Tuesday, November 27th, 2012

Chapter 6. Polar codes

I- Motivation

We will restrict our discussion to channels with binary inputs $\mathcal{X} = \{0, 1\}$ and $\mathcal{Y}$ can be anything

Among these channels there are two "extremal" channels over which it is easy to communicate at capacity:

1/ Completely useless channels where $C = 0$. These channels have $p(y/0) = p(y/1)$ (output independent of input).

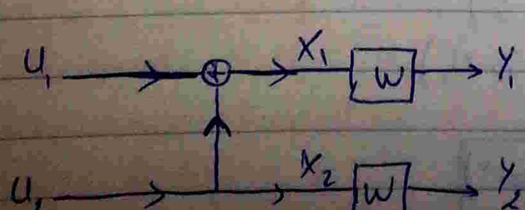
2/ Completely noiseless channels where $C = 1$. These channels have $p(y/0)\,p(y/1) = 0$ (output determines the input).

Polar codes try to turn a given channel to one of the above "extremal" channels where it is easy to communicate at capacity.

II - Polar transform

Polar transform is a method to create extremal channels from multiple-uses of a given channel W.

1. 2x2 building block



Set $X_1 = U_1 \oplus U_2$ (mod 2 sum)
$X_2 = U_2$

Suppose U_1, U_2 are independent and uniformly distributed on $\{0, 1\}$
$\Rightarrow$ (X_1, X_2) are also independent, uniformly distributed on $\{0, 1\}$

$I(U_1 U_2 ; Y_1 Y_2) = I(X_1 X_2 ; Y_1 Y_2)$ since $(U_1 U_2)$ are in one-to one correspondence with $(X_1 X_2)$

$= I(X_1 ; Y_1) + I(X_2 ; Y_2)$ since channel is memoryless and input independent

$= 2\, I(W)$

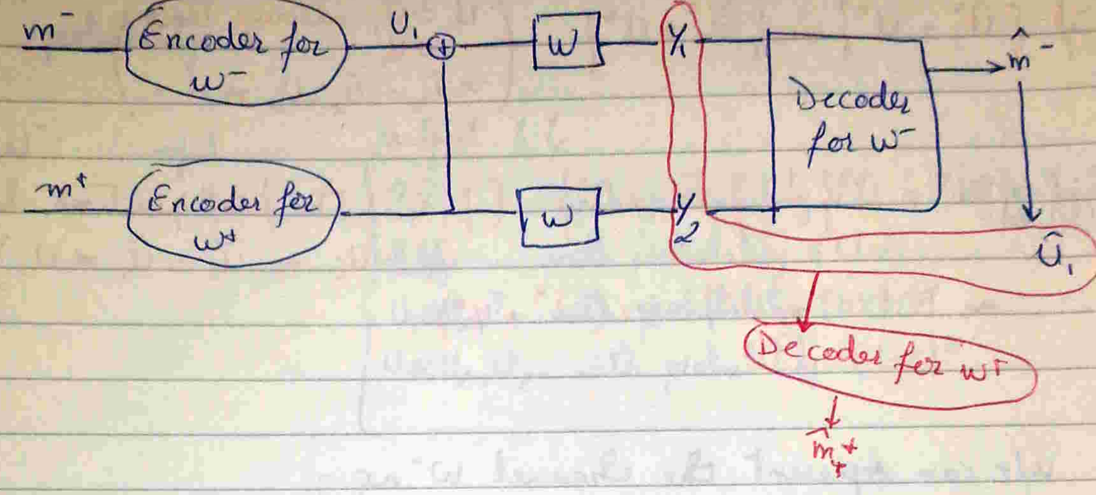
where $I(W)$ is the mutual information between input and output of W when input is uniform on $\{0, 1\}$

So far $2 I(W) = I(U_1 U_2 , Y_1 Y_2)$
$= I(U_1 , Y_1 Y_2) + I(U_2 , Y_1 Y_2 / U_1)$ chain rule
$= I(U_1 , Y_1 Y_2) + I(U_2 , Y_1 Y_2 U_1)$ since U_1, U_2 independent
$\Rightarrow I(U_2 , Y_1 Y_2 U_1)$
$= I(U_2 , U_1) + I(U_2 , Y_1 Y_2 / U_1)$

Consider the two synthetic channels:
1/ W^- : input U_1, output $(Y_1 Y_2)$
2/ W^+ : input U_2, output $(Y_1 Y_2 U_1)$

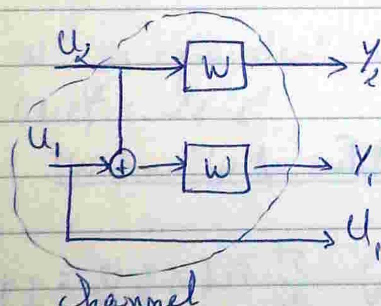
$\Rightarrow \boxed{2\, I(W) = I(W^-) + I(W^+)}$

However to obtain U_1 as an output of W^+ we can impose a successive decoder order at the receiver



