getting an output sequence $y^N = y_1, \cdots, y_N$ is $p(y^N|x^N) = \prod_{i=1}^N p(y_i|x_i)$
- Given a message $m$ (say generated from a distribution $p(m)$)
  - We will have an encoder decoder pair

# Channel coding setup



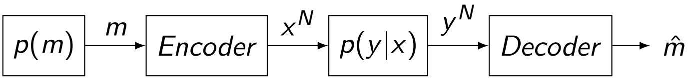

- As the name suggests, the output of a discrete memoryless channel (DMS) only depends on the current input (thus no memoryless). And both its input $X$ and output $Y$ are characterized by the conditional probability $p(y|x)$
- Given an input sequence $x^N = x_1, \cdots, x_N$, the probability of getting an output sequence $y^N = y_1, \cdots, y_N$ is $p(y^N|x^N) = \prod_{i=1}^N p(y_i|x_i)$
- Given a message $m$ (say generated from a distribution $p(m)$)
    - We will have an encoder decoder pair
    - The encoder will convert $m$ to $x^N$ suitable for transmission

# Channel coding setup



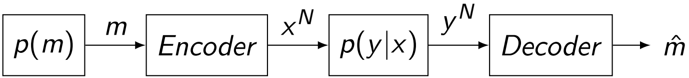

- As the name suggests, the output of a discrete memoryless channel (DMS) only depends on the current input (thus no memoryless). And both its input $X$ and output $Y$ are characterized by the conditional probability $p(y|x)$
- Given an input sequence $x^N = x_1, \cdots, x_N$, the probability of getting an output sequence $y^N = y_1, \cdots, y_N$ is $p(y^N|x^N) = \prod_{i=1}^N p(y_i|x_i)$
- Given a message $m$ (say generated from a distribution $p(m)$)
  - We will have an encoder decoder pair
  - The encoder will convert $m$ to $x^N$ suitable for transmission
  - Decoder will try to extracted the message from the channel output $y^N$

# Channel coding rate




The channel coding rate is defined as number of bits of message can be sent per channel use

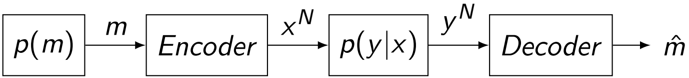
# Channel coding rate




The channel coding rate is defined as number of bits of message can be sent per channel use

- Since there is $H(M)$ bits of information for each message $M$ sent

# Channel coding rate




The channel coding rate is defined as number of bits of message can be sent per channel use

- Since there is $H(M)$ bits of information for each message $M$ sent
- $R = \frac{H(M)}{N}$

# Channel capacity

- By Shannon's channel coding theorem, the capacity of the channel (will be shown later) is given by

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

# Channel capacity

- By Shannon's channel coding theorem, the capacity of the channel (will be shown later) is given by

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

- This means that as long as the rate $R$ is less than the capacity $C$, we can find encoder-decoder pair such that the decoding error ($Pr(\hat{M} \neq M)$) can be made arbitrarily small

# Channel capacity

- By Shannon's channel coding theorem, the capacity of the channel (will be shown later) is given by

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

- This means that as long as the rate $R$ is less than the capacity $C$, we can find encoder-decoder pair such that the decoding error $(Pr(\hat{M} \neq M))$ can be made arbitrarily small
- On the other hand, if $R$ is larger than the capacity $C$, no matter how we try, it is impossible to recontruct $m$ error free

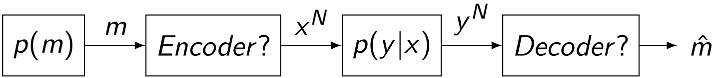
# Channel capacity

- By Shannon's channel coding theorem, the capacity of the channel (will be shown later) is given by

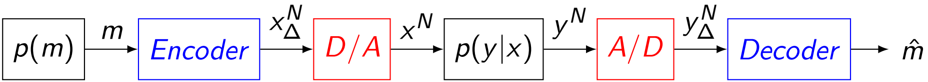
$$C = \max_{p(x)} I(X;Y)$$

- This means that as long as the rate $R$ is less than the capacity $C$, we can find encoder-decoder pair such that the decoding error $(Pr(\hat{M} \neq M))$ can be made arbitrarily small
- On the other hand, if $R$ is larger than the capacity $C$, no matter how we try, it is impossible to recontruct $m$ error free
- An intuitive interpretation is that the amount of information can be passed through a channel is just mutual information between the input and output. And since we can pick the statistics of our input, we may make our choice wisely and maximize the mutual information. And the maximum that we can attain is the capacity

