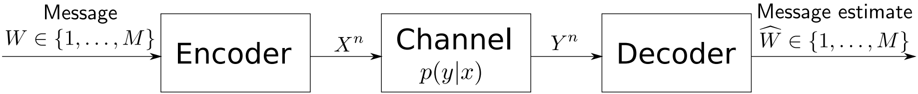
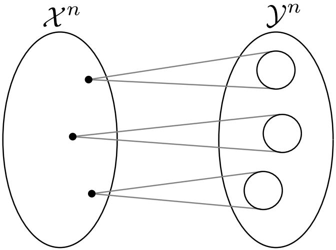
EEE 551 Information Theory (Spring 2022)

Chapter 7: Channel Capacity



Discrete memoryless channel (DMC): $$p(y^n|x^n) = \prod_{i=1}^{n} p(y_i|x_i)$$



How many balls can we pack into $\mathcal{Y}^n$?

Operational Definition of Channel Capacity

A **discrete channel**, denoted $(\mathcal{X}, p(y|x), \mathcal{Y})$, consists of an input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$, and conditional probability $p(y|x)$

An (M, n) code consists of:

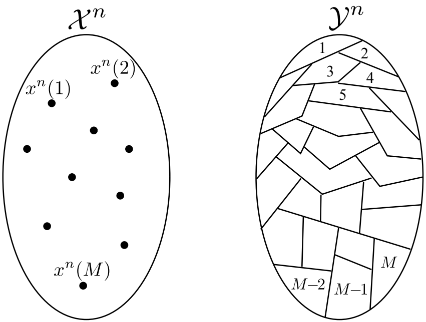
- Message set $\{1, 2, \ldots, M-1, M\}$
- Encoding function

$$x^n : \{1, 2, \ldots, M\} \to \mathcal{X}^n$$

  $x^n(1), x^n(2), \ldots, x^n(M)$ are **codewords** (which form the **codebook**)

- Decoding function

$$g : \mathcal{Y}^n \to \{1, 2, \ldots, M\}$$



- For a given message $m \in \{1, 2, \dots, M\}$, its **probability of error** is

$$\begin{aligned}\lambda_m &= \Pr\{\text{error}|m \text{ is transmitted}\} \\ &= \Pr\{g(Y^n) \neq m | X^n = x^n(m)\} \\ &= \sum_{y^n : g(y^n) \neq m} p(y^n | x^n(m))\end{aligned}$$

- The **maximal probability of error** is $\lambda^{(n)} = \max_{m \in \{1,2,\dots,M\}} \lambda_m$
- The **average probability of error** is $P_e^{(n)} = \frac{1}{M} \sum_{m=1}^{M} \lambda_m$

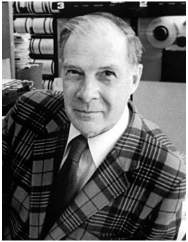
  (note that $P_e^{(n)} \leq \lambda^{(n)}$)
- The **rate** R of an (M, n) code is $R = \frac{\log M}{n}$ bits per channel use

  i.e. $M = 2^{nR}$
- A rate R is **achievable** if there exists a sequence of rate-R codes such that $\lambda^{(n)} \to 0$ as $n \to \infty$
- The **channel capacity** C is the supremum of all achievable rates

Example code: Hamming code

$(16, 7)$ code for a binary-input channel.

$n = 7$, $nR = 4$, $R = \dfrac{4}{7}$

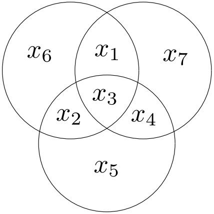


Richard Hamming
(1915–1998)

Encoding process

1. Select message $W \in \{1, \ldots, 16\}$
2. Write W as 4 bits $x_1x_2x_3x_4$
3. Codeword is $\underbrace{x_1x_2x_3x_4}_{\text{message}}\ \underbrace{x_5x_6x_7}_{\text{parity check bits}}$

   where x_5, x_6, x_7 selected so that the number of 1s in each circle is even:



Example: If $W = 13$, then $x_1x_2x_3x_4 = 1101$, and codeword is $x^n(13) = 1101000$

Full codebook:

$$
\begin{aligned}
x^n(1) &= 0001101\\
x^n(2) &= 0010111\\
x^n(3) &= 0011010\\
x^n(4) &= 0100110\\
x^n(5) &= 0101011\\
x^n(6) &= 0110001\\
x^n(7) &= 0111100\\
x^n(8) &= 1000011\\
x^n(9) &= 1001110\\
x^n(10) &= 1010100\\
x^n(11) &= 1011001\\
x^n(12) &= 1100101\\
x^n(13) &= 1101000\\
x^n(14) &= 1110010\\
x^n(15) &= 1111111\\
x^n(16) &= 0000000
\end{aligned}
$$

Decoding process for a binary symmetric channel (BSC)

1. Put all 7 bits into Venn diagram
2. Identify which circles have a parity error
3. If no parity errors, take $x_1x_2x_3x_4$ as given
4. Otherwise, identify bit in all error circles and flip it
5. Take $x_1x_2x_3x_4$ from adjusted diagram

Example: Codeword 1101000:

- 0 bit flips: $Y^n = 1101000$ decoded as 1101
- 1 bit flips: $Y^n = 1001000$ decoded as 1101
- 2 bit flips: $Y^n = 1000000$ decoded as 0000 error!
- 3 bit flips: $Y^n = 0000000$ decoded as 0000 error!

In general, error occurs if 2 or more bit flips occur

$$\lambda^{(n)} = \sum_{i=2}^{7} \Pr\{i \text{ bit flips occur}\} = \sum_{i=2}^{7} \binom{7}{i} p^i (1-p)^{7-i}$$

If $p = 0.01$, $\lambda^{(n)} = 0.00203$

The Information Channel Capacity

Given a discrete channel $(\mathcal{X}, p(y|x), \mathcal{Y})$, the **information channel capacity** is given by

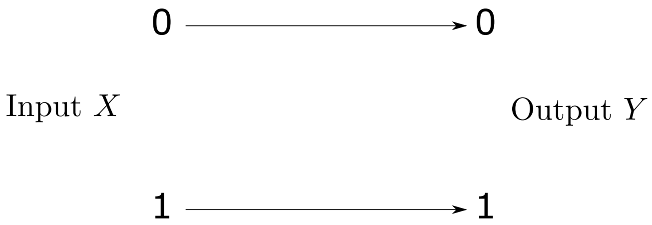
$$C^{(I)} = \max_{p(x)} I(X;Y)$$

where $(X,Y) \sim p(x)\,p(y|x)$, and the max is taken over all possible input distributions on alphabet $\mathcal{X}$

Theorem (Shannon's main theorem)