See after 2 pages

So we can represent channel W^+ as:



$$I(W^+) = I(U_2 ; Y_2 Y_1 U_1)$$
$$\geq I(U_2 ; Y_2)$$
$$= I(W)$$

$$\Rightarrow I(W^-) \leq I(W) \leq I(W^+)$$

Remark: On "helped" and "unhelped" decoding

Helped decoding:

we decode $U_1, U_2, \dots U_m$ from some observation $Y_1, Y_2, \dots Y_n$

$\hat{U}_1 = \phi_1 (Y_1, \dots Y_n)$

$\hat{U}_2 = \phi_2 (Y_1, \dots Y_n, U_1)$

$\hat{U}_i = \phi_i (Y_1, \dots Y_n, U_1, \dots U_{i-1})$

$\hat{U}_n = \phi_n (Y_1, \dots Y_n, U_1, \dots U_{n-1})$

Unhelped decoding

we decode $U_1, \dots U_m$ from some observation $Y_1, \dots Y_n$

$\tilde{U}_1 = \phi_1 (Y_1 \dots Y_n) = \hat{U}_1$

$\tilde{U}_2 = \phi_2 (Y_1, \dots Y_n, \tilde{U}_1)$

$\tilde{U}_i = \phi_i (Y_1, \dots Y_n, \tilde{U}_1, \dots \tilde{U}_{i-1})$

$\tilde{U}_n = \phi_n (Y_1, \dots Y_n, \tilde{U}_1, \dots \tilde{U}_{n-1})$

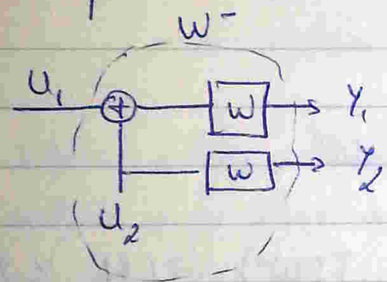
$\Pr(\hat{U}^n \neq U^n)$ = probability of error of helped system
$\Pr(\tilde{U}^n \neq U^n)$ = probability of error of unhelped system

if $\{\hat{U}^n = U^n\} \Rightarrow \{\tilde{U}^n = U^n\}$ $\left(\begin{array}{l} \tilde{U}_1 = \hat{U}_1 \text{ always true} \\ \quad = U_1 \\ \Rightarrow \tilde{U}_2 = \hat{U}_2 = U_2 \Rightarrow \dots \end{array} \right)$

if $\{\hat{U}^n \neq U^n\} \Rightarrow \{\tilde{U}^n \neq U^n\}$ (if i is the first position where $\hat{U}_i \neq U_i \Rightarrow \tilde{U}_i = \hat{U}_i \neq U_i$)

$\Rightarrow \Pr(\hat{U}^n \neq U^n) = \Pr(\tilde{U}^n \neq U^n)$

We can represent the channel W^- as:



Channel transition probabilities of W^- and W^+

$$W^-(y_1 y_2 / u_1) = \Pr(Y_1 Y_2 = y_1 y_2 / U_1 = u_1) = \sum_{u_2} \Pr(Y_1 Y_2 U_2 = y_1 y_2 u_2 / U_1 = u_1)$$

$$= \sum_{u_2} \underbrace{\Pr(U_2 = u_2 / U_1 = u_1)}_{=1/2} \Pr(Y_1 Y_2 = y_1 y_2 / U_1 U_2 = u_1 u_2)$$

$$= \frac{1}{2} \sum_{u_2} W(y_1 / u_1 \oplus u_2)\, W(y_2 / u_2)$$

$$= \frac{1}{2} W(y_1 / u_1)\, W(y_2 / 0) + \frac{1}{2} W(y_1 / u_1 \oplus 1)\, W(y_2 / 1)$$

Similarly $W^+(y_1 y_2 u_1 / u_2) = \frac{1}{2} W(y_1 / u_1 \oplus u_2)\, W(y_2 / u_2)$

2- Example 1

Let W be the Binary Erasure channel with erasure probability ε

W^-: input U_1 $\longrightarrow$ output $Y_1 Y_2$

$$\begin{cases} ?,? & \text{with probability } \varepsilon^2 \text{ (both outputs erased)} \\ ?,U_2 & \text{with probability } \varepsilon(1-\varepsilon) \\ U_1 \oplus U_2, ? & \text{with probability } (1-\varepsilon)\varepsilon \\ U_1 \oplus U_2, U_2 & \text{with probability } (1-\varepsilon)^2 \end{cases}$$

$$= \begin{cases} ? \\ ? \\ ? \\ U_1 & \text{with probability } (1-\varepsilon)^2 \end{cases}$$

So W^- is equivalent to BEC with erasure probability $1-(1-\varepsilon)^2 = 2\varepsilon - \varepsilon^2$

W^+: input U_2 $\longrightarrow$ output $U_1 Y_1 Y_2$

$$\begin{cases} U_1, ?, ? & \text{with probability } \varepsilon^2 \\ U_1, ?, U_2 & \text{with probability } (1-\varepsilon)\varepsilon \\ U_1, U_1 \oplus U_2, ? & \text{with probability } \varepsilon(1-\varepsilon) \\ U_1, U_1 \oplus U_2, U_2 & \text{with probability } (1-\varepsilon)^2 \end{cases}$$