# Continuous channel

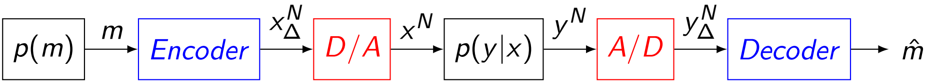


# Continuous channel




- For continuous channel, we can create a "pseudo" discrete channel using A/D and D/A converters

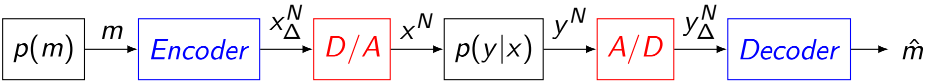
# Continuous channel




- For continuous channel, we can create a "pseudo" discrete channel using A/D and D/A converters
- The maximum information that can pass through the channel will then be

$$\begin{aligned} C_\Delta &= \max_{p(x)} I(X_\Delta; Y_\Delta) = \max_{p(x)} H(Y_\Delta) - H(Y_\Delta|X_\Delta) \\ &\approx \max_{p(x)} h(Y) - \log \Delta - h(Y|X_\Delta) + \log \Delta \\ &\approx \max_{p(x)} h(Y) - h(Y|X) = \max_{p(x)} I(X; Y) \end{aligned}$$

# Continuous channel




- For continuous channel, we can create a "pseudo" discrete channel using A/D and D/A converters
- The maximum information that can pass through the channel will then be

$$\begin{aligned} C_\Delta &= \max_{p(x)} I(X_\Delta; Y_\Delta) = \max_{p(x)} H(Y_\Delta) - H(Y_\Delta|X_\Delta) \\ &\approx \max_{p(x)} h(Y) - \log\Delta - h(Y|X_\Delta) + \log\Delta \\ &\approx \max_{p(x)} h(Y) - h(Y|X) = \max_{p(x)} I(X;Y) \end{aligned}$$

- As $\Delta \to 0$, $C = \max_{p(x)} I(X;Y)$. So expression is completely the same as the discrete case

# Example: Binary symmetric channel

- Both input and output are binary (say take value 0 or 1)

# Example: Binary symmetric channel

- Both input and output are binary (say take value 0 or 1)
- The channel is symmetric in the sense that

$$p_{Y|X}(1|0) = p_{Y|X}(0|1) = p$$

and

$$p_{Y|X}(0|0) = p_{Y|X}(1|1) = 1 - p,$$

where $p$ is known to be the cross-over probability

# Example: Binary symmetric channel

- Both input and output are binary (say take value 0 or 1)
- The channel is symmetric in the sense that

$$p_{Y|X}(1|0) = p_{Y|X}(0|1) = p$$

and

$$p_{Y|X}(0|0) = p_{Y|X}(1|1) = 1 - p,$$

where $p$ is known to be the cross-over probability

- Capacity is given by

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

# Example: Binary symmetric channel

- Both input and output are binary (say take value 0 or 1)
- The channel is symmetric in the sense that

$$p_{Y|X}(1|0) = p_{Y|X}(0|1) = p$$

and

$$p_{Y|X}(0|0) = p_{Y|X}(1|1) = 1 - p,$$

where $p$ is known to be the cross-over probability

- Capacity is given by

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

# Example: Binary symmetric channel

- Both input and output are binary (say take value 0 or 1)
- The channel is symmetric in the sense that

$$p_{Y|X}(1|0) = p_{Y|X}(0|1) = p$$

and

$$p_{Y|X}(0|0) = p_{Y|X}(1|1) = 1 - p,$$

where $p$ is known to be the cross-over probability

- Capacity is given by

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

# Example: Binary symmetric channel

- Both input and output are binary (say take value 0 or 1)
- The channel is symmetric in the sense that

$$p_{Y|X}(1|0) = p_{Y|X}(0|1) = p$$

and

$$p_{Y|X}(0|0) = p_{Y|X}(1|1) = 1 - p,$$

where $p$ is known to be the cross-over probability

- Capacity is given by

$$\begin{aligned} C &= \max_{p(x)} I(X;Y) \\ &= \max_{p(x)} H(Y) - H(Y|X) \\ &= \max_{p(x)} H(Y) - H(p) = 1 - H(p) \end{aligned}$$