$C = C^{(I)}$

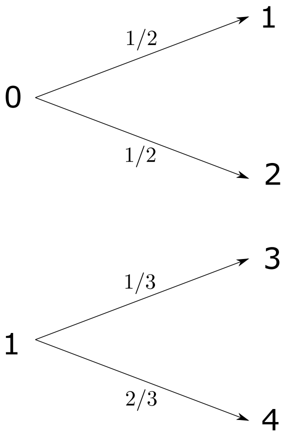
Example 1: Noiseless Binary Channel



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

achieved by $p(x) = (1/2, 1/2)$

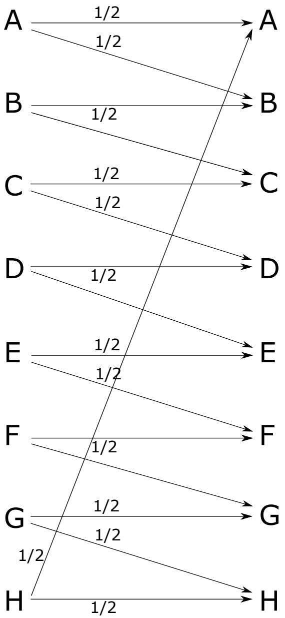
Example 2: Nonoverlapping Outputs



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

achieved by $p(x) = (1/2, 1/2)$

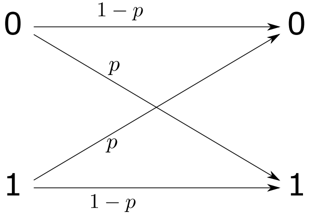
Example 3: Noisy Typewriter



$$
\begin{aligned}
C^{(I)} &= \max_{p(x)} I(X;Y) \\
&= \max_{p(x)} H(Y) - H(Y|X) \\
&= \max_{p(x)} H(Y) - 1 \\
&= \log 8 - 1 \\
&= 2 \text{ bits}
\end{aligned}
$$

achieved by uniform distribution on $\{\mathsf{A}, \mathsf{C}, \mathsf{E}, \mathsf{G}\}$, or by uniform distribution on $\{\mathsf{A}, \ldots, \mathsf{H}\}$,

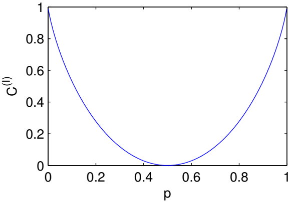
Example 4: Binary Symmetric Channel (BSC)



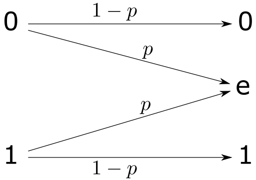
$$
\begin{aligned}
I(X;Y) &= H(Y) - H(Y|X) \\
&= H(Y) - H(p) \\
&\leq 1 - H(p)
\end{aligned}
$$

equality iff Y is uniform, which occurs if X is uniform

$$\implies \quad C^{(I)} = 1 - H(p) \text{ bits}$$



Example 5: Binary Erasure Channel (BEC)



Let $E = \begin{cases} 1, & \text{if } Y = \mathsf{e} \\ 0, & \text{if } Y \neq \mathsf{e}. \end{cases}$

$$\begin{aligned}
I(X;Y) &= I(X;Y,E) \\
&= I(X;E) + I(X;Y|E) \\
&= I(X;Y|E) \\
&= H(Y|E) - H(Y|E,X) \\
&= H(Y|E) \\
&= \Pr(E=0)H(Y|E=0) + \Pr(E=1)H(Y|E=1) \\
&= \Pr(E=0)H(X|E=0) \\
&= (1-p)H(X) \\
&\leq 1-p
\end{aligned}$$

Equality if X is uniform, so $C^{(I)} = 1 - p$

Example 6: Weakly Symmetric Channels

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

$$p(y|x) = \underbrace{\left[\begin{array}{ccc} 1/3 & 1/6 & 1/2 \\ 1/3 & 1/2 & 1/6 \end{array}\right]}_{\mathcal{Y}} \Big\} \mathcal{X}$$

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - \sum_x p(x) H(Y|X=x) \\ &= H(Y) - H(1/3, 1/2, 1/6) \\ &\leq \log 3 - H(1/3, 1/2, 1/6) \end{aligned}$$

If X is uniform, then

$$p(y) = \frac{1}{2}[1/3 \;\; 1/6 \;\; 1/2] + \frac{1}{2}[1/3 \;\; 1/2 \;\; 1/6] = [1/3 \;\; 1/3 \;\; 1/3]$$

Thus uniform X achieves uniform $Y \implies C^{(I)} = \log 3 - H(1/3, 1/2, 1/6)$

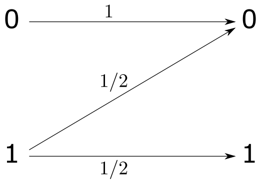
More generally:

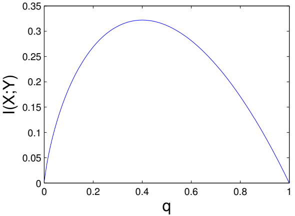
- A channel is **weakly symmetric** if rows are permutations of each other, and columns sums $\sum_x p(y|x)$ are equal
- For weakly symmetric channels,

$$C^{(I)} = \log|\mathcal{Y}| - H(\text{row of transition matrix})$$

  achieved by uniform input distribution
- BSC, noisy typewriter are special cases

Example 7: Z-Channel





Let $X \sim \text{Bern}(q)$

$$I(X;Y) = H(Y) - H(Y|X) = H(q/2) - q$$

Let $r = q/2$, so

$$C^{(I)} = \max_r H(r) - 2r = \max_r -r\log r - (1-r)\log(1-r) - 2r$$

Find optimal r:

$$0 = \frac{d}{dr} I(X;Y) = \log\left(\frac{1-r}{r}\right) - 2 \Longrightarrow r = 1/5, \quad q = 2/5$$

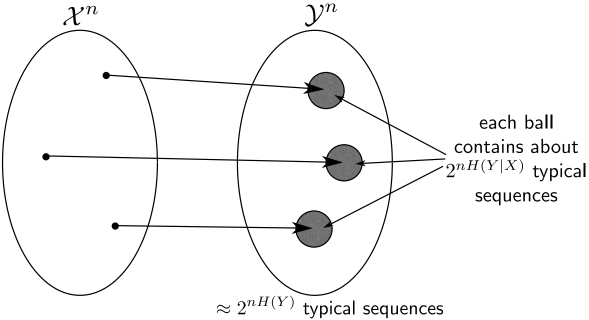
Therefore $C^{(I)} = H(1/5) - 2/5 \approx 0.322$ bits

Properties of the Information Channel Capacity

- $C^{(I)} \geq 0$, since $I(X;Y) \geq 0$
- $C^{(I)} \leq \log|\mathcal{X}|$, since $I(X;Y) \leq H(X) \leq \log|\mathcal{X}|$
- $C^{(I)} \leq \log|\mathcal{Y}|$
- Since $I(X;Y)$ is a concave function of $p(x)$, standard convex optimization techniques can be used to calculate $C^{(I)}$ numerically

Channel Capacity Intuition

For large n, all channels look like the noisy typewriter:

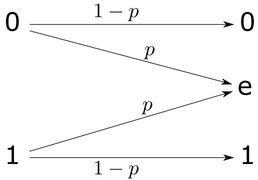


maximum number of balls that can be packed in $\mathcal{Y}^n$:

$$\frac{2^{nH(Y)}}{2^{nH(Y|X)}} = 2^{n(H(Y)-H(Y|X))} = 2^{nI(X;Y)}$$

Since we can choose $p(x)$, maximum rate is: $\max_{p(x)} I(X;Y)$

Focus on the binary erasure channel



Theorem (BEC capacity)

For a BEC(p), $C = 1 - p$.