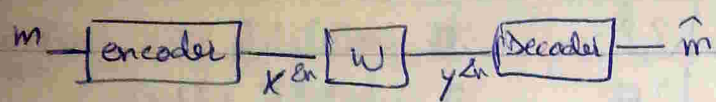
$$= \begin{cases} ? & \text{with probability } \varepsilon^2 \\ U_2 \\ U_2 \\ U_2 \end{cases}$$

So W^+ is equivalent to BEC with erasure probability ε^2

3- Explanation of successive decoder

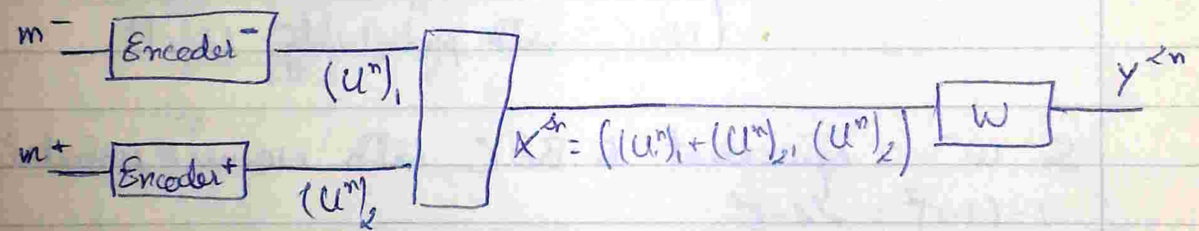
We want to develop a way to implement channels W^- and W^+.

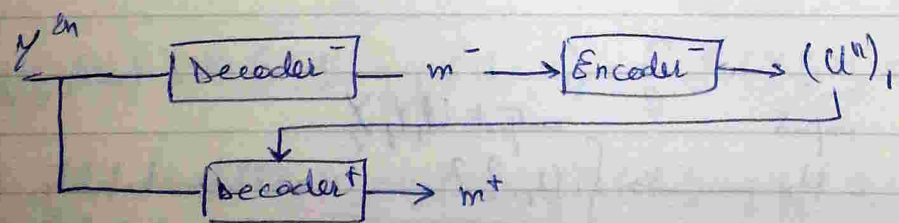
So we want to design a system



in the following way:

- Divide the message m into m^- and m^+ parts
  so $m = (m^-, m^+)$
- Consider the following construction





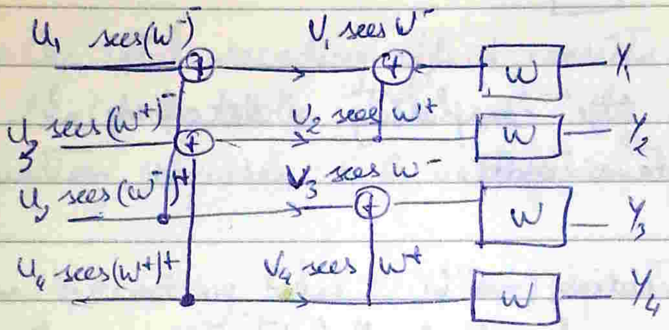
- So for $\boxed{\text{decoder}^-}$ it seems as if we have been using channel W^- since $(U^n)_2$ acts as a noise to the input $(U^n)_1$ and the output used is $Y^{<n}$.
- For $\boxed{\text{decoder}^+}$ it seems as if we have been using channel W^+ since $(U^n)_1$ acts as a noise to $(U^n)_2$ (input of the channel). Moreover $\boxed{\text{decoder}^+}$ uses both $Y^{<n}$

and $(u^n)_1$ ~~good~~ generated at receiver to decode m^*. So it is as if the output of the channel is $(Y^{2n}, (u^n)_1)$

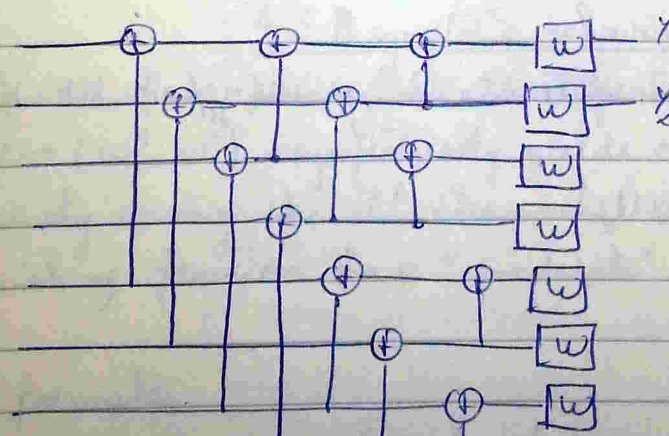
4. Recursive Channels

The hope is that the channels W^- and W^+ will be "easier" channels to communicate over, and to use polar transform recursively over them.

For example at one more level of the construction



4.1. Encoding capacity



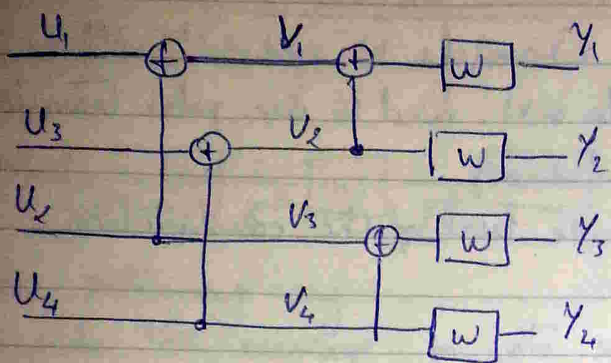
if $n = 2^l$ is the block length of the l'th level of polarization, the encoding complexity is:

$\frac{2^l}{2} l$ additions and $\frac{2^l}{2} l$ copies

so we have $\boxed{O(n \log n) \text{ operations}}$

So we see that encoding is cheap.