# Example: Gaussian channel

The channel output $Y = X + Z$, where $Z$ is a zero-mean Gaussian noise (independent of the input $X$)

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

# Example: Gaussian channel

The channel output $Y = X + Z$, where $Z$ is a zero-mean Gaussian noise (independent of the input $X$)

$$\begin{aligned} C &= \max_{p(x)} I(X;Y) \\ &= \max_{p(x)} h(Y) - h(Y|X) \end{aligned}$$

# Example: Gaussian channel

The channel output $Y = X + Z$, where $Z$ is a zero-mean Gaussian noise (independent of the input $X$)

$$\begin{aligned} C &= \max_{p(x)} I(X;Y) \\ &= \max_{p(x)} h(Y) - h(Y|X) = \max_{p(x)} h(Y) - h(X+Z|X) \end{aligned}$$

# Example: Gaussian channel

The channel output $Y = X + Z$, where $Z$ is a zero-mean Gaussian noise (independent of the input $X$)

$$\begin{aligned} C &= \max_{p(x)} I(X;Y) \\ &= \max_{p(x)} h(Y) - h(Y|X) = \max_{p(x)} h(Y) - h(X+Z|X) \\ &= \max_{p(x)} h(Y) - h(Z|X) \end{aligned}$$

# Example: Gaussian channel

The channel output $Y = X + Z$, where $Z$ is a zero-mean Gaussian noise (independent of the input $X$)

$$
\begin{aligned}
C &= \max_{p(x)} I(X;Y) \\
&= \max_{p(x)} h(Y) - h(Y|X) = \max_{p(x)} h(Y) - h(X+Z|X) \\
&= \max_{p(x)} h(Y) - h(Z|X) = \max_{p(x)} h(Y) - h(Z)
\end{aligned}
$$

# Example: Gaussian channel

The channel output $Y = X + Z$, where $Z$ is a zero-mean Gaussian noise (independent of the input $X$)

$$
\begin{aligned}
C &= \max_{p(x)} I(X;Y) \\
&= \max_{p(x)} h(Y) - h(Y|X) = \max_{p(x)} h(Y) - h(X+Z|X) \\
&= \max_{p(x)} h(Y) - h(Z|X) = \max_{p(x)} h(Y) - h(Z) \\
&= \max_{p(x)} h(Y) - \frac{1}{2}\log 2\pi e \sigma_Z^2
\end{aligned}
$$

# Example: Gaussian channel

The channel output $Y = X + Z$, where $Z$ is a zero-mean Gaussian noise (independent of the input $X$)

$$
\begin{aligned}
C &= \max_{p(x)} I(X;Y) \\
&= \max_{p(x)} h(Y) - h(Y|X) = \max_{p(x)} h(Y) - h(X+Z|X) \\
&= \max_{p(x)} h(Y) - h(Z|X) = \max_{p(x)} h(Y) - h(Z) \\
&= \max_{p(x)} h(Y) - \frac{1}{2}\log 2\pi e\sigma_Z^2 = \frac{1}{2}\log 2\pi e\sigma_Y^2 - \frac{1}{2}\log 2\pi e\sigma_Z^2
\end{aligned}
$$

# Example: Gaussian channel

The channel output $Y = X + Z$, where $Z$ is a zero-mean Gaussian noise (independent of the input $X$)

$$
\begin{aligned}
C &= \max_{p(x)} I(X;Y) \\
&= \max_{p(x)} h(Y) - h(Y|X) = \max_{p(x)} h(Y) - h(X+Z|X) \\
&= \max_{p(x)} h(Y) - h(Z|X) = \max_{p(x)} h(Y) - h(Z) \\
&= \max_{p(x)} h(Y) - \frac{1}{2}\log 2\pi e \sigma_Z^2 = \frac{1}{2}\log 2\pi e \sigma_Y^2 - \frac{1}{2}\log 2\pi e \sigma_Z^2 \\
&= \frac{1}{2}\log \frac{\sigma_X^2 + \sigma_Z^2}{\sigma_Z^2} = \frac{1}{2}\log\left(1 + \frac{\sigma_X^2}{\sigma_Z^2}\right)
\end{aligned}
$$