- Easy to see that $C \leq 1 - p$
- Easy to achieve rate $R = 1 - p$ with feedback
- How to achieve rate $R = 1 - p$ without feedback?

Polar Codes

Erdal Arıkan

There are two kinds of channels that make it easy to achieve capacity:

- **Perfect channels**
- **Useless channels**

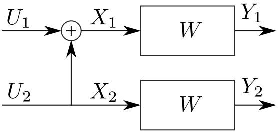
Polar codes work by **polarizing** the channel — covert the channel into a mixture of perfect or useless channels

- Polar codes can achieve the capacity of any binary-input channel where the capacity-achieving input distribution is Bern$(1/2)$ (e.g. BEC, BSC)
- Low complexity encoders and decoders
- Introduced in 2008, implemented in 5G standard in 2016

Basic Polar Transform

- Let W be the channel
- Let $I(W) = I(X;Y)$ where $X \sim \text{Bern}(1/2)$
- U_1, U_2 represent the message bits — each $\text{Bern}(1/2)$
- Transmitted bits are formed by

$$X_1 = U_1 \oplus U_2, \qquad X_2 = U_2, \qquad \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \end{bmatrix}$$



- Since X_1, X_2 are independent:

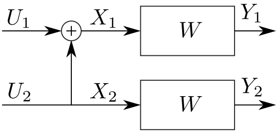
$$\begin{aligned} 2I(W) &= I(X_1, X_2\,;\, Y_1, Y_2) \\ &= I(U_1, U_2\,;\, Y_1, Y_2) \\ &= I(U_1\,;\, Y_1, Y_2) + I(U_2\,;\, Y_1, Y_2 | U_1) \\ &= I(U_1\,;\, Y_1, Y_2) + I(U_2\,;\, Y_1, Y_2, U_1) \\ &= I(W^-) + I(W^+) \end{aligned}$$

- W^- is the channel from $U_1 \to Y_1, Y_2$
- W^+ is the channel from $U_2 \to Y_1, Y_2, U_1$

Basic Polar Transform on the BEC

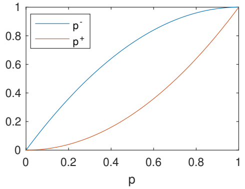
- W^- is the channel from $U_1 \to Y_1, Y_2$
- U_1 can be decoded only if neither Y_1, Y_2 are erased
- W^- is equivalent to a BEC(p^-) where

$$p^- = 1 - (1-p)^2 = 2p - p^2$$



- W^+ is the channel from $U_2 \to Y_1, Y_2, U_1$
- U_2 can be decoded if either of Y_1, Y_2 are un-erased
- W^+ is equivalent to a BEC(p^+) where

$$p^+ = p^2$$

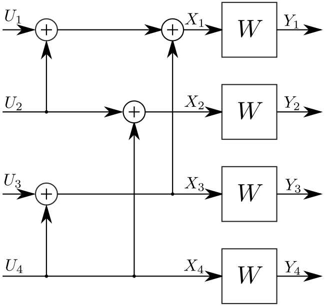


- $p^- > p > p^+$
- W^- is a worse channel than W, and W^+ is better
  Polarization!

2nd generation polar transform

- Given two copies of W, we fabricated W^- and W^+
- We can duplicate W^- and W^+, and fabricate

$$W^{--} : U_1 \to Y_1, Y_2, Y_3, Y_4$$
$$W^{-+} : U_2 \to Y_1, Y_2, Y_3, Y_4, U_1$$
$$W^{+-} : U_3 \to Y_1, Y_2, Y_3, Y_4, U_1, U_2$$
$$W^{++} : U_4 \to Y_1, Y_2, Y_3, Y_4, U_1, U_2, U_3$$



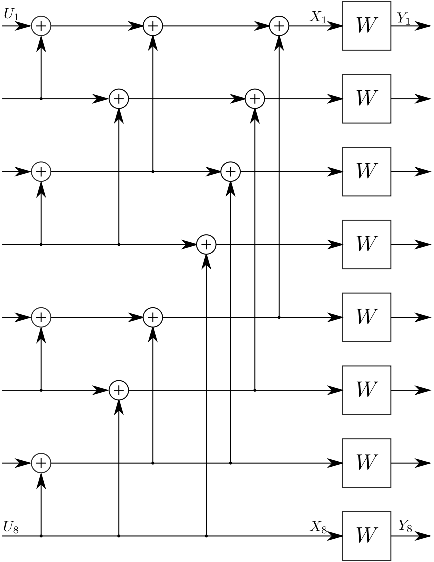
- W^{--} is equivalent to a BEC(p^{--}) where

$$p^{--} = 2p^{-} - (p^{-})^2$$

- Similarly

$$p^{-+} = (p^{-})^2$$
$$p^{+-} = 2p^{+} - (p^{+})^2$$
$$p^{++} = (p^{+})^2$$

3rd generation polar transform



- Can continue t generations, to create a $n = 2^t$ length code
- Encoding can be done in $nt = n \log n$ time (same for decoding, but less obvious)

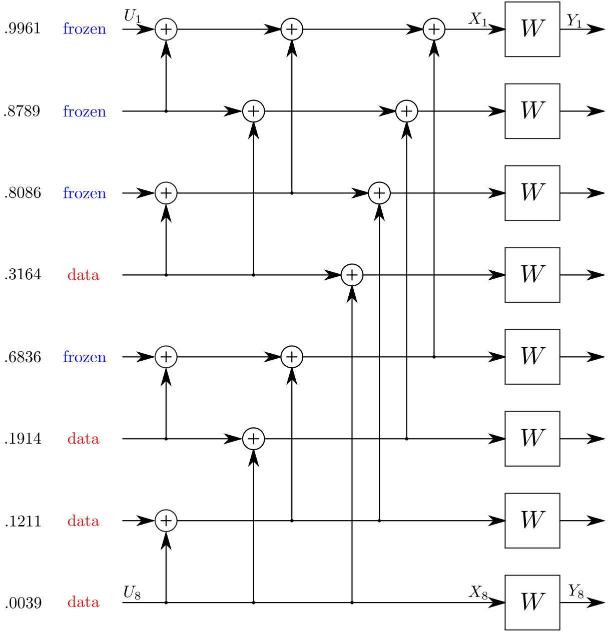
Polar coding

How we actually use the polar transform to code:

- Fix target rate $R < 1 - p$
- Apply the polar transform for t generations to synthesize the $n = 2^t$ channels $W^{+\cdots+}, \dots, W^{-\cdots-}$
- For $s^t = (s_1, \dots, s_t) \in \{+,-\}^t$, W^{s^t} is equivalent to a BEC(p^{s^t})
- Set the inputs of the best nR channels to uncoded data
- Freeze the inputs of the remaining channels to 0
- At the receiver, successively decode $U_1, \dots, U_n$. These must be decoded in order, so when decoding each channel, the previous channel inputs are available.
  For a frozen input i, we can assume $U_i = 0$
- The error probability is upper bounded by

$$\sum_{nR \text{ best channels } s^t \in \{+,-\}^t} p^{s^t}$$

Polar coding for 3rd generation transform



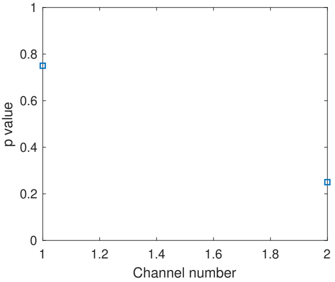
Polarization Phenomenon

- For $s^t = (s_1, \ldots, s_t) \in \{+,-\}^t$,

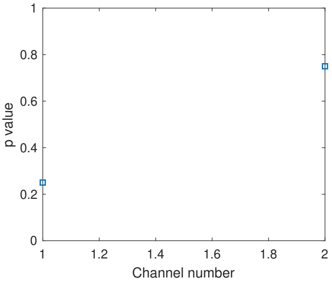
$$p^{s^t} = \begin{cases} 2p^{s^{t-1}} - (p^{s^{t-1}})^2, & s_t = - \\ (p^{s^{t-1}})^2, & s_t = + \end{cases}$$

- Note that $p^{s^{t-1}} = \frac{1}{2}\left(p^{s^{t-1}-} + p^{s^{t-1}+}\right)$
(i.e., conservation of capacity)
- $p = 0.5$:

Unsorted:



Sorted:



$t = 1$

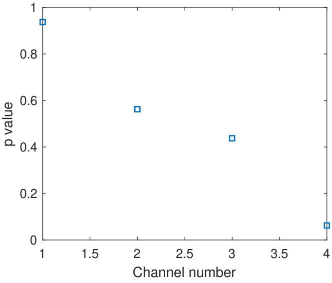
Polarization Phenomenon

- For $s^t = (s_1, \ldots, s_t) \in \{+, -\}^t$,

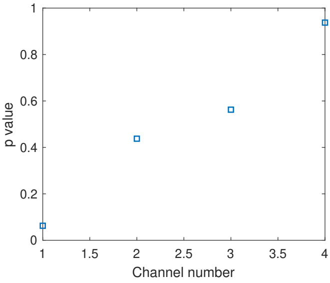
$$p^{s^t} = \begin{cases} 2p^{s^{t-1}} - (p^{s^{t-1}})^2, & s_t = - \\ (p^{s^{t-1}})^2, & s_t = + \end{cases}$$

- Note that $p^{s^{t-1}} = \frac{1}{2}\left(p^{s^{t-1}-} + p^{s^{t-1}+}\right)$ (i.e., conservation of capacity)
- $p = 0.5$:

Unsorted:



Sorted:



$t = 2$

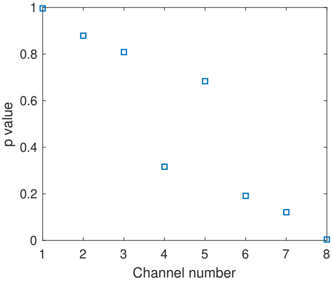
Polarization Phenomenon

- For $s^t = (s_1, \dots, s_t) \in \{+,-\}^t$,

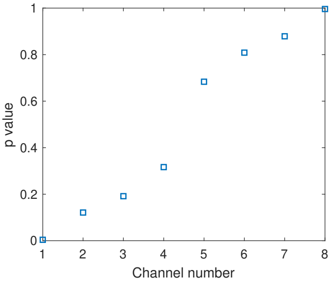
$$p^{s^t} = \begin{cases} 2p^{s^{t-1}} - (p^{s^{t-1}})^2, & s_t = - \\ (p^{s^{t-1}})^2, & s_t = + \end{cases}$$

- Note that $p^{s^{t-1}} = \frac{1}{2}\left(p^{s^{t-1}-} + p^{s^{t-1}+}\right)$ (i.e., conservation of capacity)
- $p = 0.5$:

Unsorted:



Sorted:



$t = 3$

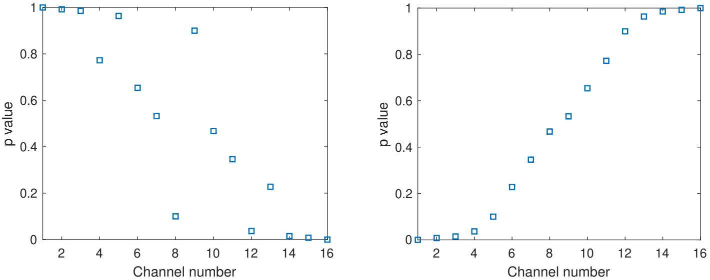
Polarization Phenomenon

- For $s^t = (s_1, \ldots, s_t) \in \{+, -\}^t$,

$$p^{s^t} = \begin{cases} 2p^{s^{t-1}} - (p^{s^{t-1}})^2, & s_t = - \\ (p^{s^{t-1}})^2, & s_t = + \end{cases}$$

- Note that $p^{s^{t-1}} = \frac{1}{2}\left(p^{s^{t-1}-} + p^{s^{t-1}+}\right)$ (i.e., conservation of capacity)
- $p = 0.5$:

Unsorted: Sorted:



$t = 4$

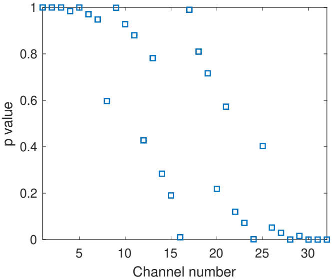
Polarization Phenomenon

- For $s^t = (s_1, \ldots, s_t) \in \{+, -\}^t$,

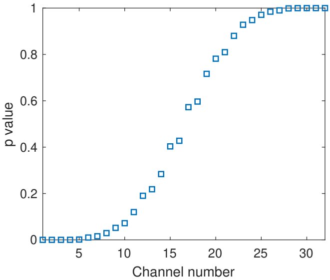
$$p^{s^t} = \begin{cases} 2p^{s^{t-1}} - (p^{s^{t-1}})^2, & s_t = - \\ (p^{s^{t-1}})^2, & s_t = + \end{cases}$$