4.2. Decoding complexity



In order to calculate the complexity of decoding we need to follow a certain order.

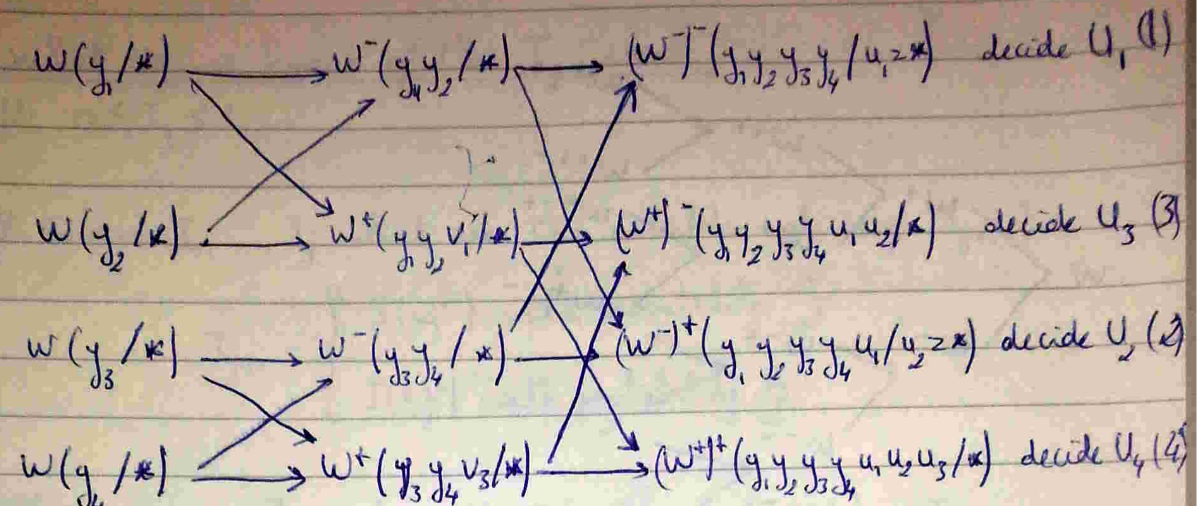
- First we start by deciding on u_1
  To decide on u_1 we need to find $(W^-)^-(y_1 y_2 y_3 y_4 / u_1 = *)$
This quantity is determined by $W^-(y_1 y_2 / *)$ and $W^-(y_3 y_4 / *)$

Similarly $W^-(y_1 y_2 / *)$ is determined by $W(y_1 / *)$ and $W(y_2 / *)$

- Second we decide on u_2

- we decide on u_3

- We decide u_4



So the total decoding effect consists of ~~fitting~~ filling in a $2^\ell \times (\ell+1)$ table with $W^{+--+}(y_1 \dots y_\ell / *)$, each filling equation requires ≤ 6 arithmetic equations

$\Rightarrow$ Decoding takes $O(n \log n)$ equations
So decoding is also cheap.

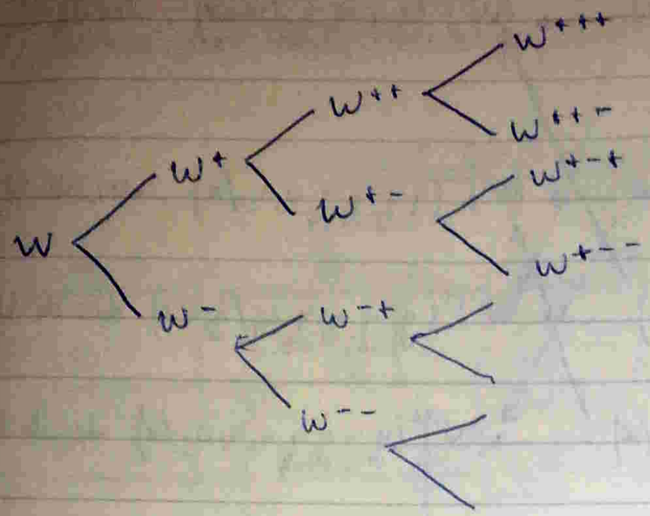
4.3 Asymptotic behavior of recursive channels

We will now show that the repeated application of the $W \to (W^+, W^-)$ asymptotically yields extremal channels. Namely among the 2^ℓ channels $W^{\overbrace{-\cdots-}^{\ell}}, \dots, W^{\overbrace{++\cdots+}^{\ell}}$ a vanishing fraction have "moderate" value of $I(W)$.