# Example: Gaussian channel

The channel output $Y = X + Z$, where $Z$ is a zero-mean Gaussian noise (independent of the input $X$)

$$
\begin{aligned}
C &= \max_{p(x)} I(X;Y) \\
&= \max_{p(x)} h(Y) - h(Y|X) = \max_{p(x)} h(Y) - h(X+Z|X) \\
&= \max_{p(x)} h(Y) - h(Z|X) = \max_{p(x)} h(Y) - h(Z) \\
&= \max_{p(x)} h(Y) - \frac{1}{2}\log 2\pi e \sigma_Z^2 = \frac{1}{2}\log 2\pi e \sigma_Y^2 - \frac{1}{2}\log 2\pi e \sigma_Z^2 \\
&= \frac{1}{2}\log \frac{\sigma_X^2 + \sigma_Z^2}{\sigma_Z^2} = \frac{1}{2}\log\left(1 + \frac{\sigma_X^2}{\sigma_Z^2}\right) = \frac{1}{2}\log(1 + SNR),
\end{aligned}
$$

where $SNR$ is the signal to noise ratio

# Example: Additive White Gaussian Noise (AWGN) channel

Consider an AWGN channel with bandwidth $W$ and two-sided power spectrum density of $N_0/2$

# Example: Additive White Gaussian Noise (AWGN) channel

Consider an AWGN channel with bandwidth $W$ and two-sided power spectrum density of $N_0/2$

- From the result of Nyquist and Shannon, a signal of bandwidth $W$ will need to at least $2W$ samples per second to be fully reconstructed

# Example: Additive White Gaussian Noise (AWGN) channel

Consider an AWGN channel with bandwidth $W$ and two-sided power spectrum density of $N_0/2$

- From the result of Nyquist and Shannon, a signal of bandwidth $W$ will need to at least $2W$ samples per second to be fully reconstructed
- Per each second, $2W$ samples needed to recover the signal

# Example: Additive White Gaussian Noise (AWGN) channel

Consider an AWGN channel with bandwidth $W$ and two-sided power spectrum density of $N_0/2$

- From the result of Nyquist and Shannon, a signal of bandwidth $W$ will need to at least $2W$ samples per second to be fully reconstructed
- Per each second, $2W$ samples needed to recover the signal
- Per each second, $2W$ degrees of freedom exists $\Rightarrow 2W$ parallel Gaussian channel per second

# Example: Additive White Gaussian Noise (AWGN) channel

Consider an AWGN channel with bandwidth $W$ and two-sided power spectrum density of $N_0/2$

- From the result of Nyquist and Shannon, a signal of bandwidth $W$ will need to at least $2W$ samples per second to be fully reconstructed
- Per each second, $2W$ samples needed to recover the signal
- Per each second, $2W$ degrees of freedom exists $\Rightarrow 2W$ parallel Gaussian channel per second
- Given $N_0$, $SNR = \frac{\sigma_X^2}{WN_0} = \frac{P}{WN_0}$

# Example: Additive White Gaussian Noise (AWGN) channel

Consider an AWGN channel with bandwidth $W$ and two-sided power spectrum density of $N_0/2$

- From the result of Nyquist and Shannon, a signal of bandwidth $W$ will need to at least $2W$ samples per second to be fully reconstructed
- Per each second, $2W$ samples needed to recover the signal
- Per each second, $2W$ degrees of freedom exists $\Rightarrow 2W$ parallel Gaussian channel per second
- Given $N_0$, $SNR = \frac{\sigma_X^2}{WN_0} = \frac{P}{WN_0}$

$$C = 2W\frac{1}{2}\log(1+SNR) = W\log\left(1+\frac{P}{WN_0}\right)$$

# Codebook construction

## Forward statement

If the code rate $R < C = \max_{p(x)} I(X;Y)$, according to the Channel Coding Theorem, we should be able to find a code with encoding mapping $\mathbf{c} : m \in \{1, 2, \cdots, 2^{NR}\} \rightarrow \{0, 1\}^N$ and the error probability of transmitting any message $m \in \{1, 2, \cdots, 2^{NR}\}$, $p_e(m)$, is arbitrarily small

# Codebook construction

**Forward statement**

If the code rate $R < C = \max_{p(x)} I(X;Y)$, according to the Channel Coding Theorem, we should be able to find a code with encoding mapping $\mathbf{c} : m \in \{1, 2, \cdots, 2^{NR}\} \rightarrow \{0,1\}^N$ and the error probability of transmitting any message $m \in \{1, 2, \cdots, 2^{NR}\}$, $p_e(m)$, is arbitrarily small

