Introduction to Information Theory

Lecture 12 - Polar Codes Part a

Lecturer: Haim Permuter *Scribe: Yhonatan Gayer, Yakov Gusakov*

I. Intro to Polar Codes

In this lecture, we present an introduction to Polar Codes. Polar Codes were first introduced by Arikan Erdal in 2009 [1]. In 2016 they were already accepted as part of 5G channel coding technology. Over the course of three lectures, we will explore the effectiveness of this code, and study how it is used to enhance capacity and reliability. Our discussion begins with a motivation for polar coding, followed by a step-by-step construction of the initial polarization block using an intuitive example on the erasure channel (BEC). Hence this lecture will be on polar codes in general but will focus on the example of BEC. Subsequently, we generalize this algorithm to encompass n inputs and any type of channel. The content of these lecture notes is based on the foundational work by Arikan [1], complemented by the insights from [2] and [3].

II. Repetition Coding

To improve the accuracy of estimation of channel input, a proposed approach involves utilizing multiple instances of the same channel. By repeating the use of the channel a total of n times, we can increase our confidence in the correctness of the estimation. This strategy is depicted in Fig. 1 as the Repetition Coding Scheme, where W represents the channel.

The code can be analyzed in terms of the error probability P_e and the rate. To assess the reliability of the channel, we'll be examining the error probability P_e, for which $\lim_{n\to\infty} P_e = 0$. This property holds true for most channels, including BSC, BEC. Regarding the rate, each bit of information is transmitted approximately n times (depending on the specific channel). Consequently, the rate is significantly lower than

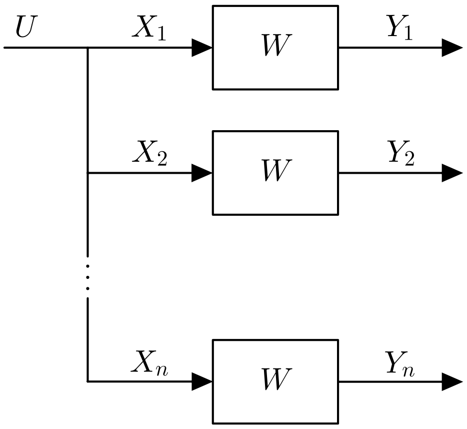

Fig. 1: Repetition Coding Scheme.

the capacity, which is given by $C_{total} = nC(W)$, where $C(W)$ denotes the capacity of the channel W.

Example 1 (Repetition Code with $n = 2$ and $BEC\,(p)$ Channels) Consider a repetition code with $n = 2$ and $BEC\,(p)$ channels, meaning we transmit each symbol twice. We also assume that $U \sim Ber\left(\frac{1}{2}\right)$. For example

$$0011 \mapsto 00001111.$$

This scenario is illustrated in Fig. 2.

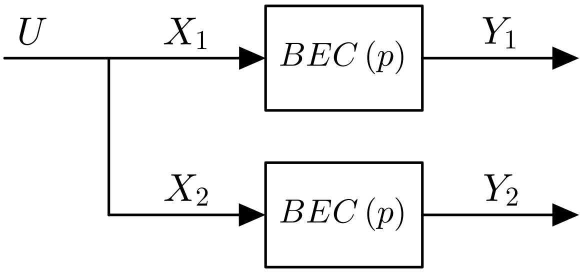

Fig. 2: Repetition code scheme with two $BEC\,(p)$ channels.

Example 2 (Rate and Reliability of Repetition code with $BEC\,(p)$ Channel) First we will examine the reliability,

$$\begin{aligned}\lim_{n\to\infty} P_e &= \lim_{n\to\infty} P(\hat{U} \neq U)\\ &\overset{(a)}{=} \lim_{n\to\infty} P(Y_1 =?, ..., Y_n =?)\\ &\overset{(b)}{=} \lim_{n\to\infty} (p)^n\\ &\overset{(c)}{=} 0,\end{aligned} \tag{1}$$

where

(a) follows from the requirement that $Y_k \neq ?$ for some k to determine U's value.