In formula,

$$\forall \delta > 0, \quad \frac{1}{2^\ell} \# \left\{ \bar{s} \in \{+,-\}^\ell \,/\, I(W^{\bar{s}}) \in (\delta, 1-\delta) \right\} \to 0 \text{ as } \ell \text{ gets large}$$

For this, let us first organize the $W^{\bar{s}}$'s in a binary tree:



Consider climbing the tree randomly:

$$W_0 = W$$

$$W_{i+1} = \begin{cases} W_i^+ & \text{with probability } \frac{1}{2} \\ W_i^- & \text{with probability } \frac{1}{2} \end{cases}$$

W_ℓ is a randomly chosen channel among the 2^ℓ channels at level ℓ of the tree.

So $I_\ell = I(W_\ell)$ is a random variable taking values in $[0, 1]$

and $Pr\left(I(W_\ell) \in (\delta, 1-\delta) \right) = \frac{1}{2^\ell} \# \left\{ \bar{s} \in \{+,-\}^\ell / I(W^{\bar{s}}) \in (\delta, 1-\delta) \right\}$

<u>Example</u>: $I_0 = I(W)$, $I_1(W) = \begin{cases} I(W^+) & \text{with probability } \frac{1}{2} \\ I(W^-) & \text{with probability } \frac{1}{2} \end{cases}$

$$I_2(W) = \begin{cases} I(W^{++}) & \text{with probability } \frac{1}{4} \\ \vdots & \\ I(W^{--}) & \text{with probability } \frac{1}{4} \end{cases}$$

What do we know of the I_j process?

1) $0 \leq I_j \leq 1$

2) Given $I_0, \dots, I_j$ what are the possible values of I_{j+1}?

$$I_{j+1} = \begin{cases} I(w_j^+) \text{ with probability } \frac{1}{2} \\ I(w_j^-) \text{ with probability } \frac{1}{2} \end{cases}$$

$$E\left[I_{j+1} / I_0 \dots I_j\right] = \frac{1}{2}\left[I(w_j^+) + I(w_j^-)\right] = \frac{1}{2} 2 I(w_j) = I_j$$

From the homework we know that for a real process

$$E\left[(I_{k+1} - I_k)(I_{j+1} - I_j)\right] = 0 \quad \text{when } k \neq j$$

Consequently $\quad 1 \geq |I_\ell - I_0|^2 \quad$ because both I_0 and I_ℓ are in $[0, 1]$

$$\Rightarrow E(1) \geq E\left[|I_\ell - I_0|^2\right]$$

$$= E\left[\left(\sum_{j=0}^{\ell-1} I_{j+1} - I_j\right)^2\right]$$

$$= \underbrace{\sum_{j=0}^{\ell-1} E\left[(I_{j+1} - I_j)^2\right]}_{\geq 0} + \sum_{j \neq k} \cancelto{0}{E\left[(I_{j+1} - I_j)(I_{k+1} - I_k)\right]}$$

$$\Rightarrow 1 \geq \sum_{j=0}^{\ell-1} E\left[(I_{j+1} - I_j)^2\right] \quad \text{for any } \ell.$$

Let $\sigma_j^2 = E\left[(I_{j+1} - I_j)^2\right]$

$$\Rightarrow \sum_{j=0}^{\ell-1} \sigma_j^2 \leq 1 \quad \text{for every } \ell$$

$$\Rightarrow \lim_{\ell \to \infty} \sigma_\ell^2 = 0$$

$$\Rightarrow E\left[(I_{\ell+1} - I_\ell)^2\right] \longrightarrow 0$$

$\Rightarrow \cancel{\Pr\left(|I_{\ell+1} - I_\ell|^2\right)}$

$$\Rightarrow \Pr\left(|I_{\ell+1} - I_\ell| > \varepsilon\right) \leq \frac{E\left((I_{\ell+1} - I_\ell)^2\right)}{\varepsilon^2}$$

$\Rightarrow$ for any $\varepsilon > 0$, $\Pr\left(|I_{\ell+1} - I_\ell| > \varepsilon\right) \longrightarrow 0$ as ℓ gets large

this is equivalent to $\frac{1}{2^\ell} \#\left\{\bar{s} \in \{+,-\}^\ell \,/\, I(W^{\bar{s}}) - I(W^{\bar{s}-}) > \varepsilon\right\} \to 0$ as ℓ gets large

Interpretation: At level ℓ the next channel is very close to the present one in terms of mutual information

Lemma 1

$\forall \delta > 0$, $\exists\, \varepsilon > 0$ such that

$$I(W) - I(W^-) \leq \varepsilon$$

$$\Rightarrow I(W) \notin (\delta, 1-\delta)$$ (Proof after 3 pages)

This means that if the channel does not "move by much" then it must be an extremal channel (either good or bad).

Using this lemma 1 we can deduce that

$$\frac{1}{2^\ell} \#\left\{\bar{s} \in \{+,-\}^\ell \,/\, I(W^{\bar{s}}) \in (\delta, 1-\delta)\right\} \to 0 \text{ as } \ell \text{ gets large}$$

$\Rightarrow$ Polarization happens

So as ℓ gets large the channel becomes extremal and hence their polarization

Further note: $I(W^-) + I(W^+) = 2I(W)$

$$\underbrace{I(W^{--}) + I(W^{-+})}_{2I(W^-)} + \underbrace{I(W^{+-}) + I(W^{++})}_{2I(W^+)} = 4I(W)$$

$$\rightarrow \frac{1}{2^l} \sum_{\bar{s} \in \{+,-\}^l} I(W^{\bar{s}}) = I(W) \quad \text{for every } l.$$