- Note that $p^{s^{t-1}} = \frac{1}{2}\left(p^{s^{t-1}-} + p^{s^{t-1}+}\right)$ (i.e., conservation of capacity)
- $p = 0.5$:

Unsorted:



Sorted:



$t = 5$

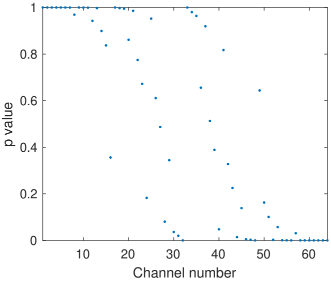
Polarization Phenomenon

- For $s^t = (s_1, \ldots, s_t) \in \{+,-\}^t$,

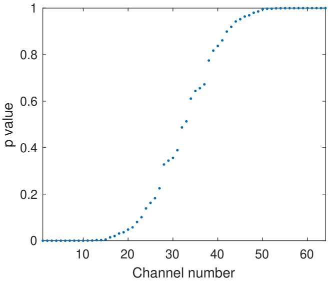
$$p^{s^t} = \begin{cases} 2p^{s^{t-1}} - (p^{s^{t-1}})^2, & s_t = - \\ (p^{s^{t-1}})^2, & s_t = + \end{cases}$$

- Note that $p^{s^{t-1}} = \frac{1}{2}\left(p^{s^{t-1}-} + p^{s^{t-1}+}\right)$
(i.e., conservation of capacity)
- $p = 0.5$:

Unsorted:



Sorted:



$t = 6$

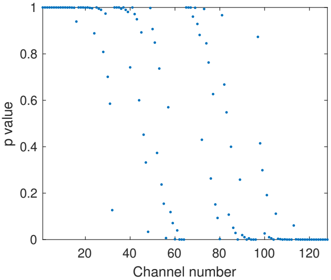
Polarization Phenomenon

- For $s^t = (s_1, \ldots, s_t) \in \{+, -\}^t$,

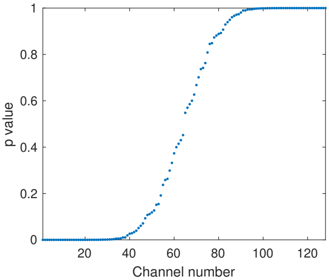
$$p^{s^t} = \begin{cases} 2p^{s^{t-1}} - (p^{s^{t-1}})^2, & s_t = - \\ (p^{s^{t-1}})^2, & s_t = + \end{cases}$$

- Note that $p^{s^{t-1}} = \frac{1}{2}\left(p^{s^{t-1}-} + p^{s^{t-1}+}\right)$ (i.e., conservation of capacity)
- $p = 0.5$:

Unsorted:



Sorted:



$t = 7$

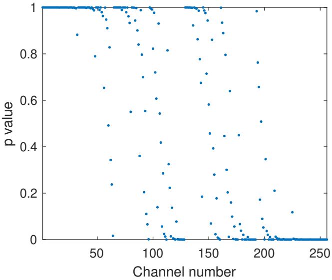
Polarization Phenomenon

- For $s^t = (s_1, \ldots, s_t) \in \{+, -\}^t$,

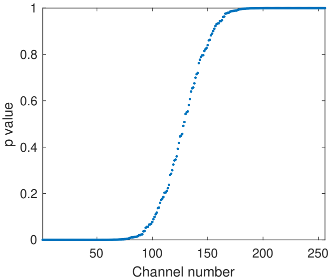
$$p^{s^t} = \begin{cases} 2p^{s^{t-1}} - (p^{s^{t-1}})^2, & s_t = - \\ (p^{s^{t-1}})^2, & s_t = + \end{cases}$$

- Note that $p^{s^{t-1}} = \frac{1}{2}\left(p^{s^{t-1}-} + p^{s^{t-1}+}\right)$
  (i.e., conservation of capacity)
- $p = 0.5$:

Unsorted:



Sorted:



$t = 8$

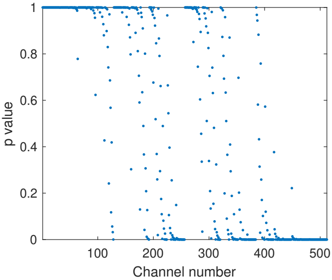
Polarization Phenomenon

- For $s^t = (s_1, \ldots, s_t) \in \{+,-\}^t$,

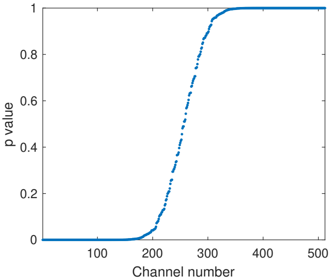
$$p^{s^t} = \begin{cases} 2p^{s^{t-1}} - (p^{s^{t-1}})^2, & s_t = - \\ (p^{s^{t-1}})^2, & s_t = + \end{cases}$$

- Note that $p^{s^{t-1}} = \frac{1}{2}\left(p^{s^{t-1}-} + p^{s^{t-1}+}\right)$ (i.e., conservation of capacity)
- $p = 0.5$:

Unsorted:



Sorted:



$t = 9$

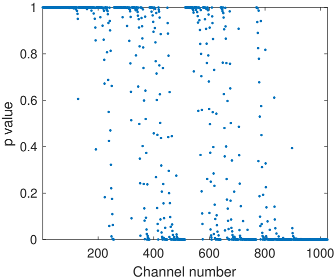
Polarization Phenomenon

- For $s^t = (s_1, \ldots, s_t) \in \{+, -\}^t$,

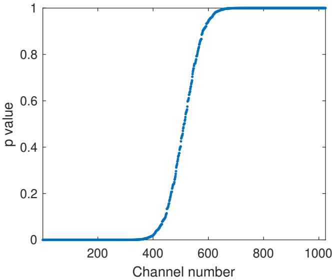
$$p^{s^t} = \begin{cases} 2p^{s^{t-1}} - (p^{s^{t-1}})^2, & s_t = - \\ (p^{s^{t-1}})^2, & s_t = + \end{cases}$$

- Note that $p^{s^{t-1}} = \frac{1}{2}\left(p^{s^{t-1}-} + p^{s^{t-1}+}\right)$ (i.e., conservation of capacity)
- $p = 0.5$:

Unsorted:



Sorted:



$t = 10$

BEC Polarization Theorem

Theorem

Let $\mu_t(\delta)$ be the fraction of δ**-mediocre channels** *after t generations of the polar transform:*

$$\mu_t(\delta) = \frac{1}{2^t} \sum_{s^t \in \{+,-\}^t} \mathbf{1}[p^{s^t} \in (\delta, 1-\delta)].$$

For any $\delta > 0$, $\lim_{t\to\infty} \mu_t(\delta) = 0$.