- The main tool of the proof is random coding

# Codebook construction

## Forward statement

If the code rate $R < C = \max_{p(x)} I(X;Y)$, according to the Channel Coding Theorem, we should be able to find a code with encoding mapping $\mathbf{c} : m \in \{1, 2, \cdots, 2^{NR}\} \to \{0, 1\}^N$ and the error probability of transmitting any message $m \in \{1, 2, \cdots, 2^{NR}\}$, $p_e(m)$, is arbitrarily small

- The main tool of the proof is random coding
- Let $p^*(x) = \arg\max_{p(x)} I(X;Y)$. Generate codewords from the DMS $p^*(x)$ by sampling $2^n$ length-$n$ sequences from the source:

$$
\begin{aligned}
\mathbf{c}(1) =&(x_1(1), x_2(1), \cdots, x_N(1))\\
\mathbf{c}(2) =&(x_1(2), x_2(2), \cdots, x_N(2))\\
&\cdots\\
\mathbf{c}(2^{NR}) =&(x_1(2^{NR}), x_2(2^{NR}), \cdots, x_N(2^{NR}))
\end{aligned}
$$

# Encoding and decoding

The encoding and decoding procedures will be as follows.

# Encoding and decoding

The encoding and decoding procedures will be as follows.

## Encoding

For input message $m$, output $\mathbf{c}(m) = (x_1(m), x_2(m), \cdots, x_N(m))$

# Encoding and decoding

The encoding and decoding procedures will be as follows.

## Encoding

For input message $m$, output $\mathbf{c}(m) = (x_1(m), x_2(m), \cdots, x_N(m))$

## Decoding

Upon receiving sequence $\mathbf{y} = (y_1, y_2, \cdots, y_N)$, pick the sequence $\mathbf{c}(m)$ from $\{\mathbf{c}(1), \cdots, \mathbf{c}(2^{NR})\}$ such that $(\mathbf{c}(m), \mathbf{y})$ are jointly typical. That is $p_{X^N,Y^N}(\mathbf{c}(m), \mathbf{y}) \sim 2^{-nH(X,Y)}$. If no such $\mathbf{c}(m)$ exists or more than one such sequence exist, announce error. Otherwise output the decoded message as $m$

# Average performance

Without loss of generality, let us assume $M = 1$, decoding error occurs when:

# Average performance

Without loss of generality, let us assume $M = 1$, decoding error occurs when:

1. $P_1 = Pr(\mathbf{C}(1), \mathbf{Y}) \notin A_\epsilon^N(X, Y))$

# Average performance

Without loss of generality, let us assume $M = 1$, decoding error occurs when:

1. $P_1 = Pr(\mathbf{C}(1), \mathbf{Y}) \notin A_\epsilon^N(X, Y))$
2. $P_2 : \exists M' \neq 1$ and $(\mathbf{c}(M'), \mathbf{Y}) \in A_\epsilon^N(X, Y)$

Thus $P(error) = P(error|M = 1) \leq P_1 + P_2$

# Average performance

Without loss of generality, let us assume $M = 1$, decoding error occurs when:

1. $P_1 = Pr(\mathbf{C}(1), \mathbf{Y}) \notin A_\epsilon^N(X, Y))$
2. $P_2 : \exists M' \neq 1$ and $(\mathbf{c}(M'), \mathbf{Y}) \in A_\epsilon^N(X, Y)$

Thus $P(error) = P(error|M = 1) \leq P_1 + P_2$

1. Since $(\mathbf{C}(1), \mathbf{Y})$ is coming out of the joint source $X, Y$, $P_1 \to 0$ as $n \to \infty$

# Average performance

Without loss of generality, let us assume $M = 1$, decoding error occurs when:

1. $P_1 = Pr(\mathbf{C}(1), \mathbf{Y}) \notin A_\epsilon^N(X, Y))$
2. $P_2 : \exists M' \neq 1$ and $(\mathbf{c}(M'), \mathbf{Y}) \in A_\epsilon^N(X, Y)$

Thus $P(error) = P(error|M = 1) \leq P_1 + P_2$

1. Since $(\mathbf{C}(1), \mathbf{Y})$ is coming out of the joint source $X, Y$, $P_1 \to 0$ as $n \to \infty$
2. Note that $\mathbf{C}(M')$ and $\mathbf{Y}$ are independent and thus by Packing lemma,

$$P_2 \leq 2^{-N(I(X;Y)-R+3\epsilon)} \tag{1}$$

# Average performance

Without loss of generality, let us assume $M = 1$, decoding error occurs when:

1. $P_1 = Pr(\mathbf{C}(1), \mathbf{Y}) \notin A_\epsilon^N(X, Y))$
2. $P_2 : \exists M' \neq 1$ and $(\mathbf{c}(M'), \mathbf{Y}) \in A_\epsilon^N(X, Y)$

Thus $P(error) = P(error|M = 1) \leq P_1 + P_2$

1. Since $(\mathbf{C}(1), \mathbf{Y})$ is coming out of the joint source $X, Y$, $P_1 \to 0$ as $n \to \infty$
2. Note that $\mathbf{C}(M')$ and $\mathbf{Y}$ are independent and thus by Packing lemma,

$$P_2 \leq 2^{-N(I(X;Y)-R+3\epsilon)} \tag{1}$$

Since $\epsilon$ can be made arbitrarily small as $N$ increase, as long as $I(X; Y) > R$, we can make $P_2$ arbitrarily small also given a sufficiently large $N$

# A bit more caveat

- We show that the average error over all random codes can be made arbitrarily small

# A bit more caveat

- We show that the average error over all random codes can be made arbitrarily small
- But we really *want* is to find a code $\mathbf{c}^*(\cdot)$ and ensure that $Pr(error|\mathbf{c}^*, m) \to 0$ no matter what message $m$ is sent

# A bit more caveat

- We show that the average error over all random codes can be made arbitrarily small
- But we really *want* is to find a code $\mathbf{c}^*(\cdot)$ and ensure that $Pr(error|\mathbf{c}^*, m) \to 0$ no matter what message $m$ is sent
- Let say for a finite $N$, the average error is $\delta$. Then, we should be able to find a code $\mathbf{c}^*$ such that it has average error at least equal to $\delta$

# A bit more caveat

- We show that the average error over all random codes can be made arbitrarily small
- But we really *want* is to find a code $\mathbf{c}^*(\cdot)$ and ensure that $Pr(error|\mathbf{c}^*, m) \to 0$ no matter what message $m$ is sent
- Let say for a finite $N$, the average error is $\delta$. Then, we should be able to find a code $\mathbf{c}^*$ such that it has average error at least equal to $\delta$
- Without loss of generality and for simplicity, assume that all messages are equally likely $Pr(error|\mathbf{c}^*) = \frac{1}{2^{NR}} \sum_m Pr(error|\mathbf{c}^*, m) \leq \delta$

# A bit more caveat

- We show that the average error over all random codes can be made arbitrarily small
- But we really *want* is to find a code $\mathbf{c}^*(\cdot)$ and ensure that $Pr(error|\mathbf{c}^*, m) \to 0$ no matter what message $m$ is sent
- Let say for a finite $N$, the average error is $\delta$. Then, we should be able to find a code $\mathbf{c}^*$ such that it has average error at least equal to $\delta$
- Without loss of generality and for simplicity, assume that all messages are equally likely $Pr(error|\mathbf{c}^*) = \frac{1}{2^{NR}} \sum_m Pr(error|\mathbf{c}^*, m) \leq \delta$
- If we discard the worse half of the codewords, for any remaining message $m$, we have $Pr(error|\mathbf{c}^*, m) \leq 2Pr(error|\mathbf{c}^*) \leq 2\delta \to 0$ as $N \to \infty$

# A bit more caveat

- We show that the average error over all random codes can be made arbitrarily small
- But we really *want* is to find a code $\mathbf{c}^*(\cdot)$ and ensure that $Pr(error|\mathbf{c}^*, m) \to 0$ no matter what message $m$ is sent
- Let say for a finite $N$, the average error is $\delta$. Then, we should be able to find a code $\mathbf{c}^*$ such that it has average error at least equal to $\delta$
- Without loss of generality and for simplicity, assume that all messages are equally likely $Pr(error|\mathbf{c}^*) = \frac{1}{2^{NR}} \sum_m Pr(error|\mathbf{c}^*, m) \leq \delta$
- If we discard the worse half of the codewords, for any remaining message $m$, we have $Pr(error|\mathbf{c}^*, m) \leq 2Pr(error|\mathbf{c}^*) \leq 2\delta \to 0$ as $N \to \infty$
- Even though the rate reduces from $N$ to $R - \frac{1}{N}$ (number of messages from $2^{NR} \to 2^{NR-1}$). But we can still make the final rate arbitrarily close to the capacity as $N \to \infty$