Theorem 1:

For any $\delta > 0$,

(1) $\frac{1}{2^l} \# \{ \bar{s} \in \{+,-\}^l / I(W^{\bar{s}}) \in [1-\delta, 1] \} \longrightarrow I(W)$ (Number of good channels)

(2) $\frac{1}{2^l} \# \{ \bar{s} \in \{+,-\}^l / I(W^{\bar{s}}) \in [0, \delta] \} \longrightarrow 1 - I(W)$ (Number of bad channels)

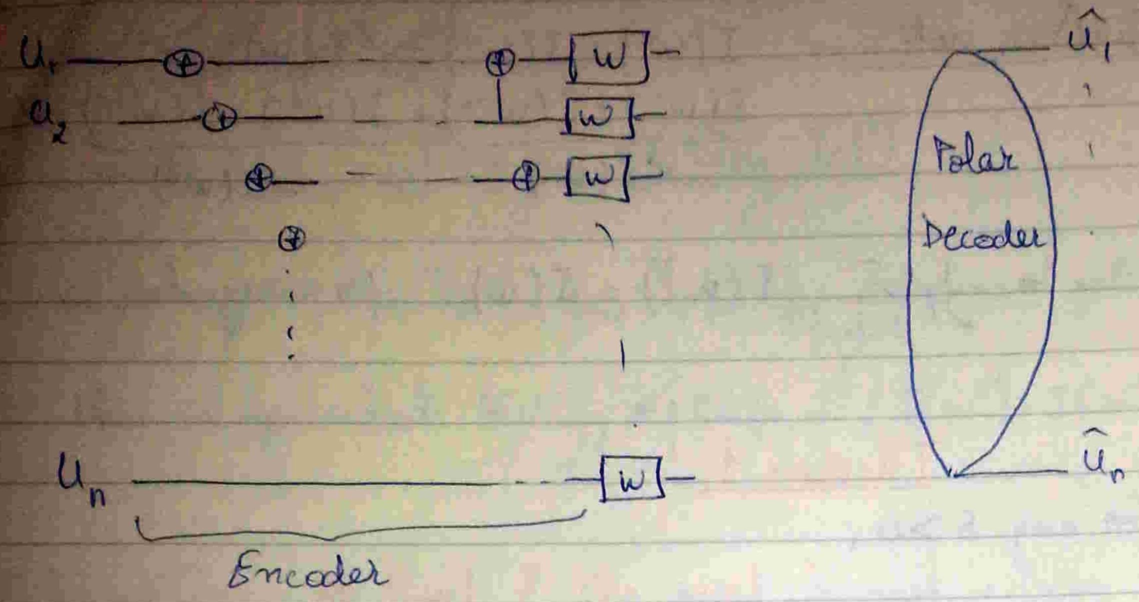
(3) $\frac{1}{2^l} \# \{ \bar{s} \in \{+,-\}^l / I(W^{\bar{s}}) \in (\delta, 1-\delta) \} \longrightarrow 0$ (Number of moderate ~~chan~~ channels)

4.4. Alternative design for communication systems

The results of the previous section suggest a method of designing a communication system to transmit data over W:

Given $R < I(W)$, $\delta > 0$, pick l large enough so that $\frac{1}{2^l} \# \{ \bar{s} \in \{+,-\}^l / I(W^{\bar{s}}) \geq 1-\delta \} \geq R$

Then construct a polar encoder/decoder for polarization depth l, with $n = 2^l$



- Once this is done, we will have i's such that u_i sees a good channel (there are $\geq R2^l$ of them). Use these indices to send data.
- The remaining u_i's freeze them to randomly chosen values in $\{0,1\}$ known to both sender and receiver.

This gives us a code that operates at rate $\geq R$ and where each data bit travels on a synthetic channel with mutual information $\geq 1-\delta$.

<u>N.B</u>: The input to bad channels are chosen randomly rather than in a deterministic way since in our calculations of $I(W^-)$ and $I(W^+)$ we assumed the inputs u_1 and u_2 to be iid drawn from $\{0,1\}$ according to uniform distribution

Remarks

1/ For "symmetric channels there is no need to choose the frozen U_i's by a random experiment. They can be chosen to equal 0.

Definition 1

A binary input channel W is said to be symmetric if there is a function

$$\pi : \mathcal{Y} \longrightarrow \mathcal{Y} \text{ with}$$

(1) $\pi(\pi(y)) = y$ (π is its own inverse)

(2) $W(y/0) = W(\pi(y)/1)$

Example:

- BSC is symmetric $\left(\pi : \begin{array}{l} 0 \to 1 \\ 1 \to 0 \end{array}\right)$
- BEC is symmetric $\left(\pi : \begin{array}{l} 0 \to 1 \\ 1 \to 0 \end{array} \quad \mathcal{E} \to \mathcal{E}\right)$
- Binary input additive Gaussian channel

$$\mathcal{X} = \{0, 1\}, \quad \mathcal{Y} = \mathbb{R}$$

$$Y = (-1)^x + Z \qquad Z \sim \mathcal{N}(0, \sigma^2)$$

Check that ~~$\pi(y) = y$~~ $\pi(y) = -y$ satisfies (1) and (2) in the symmetry definition, i.e this channel is also symmetric.