Implies that roughly $(1-p)2^t$ channels have $p^{s^t} < \delta$ and $p\,2^t$ channels have $p^{s^t} > 1-\delta$

Proof:

- For erasure probability q, define its **ugliness**, $\text{ugly}(q) = \sqrt{q(1-q)}$
- $\mathbf{1}[q \in (\delta, 1-\delta)] \leq \dfrac{\text{ugly}(q)}{\sqrt{\delta(1-\delta)}}$
- Thus it is enough to prove that

$$\lim_{t\to\infty} \frac{1}{2^t} \sum_{s^t \in \{+,-\}^t} \text{ugly}(p^{s^t}) = 0.$$

Recall $\mathsf{ugly}(q) = \sqrt{q(1-q)}$

- For the two channels descended from q:

$$\begin{aligned}
\mathsf{ugly}(q^{+}) &= \mathsf{ugly}(q^{2}) & \mathsf{ugly}(q^{-}) &= \mathsf{ugly}(2q - q^{2}) \\
&= \sqrt{(q^2)(1-q^2)} & &= \sqrt{(2q-q^2)(1-2q+q^2)} \\
&= \sqrt{q^2(1-q)(1+q)} & &= \sqrt{q(2-q)(1-q)^2} \\
&= \mathsf{ugly}(q)\sqrt{q(1+q)} & &= \mathsf{ugly}(q)\sqrt{(2-q)(1-q)}
\end{aligned}$$

- Average of the ugliness of two descendents:

$$\begin{aligned}
\frac{1}{2}\mathsf{ugly}(q^{+}) + \frac{1}{2}\mathsf{ugly}(q^{-}) &= \mathsf{ugly}(q)\frac{1}{2}\left(\sqrt{q(1+q)} + \sqrt{(2-q)(1-q)}\right) \\
&\le \mathsf{ugly}(q)\sqrt{\frac{3}{4}}
\end{aligned}$$

- $\displaystyle \frac{1}{2^t}\sum_{s^t \in \{+,-\}^t} \mathsf{ugly}(p^{s^t}) \le \mathsf{ugly}(p)\left(\frac{3}{4}\right)^{t/2} \to 0$

Channel Coding Converse

- Consider any channel $p(y|x)$
- Assume there exists a sequence of $(2^{nR}, n)$ codes such that avg. probability of error $P_e^{(n)} \to 0$
- We want to show $R \leq C^{(I)}$

Notes:

- We use avg. probability of error, since it's a weaker condition than max. probability of error (i.e $\lambda^{(n)} \to 0$ implies $P_e^{(n)} \to 0$)
- Equivalent to: if $R > C^{(I)}$, then for any sequence of $(2^{nR}, n)$ codes, $P_e^{(n)}$ is bounded away from 0

Proof:

- $W \to X^n \to Y^n \to \widehat{W}$ is a Markov chain
- $\Pr\{W = m\} = 2^{-nR}$, and $P_e^{(n)} = \Pr\{W \neq \widehat{W}\}$
- By Fano's inequality,

$$\begin{aligned} H(W|Y^n) &\leq 1 + P_e^{(n)} \log(2^{nR}) \\ &= 1 + P_e^{(n)} nR \\ &= n\left(\frac{1}{n} + P_e^{(n)} R\right) \\ &= n\epsilon_n \end{aligned}$$

where $\epsilon_n \to 0$ as $n \to \infty$

$$
\begin{aligned}
nR &= H(W) \\
&= I(W;Y^n) + H(W|Y^n) \\
&\le I(W;Y^n) + n\epsilon_n \\
&\le I(X^n;Y^n) + n\epsilon_n \\
&= H(Y^n) - H(Y^n|X^n) + n\epsilon_n \\
&= H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1,\dots,Y_{i-1},X^n) + n\epsilon_n \\
&= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) + n\epsilon_n \\
&\le \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) + n\epsilon_n \\
&= \sum_{i=1}^{n} I(X_i;Y_i) + n\epsilon_n \\
&\le nC^{(I)} + n\epsilon_n
\end{aligned}
$$

Therefore $R \le C^{(I)} + \epsilon_n$. Taking the limit yields $R \le C^{(I)}$

Toward Achievability: Jointly Typical Sequences

Given a joint distribution $p(x, y)$, the **jointly typical set** is

$$A_\epsilon^{(n)} = \Big\{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \left|-\tfrac{1}{n}\log p(x^n) - H(X)\right| \le \epsilon,$$
$$\left|-\tfrac{1}{n}\log p(y^n) - H(Y)\right| \le \epsilon,$$
$$\left|-\tfrac{1}{n}\log p(x^n, y^n) - H(X, Y)\right| \le \epsilon\Big\}$$

Properties (joint AEP)

- If $(X^n, Y^n) \overset{\text{iid}}{\sim} p(x, y)$, then $\Pr\left\{(X^n, Y^n) \in A_\epsilon^{(n)}\right\} \to 1$ as $n \to \infty$
- $|A_\epsilon^{(n)}| \le 2^{n(H(X,Y)+\epsilon)}$,
  $|A_\epsilon^{(n)}| \ge (1-\epsilon)2^{n(H(X,Y)-\epsilon)}$ for sufficiently large n
- If $(\tilde{X}^n, \tilde{Y}^n) \overset{\text{iid}}{\sim} p(x)p(y)$, then

$$\Pr\left\{(\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\right\} \le 2^{-n(I(X;Y)-3\epsilon)}$$
$$\Pr\left\{(\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\right\} \ge (1-\epsilon)2^{-n(I(X;Y)+3\epsilon)} \text{ for sufficiently large } n$$

Properties 1 and 2 follow from the same arguments as standard AEP

Proof of Property 3

Let $(\tilde{X}^n, \tilde{Y}^n) \overset{\text{iid}}{\sim} p(x)p(y)$

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

Lower bound follows similarly

Achievability Proof of Channel Coding Theorem

Key idea: **Random coding**

1. Construct probability distribution on code
2. Calculate average probability of error of randomly chosen code
3. Therefore, there is at least one code with probability of error below average

> "All codes are good, except those we can think of." – Gérard Battail
>
> "I thought of one." – Erdal Arıkan (not actually a quote)

Proof Setup

- We want to prove that $C \geq C^{(I)}$. It's enough to prove that all $R < C^{(I)}$ are achievable
- Let $p(x)$ be a distribution achieving the maximum in $C^{(I)}$, so if $(X, Y) \sim p(x)p(y|x)$, then $C^{(I)} = I(X; Y)$
- Fix any $R < I(X; Y)$. We prove R is achievable. This requires proving that there exists codes at rate R with arbitrarily small probability of error for sufficiently large blocklength n
- Fix n and $\epsilon > 0$