III. On the error probability of polar codes

Recall the difference between decoding with help and the unhelped decoding

$$\hat{u}_1 = \phi_1(Y_1 \dots Y_n)$$
$$\hat{u}_2 = \phi_2(Y_1 \dots Y_n\, u_1)$$
$$\vdots$$
$$\hat{u}_n = \phi_n(Y_1 \dots Y_n\, u_1 \dots u_{n-1})$$

so $P_r(\hat{u}_i \neq u_i)$ = probability of error if a bit is sent on $W^{\bar{s}}$ (where $\bar{s}$ is the sequence of $\{+,-\}$ corresponding to u_i)

<u>Example:</u> $n=8$, u_1 travels on W^{---}, $u_1 \to Y^8$
u_2 travels on W^{--+}, $u_2 \to Y^8 u_1$
u_3 travels on W^{-+-}, $u_3 \to Y^8 u_1 u_2$
$\vdots$

The unhelped decoder (implemented by the polar decoder).

$$\tilde{u}_1 = \phi_1(Y^n)$$
$$\tilde{u}_2 = \phi_2(Y^n \tilde{u}_1)$$
$$\vdots$$
$$\tilde{u}_n = \phi_n(Y^n \tilde{u}^{n-1})$$

we had shown that

$$P_r(\tilde{u}^n \neq u^n) = P_r(\hat{u}^n \neq u^n) \leq \sum_{i=1}^{n} P_r(\hat{u}_i \neq u_i)$$

$$P_r(\hat{u}_i \neq u_i) = \begin{cases} 0 & i \text{ is frozen} \\ P_e(W^{\bar{s}}) & \bar{s} \text{ is the index of the channel on which } u_i \text{ is sent.} \end{cases}$$

with $P_e(W) = \frac{1}{2}\sum_y W(y|0)\, \mathbb{1}\{W(y|1) \geq W(y|0)\} + \frac{1}{2}\sum_y W(y|1)\, \mathbb{1}\{W(y|0) \geq W(y|1)\}$

It is not diffult to show that

$$I(W) \geq 1-\delta \Rightarrow P_e(W) \leq \delta$$

Consequently,

$P(\tilde{U}^n \neq U^n)$ of the polar codes is upper bounded by:

$$\underset{\substack{\bar{s}:\bar{s} \text{ is} \\ \text{an index} \\ \text{that is used} \\ \text{for data} \\ \text{transmission}}}{\sum{}'} \underbrace{\left(1 - I(W^{\bar{s}})\right)}_{\leq \delta_\ell}$$

It turns out ~~that~~ the polarization takes place faster than we proved. Indeed the following holds:

$$\frac{1}{2^\ell} \# \left\{ \bar{s} \in \{+,-\}^\ell \,/\, I(W^{\bar{s}}) \in (\delta_\ell, 1-\delta_\ell) \right\} \to 0$$

$$\text{and} \quad \delta_\ell = \frac{1}{2^{2^{\beta\ell}}} \qquad \beta < \frac{1}{2}$$

So by increasing the range of moderate codes we still get fast polarization

Consequently, $(n = 2^\ell)$

$$P(\tilde{U}^n \neq U^n) \leq 2^\ell \frac{1}{2^{2^{\beta\ell}}} \to 0 \quad \text{as } \ell \to \infty$$

Corner Proof:

Proof of Lemma 1

Lemma 1 can be stated as follows:

if $I(W) - I(W^-)$ is small then $I(W)$ is extremal

Proof:

Lemma 2: if (X_1, Y_1) and (X_2, Y_2) are independent and if

X_1 and X_2 are $\mathbb{F}_2$ valued ($\{0,1\}$)

and suppose $H(X_1/Y_1) = h_2(p_1)$

$H(X_2/Y_2) = h_2(p_2)$

then $H(X_1 \oplus X_2 / Y_1 Y_2) \geq h_2(p_1 * p_2)$

where $p_1 * p_2 = p_1(1-p_2) + p_2(1-p_1)$

Before we prove Lemma 2, let's see how we can use it:

Recall $W: X \to Y$

$$J(W) = I(X,Y)\Big|_{X\sim \text{uniform}\{0,1\}} = H(X) - H(X/Y) = 1 - H(X/Y)$$

$W^-: U_1 \to Y_1 Y_2$

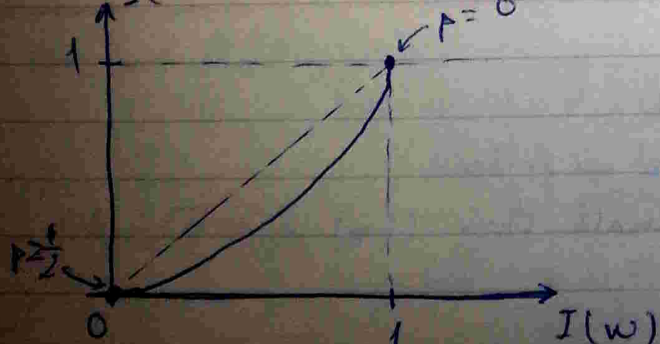
$$I(W^-) = H(U_1) - H(U_1/Y_1Y_2)$$

$$= 1 - \underbrace{H(X_1 \oplus X_2/Y_1 Y_2)}_{\substack{(X_1,Y_1) \text{ and } (X_2,Y_2) \text{ are independent} \\ \text{identically distributed as } (X,Y)}}$$

since $U_1 = X_1 \oplus X_2$ is uniform $\{0,1\}$

Thus if $H(X/Y) = h(p)$

then $H(X_1 \oplus X_2 / Y_1 Y_2) \geq h_2(p * p)$ (by lemma 2 called Mrs Gerber's lemma)



if $I(W) = 1 - h_2(p)$

$\Rightarrow I(W^-) \leq 1 - h_2(p * p)$

$\Rightarrow$ for every $\delta > 0$, $\exists \varepsilon > 0$ such that $I(W^-) \geq I(W) - \varepsilon \Rightarrow I(W) \notin (\delta, 1-\delta)$

This proves Lemma 1. So Mrs. Gerber's is sufficient to prove "polarization happens".

<u>Proof of Mrs. Gerber's lemma:</u>

Define $\eta_1(y_1) = H(X_1 / Y_1 = y_1)$ so that $\sum_{y_1} p_{Y_1}(y_1)\,\eta_1(y_1) = h_2(p_1)$

$\eta_2(y_2) = H(X_2 / Y_2 = y_2)$ so that $\sum_{y_2} p_{Y_2}(y_2)\,\eta_2(y_2) = h_2(p_2)$

$$H(X_1 \oplus X_2 / Y_1 = y_1, Y_2 = y_2) = h_2\big(\underbrace{h_2^{-1}(\eta_1(y_1))} * \underbrace{h_2^{-1}(\eta_2(y_2))}\big)$$

$Pr(X_1 \oplus X_2 = 1 / y_1, y_2) = Pr(X_1 = 1 / Y_1 = y_1) * Pr(X_2 = 1 / Y_2 = y_2)$

$$\text{so } H(X_1 \oplus X_2 / Y_1 Y_2) = \sum_{y_1} p_{Y_1}(y_1) \sum_{y_2} p_{Y_2}(y_2)\, h_2\big(h_2^{-1}(\eta_1(y_1)) * h_2^{-1}(\eta_2(y_2))\big)$$

$$\overset{(a)}{\geq} \sum_{y_1} p_{Y_1}(y_1)\, h_2\Big(h_2^{-1}(\eta_1(y_1)) * h_2^{-1}\big(\sum_{y_2} p_{Y_2}(y_2)\,\eta_2(y_2)\big)\Big)$$

$$= \sum p_{Y_1}(y_1)\, h_2\big(h_2^{-1}(\eta_1(y_1)) * h_2^{-1}(h_2(p_2))\big)$$

$$= \sum p_{Y_1}(y_1)\, h_2\big(h_2^{-1}(\eta_1(y_1)) * p_2\big)$$

$$\overset{(b)}{\geq} h_2\Big(h_2^{-1}\big(\underbrace{\textstyle\sum_{y_1} p_{Y_1}(y_1)\,\eta_1(y_1)}_{h_2(p_1)}\big) * p_2\Big)$$

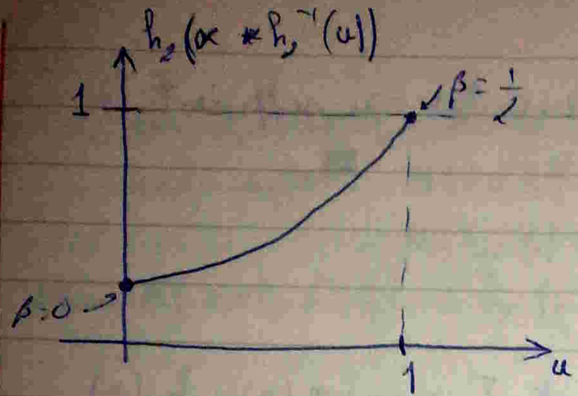
$$= h_2(p_1 * p_2)$$

(a) and (b) would be resolved if the function

$u \longrightarrow h_2(\alpha * h_2^{-1}(u))$ is convex $\cup$ in u

because then we would have $\sum_u p(u)\, h_2(\alpha * h_2^{-1}(u))$

$$\geq h_2\big(\alpha * h_2^{-1}\big(\textstyle\sum_u p(u)\,u\big)\big)$$



Set $u = h_2(\beta)$ and plot $h_2(\beta)$ on the horizontal axis and $h_2(\alpha * \beta)$ in the vertical axis and sweep β from 0 to $\frac{1}{2}$

The function $h_2(\alpha * h_2^{-1}(h(\beta))) = h_2(\alpha * \beta)$ is convex

$\equiv$ slope is increasing as we move from left to right

$$\equiv \frac{\frac{\partial}{\partial\beta} h_2(\alpha * \beta)}{\frac{\partial}{\partial\beta} h_2(\beta)} \text{ is increasing with } \beta$$

But $\frac{\partial}{\partial\beta} h_2(\beta) = \log \frac{1-\beta}{\beta}$

and $\frac{\partial}{\partial\beta} h_2(\alpha * \beta) = (1 - 2\alpha) \log \frac{1-(\alpha * \beta)}{\alpha * \beta}$

and it is easy to verify that the ratio is increasing in β ∎