Random codebook generation

For $m \in \{1, 2, \ldots, 2^{nR}\}$, generate codeword $X^n(m) \overset{\text{iid}}{\sim} p(x)$

$$\text{Codebook } \mathcal{C} = \begin{bmatrix} X_1(1) & X_2(1) & \cdots & X_n(1) \\ X_1(2) & X_2(2) & \cdots & X_n(2) \\ \vdots & \vdots & \ddots & \vdots \\ X_1(2^{nR}) & X_2(2^{nR}) & \cdots & X_n(2^{nR}) \end{bmatrix}$$

$$\Pr\{\mathcal{C} = c\} = \prod_{m=1}^{2^{nR}} \prod_{i=1}^{n} p_X\big(x_i(m)\big)$$

Encoding Process

- Message W selected at random with $\Pr\{W = m\} = 2^{-nR}$ for all m
- Encoder transmits $X^n(W)$

Decoding Process

- Decoder receives Y^n
- Decoder declares $\widehat{W}$ to be the smallest m such that

$$(X^n(m), Y^n) \in A_\epsilon^{(n)}$$

- If there is no such m, decoder declares an error

Analysis of Probability of Error

Note there are three independent sources of randomness:

- Generation of $\mathcal{C}$
- Selection of the message W
- Channel behavior

Let $\lambda_m(c)$ be the error probability for the mth codeword with code c (includes randomness only from the channel), i.e.

$$\lambda_m(c) = \Pr\left\{\widehat{W} \neq m | W = m, \mathcal{C} = c\right\}$$

The average probability of error for code c is

$$P_e^{(n)}(c) = \Pr\{\widehat{W} \neq W | \mathcal{C} = c\} = \frac{1}{2^{nR}} \sum_{m=1}^{2^{nR}} \lambda_m(c)$$

The probability of error averaged over all codebooks is

$$\begin{aligned}\bar{P}_e^{(n)} &= \sum_c p(c) P_e^{(n)}(c) \\ &= \sum_c p(c) \sum_m 2^{-nR} \lambda_m(c) \\ &= \sum_m 2^{-nR} \underbrace{\sum_c p(c) \Pr\left\{\widehat{W} \neq m | W = m, \mathcal{C} = c\right\}}_{= B_m = \Pr(\widehat{W} \neq m | W = m)}\end{aligned}$$

Define the following events:

$$\mathcal{E}_i = \left\{ (X^n(i), Y^n) \in A_\epsilon^{(n)} \right\} \text{ for } i = 1, 2, \ldots, 2^{nR}$$

If $W = m$, an error may occur only if

- $(X^n(m), Y^n) \notin A_\epsilon^{(n)}$, i.e. $\mathcal{E}_m^c$
- or, there exists $m' \neq m$ such that $(X^n(m'), Y^n) \in A_\epsilon^{(n)}$, i.e. $\displaystyle\bigcup_{m' \neq m} \mathcal{E}_{m'}$

Thus

$$\begin{aligned} B_m &\leq \Pr\left\{ \mathcal{E}_m^c \cup \bigcup_{m' \neq m} \mathcal{E}_{m'} \middle| W = m \right\} \\ &\leq \Pr\{\mathcal{E}_m^c | W = m\} + \sum_{m' \neq m} \Pr\{\mathcal{E}_{m'} | W = m\} \end{aligned}$$

- If $W = m$, $(X^n(m), Y^n) \overset{\text{iid}}{\sim} p(x, y)$, so by joint AEP

$$\Pr\{\mathcal{E}_m^c | W = m\} \leq \epsilon \text{ for } n \text{ sufficiently large}$$

- If $W = m$ and $m' \neq m$, $(X^n(m'), Y^n) \overset{\text{iid}}{\sim} p(x)p(y)$, so

$$\Pr\{\mathcal{E}_{m'} | W = m\} \leq 2^{-n(I(X;Y) - 3\epsilon)}$$

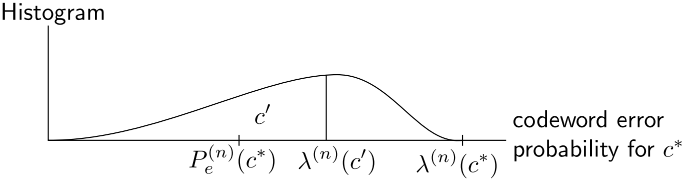
- For n sufficiently large

$$\begin{aligned} B_m &\le \epsilon + \sum_{m' \ne m} 2^{-n(I(X;Y)-3\epsilon)} \\ &= \epsilon + (2^{nR}-1)2^{-n(I(X;Y)-3\epsilon)} \\ &\le \epsilon + 2^{-n(I(X;Y)-R-3\epsilon)} \\ &\le 2\epsilon \end{aligned}$$

  assuming ϵ is small enough so that $3\epsilon < I(X;Y) - R$ (recall $R < I(X;Y)$)
- Therefore $\bar{P}_\epsilon^{(n)} = \sum_m 2^{-nR} B_m \le 2\epsilon$
- Since $\bar{P}_e^{(n)} = \sum_c p(c) P_e^{(n)}(c) \le 2\epsilon$, there exists at least one codebook c^* such that $P_e^{(n)}(c^*) \le 2\epsilon$, can be made arbitrarily small

Maximal probability of error

- Let c' be a code containing the best half of the codewords from c^*.
  i.e. the ones with smallest probability of error



- c' is a $(2^{nR}/2, n)$ code, rate is $R - \dfrac{1}{n}$
- Let $\lambda^{(n)}(c')$ be the maximal probability of error for c'

$$
\begin{aligned}
2\epsilon \geq P_e^{(n)}(c^*) &= \frac{1}{2^{nR}} \sum_{m=1}^{2^{nR}} \lambda_m(c^*) \\
&\geq \frac{1}{2^{nR}} \sum_{\substack{m:2^{nR}/2 \text{ worst} \\ \text{codewords in } c^*}} \lambda_m(c^*) \\
&\geq \frac{1}{2^{nR}} \cdot \frac{2^{nR}}{2} \lambda^{(n)}(c') = \frac{\lambda^{(n)}(c')}{2}
\end{aligned}
$$

- Thus $\lambda^{(n)}(c') \leq 4\epsilon$, can be made arbitrarily small

Equality in the Channel Coding Converse

Repeating the steps in the converse, assuming $P_e^{(n)} = 0$:

$$\begin{aligned}
nR &= I(W; Y^n) \\
&\overset{\text{(a)}}{\leq} I(X^n; Y^n) \\
&= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) \\
&\overset{\text{(b)}}{\leq} \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) \\
&= \sum_{i=1}^{n} I(X_i; Y_i) \\
&\overset{\text{(c)}}{\leq} nC^{(I)}
\end{aligned}$$

When does equality occur?

$$I(W;Y^n) \overset{(a)}{\leq} I(X^n;Y^n)$$

- Note that $W \to X^n \to Y^n$, but also $X^n \to W \to Y^n$ since X^n is a function of W
- $I(W;Y^n) \leq I(X^n;Y^n)$ and $I(X^n;Y^n) \leq I(W;Y^n)$, so $I(W;Y^n) = I(X^n;Y^n)$
- Equality always holds!

$$H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) \overset{(b)}{\leq} \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i)$$

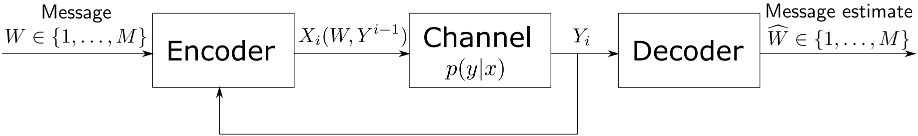
- Equality if $\{Y_i\}$ are independent

$$\sum_{i=1}^{n} I(X_i;Y_i) \overset{(c)}{\leq} nC^{(I)}$$

- Equality if the distribution of X_i is $p^*(x) = \arg\max_{p(x)} I(X;Y)$ for all i

Therefore, Y^n must be i.i.d. with distribution $p^*(y) = \sum_x p^*(x)p(y|x)$

Feedback Capacity



Notation: $Y^{i-1} = (Y_1, \ldots, Y_{i-1})$

An (M, n) **feedback code** consists of

- a sequence of encoding functions $x_i(W, Y^{i-1})$
- a decoding function $g : \mathcal{Y}^n \to \{1, 2, \ldots, M\}$

Probability of error and achievable rates are defined as for ordinary capacity.

The **feedback capacity** C_{FB} is the supremum of all rates achievable by feedback codes

Theorem

$C_{FB} = C^{(I)}$.

i.e. feedback does not increase capacity

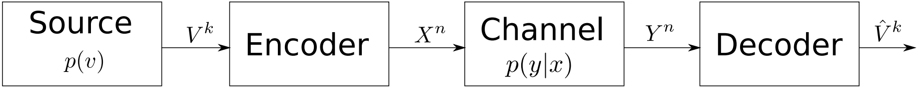
Achievability proof: Easy, since $C_{\mathsf{FB}} \geq C$

Converse proof: Assume there exists a sequence of $(2^{nR}, n)$ feedback codes with avg. probability of error $P_e^{(n)} \to 0$

Note that $W \to Y^n \to \widehat{W}$ is a Markov chain but $W \to X^n \to Y^n$ is not

$$
\begin{aligned}
nR = H(W) &= I(W;Y^n) + H(W|Y^n) \\
&\leq I(W;Y^n) + n\epsilon_n && \text{Fano's inequality} \\
&= H(Y^n) - H(Y^n|W) + n\epsilon_n \\
&= H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y^{i-1}, W) + n\epsilon_n \\
&= H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y^{i-1}, W, X_i) + n\epsilon_n \\
&= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) + n\epsilon_n \\
&\leq \sum_{i=1}^{n} I(X_i;Y_i) + n\epsilon_n \\
&\leq nC^{(I)} + n\epsilon_n
\end{aligned}
$$

Joint Source-Channel Coding



The source $V^k \overset{\text{iid}}{\sim} p(v)$

A (k, n) joint source-channel code consists of

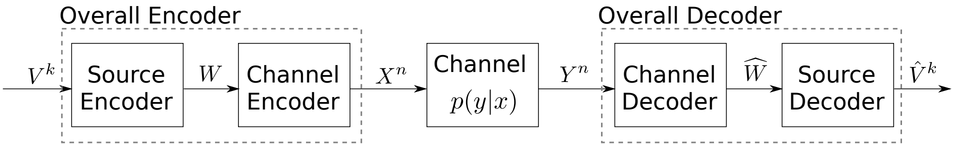
- An encoder $x^n : \mathcal{V}^k \to \mathcal{X}^n$
- A decoder $g : \mathcal{Y}^n \to \mathcal{V}^k$

The **rate** of a joint source-channel code is $R = \dfrac{k}{n}$

A rate R is **achievable** if there exists a sequence of (nR, n) codes such that $\Pr\{g(Y^n) \neq V^k\} \to 0$ as $n \to \infty$

The **joint source-channel coding capacity** C_{JSCC} is the supremum of all achievable rates

The "off the shelf" strategy



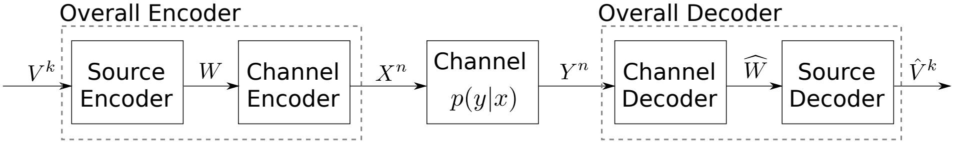
$W, \widehat{W} \in \{1, \ldots, 2^{n\tilde{R}}\}$

Achieves low probability of error if

$$kH(V) < n\tilde{R} < nC \qquad \text{i.e. } \frac{k}{n} < \frac{C}{H(V)}$$

where C is capacity of $p(y|x)$ and $H(V)$ is entropy of $p(v)$

The "off the shelf" strategy



$W, \widehat{W} \in \{1, \ldots, 2^{n\tilde{R}}\}$

Achieves low probability of error if

$$kH(V) < n\tilde{R} < nC \qquad \text{i.e. } \frac{k}{n} < \frac{C}{H(V)}$$

where C is capacity of $p(y|x)$ and $H(V)$ is entropy of $p(v)$

Theorem (Source-channel separation theorem)

$$C_{\text{JSCC}} = \frac{C}{H(V)}$$

This is an example of a **separation principle** — each component may be designed independently without loss of optimality

Achievability proof: Easy, using "off the shelf" strategy

Converse proof: Assume a sequence of (k, n) codes where $k = nR$, with probability of error $P_e^{(n)} \to 0$. We prove $R \leq \frac{C}{H(V)}$

By Fano's inequality,

$$\begin{aligned} H(V^k|Y^n) &\leq 1 + P_e^{(n)} \log |\mathcal{V}^k| \\ &= 1 + P_e^{(n)} k \log |\mathcal{V}| \\ &= 1 + P_e^{(n)} nR \log |\mathcal{V}| \\ &= n\epsilon_n \end{aligned}$$

where $\epsilon_n \to 0$ as $n \to \infty$.

$$\begin{aligned} kH(V) &= H(V^k) \\ &= I(V^k; Y^n) + H(V^k|Y^n) \\ &\leq I(V^k; Y^n) + n\epsilon_n \\ &\leq I(X^n; Y^n) + n\epsilon_n \\ &\leq nC + n\epsilon_n \end{aligned}$$

Thus $RH(V) = \dfrac{k}{n} H(V) \leq C + \epsilon_n$, so taking the limit gives $R \leq \dfrac{C}{H(V)}$