(b) follows from BEC channel having $P(Y_1 =?) = p$, and in a memory-less channel each use of the channel is independent.

(c) holds since $0 < p < 1$.

Now we will examine its rate for a constant n,

$$R = 1 << n(1-p) = nC_{BEC}, \tag{2}$$

where nC_{BEC} is the highest achievable rate when using n $BEC\,(p)$ channels. Hence, it can be inferred that Repetition coding offers reliability at the cost of reduced speed.

III. A Building Block Of Polar Coding

We saw that the repetion code can transform a channe W with erasure p into a clean channel with erasure arbitrary close to zero. However, the repetition coding is not optimal in terms of rate; it decreases the rate signifiacntly. We will now propose an alternative - polar codes. The construction of polar code is based on multiple recursive concatenations of input manipulations which transform the physical channels into virtual channels with capacities [1] that either goes to 1 or to 0. Namely the virtual channel polarize to 0 or 1. On the channels that the capacity is 1 we will transmit the our bits of information and the channels with capacity 0 we will just send freezed bits, namely, fixed bit determined a head of time with no information. Like all 0. We will see that the polar codes is an optimal code, namely, the portion of virtual channels with capacity 1 is $C(W)$, i.e., the capacity of W, and the rest of the virtual channel have a capacity 0 hence wont be used to transmit information.

We begin with the basic polarization building block for two usage of the channel,i.e., $n = 2$, and later see how it generalize it to any dimension n. In this section we focus of polar coding scheme with $n = 2$ inputs as depicted in Fig. 3 which is the building block of any general polar code. The inputs are two independent uniformly distributed variables $U \sim Ber\left(\frac{1}{2}\right)$ and $V \sim Ber\left(\frac{1}{2}\right)$. A basic manipulation on the two binary inputs U and V is

$$\begin{aligned} X_1 &= U \oplus V \\ X_2 &= V, \end{aligned} \tag{3}$$

or in vector notation:

$$\begin{bmatrix} X_1 & X_2 \end{bmatrix} = \begin{bmatrix} U & V \end{bmatrix} \underbrace{\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}}_{G}. \tag{4}$$

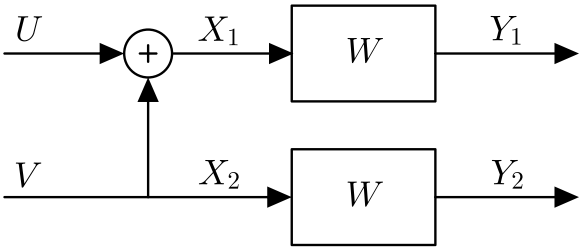

Fig. 3: Polar Code Scheme with two usage of the channel W. For instance, the channel W maybe the BEC.

The signals X_1 and X_2 are transmitted over identical and independent channel W which can be for example the $BEC\,(p)$ channels. The output of the channel are Y_1 and Y_2. The transformation of the vector $[U, V]$ to $[X_1, X_2]$ is done using a multiplication with the matrix G on the Binary field $\mathbb{F}_2$, where the alphabet is 0 and 1 and the summation is mod 2 summation. Note that the matrix G is invertible and $G \cdot G = I$, hence, $G^{-1} = G$.

Conservation of information: Since (U, V) has a one-to-one transformation to (X_1, X_2) the total mutual information does not change, namely

$$I\,(U, V; Y_1, Y_2) = I\,(X_1, X_2; Y_1, Y_2)\,. \tag{5}$$

This follow from the following lemma:

Lemma 1 (One-to-one transformation does not change mutual information) Let $z = f(x)$ be one to one transformation. Namely, there exists an f^{-1} s.t. $x = f^{-1}(z)$. Then for any r..v Y,

$$I(Z;Y) = I(X;Y) \tag{6}$$

Prove this lemma as an exercise.

Now, lets show that the transformation can achieve twice the capacity of the channel for the case of $n = 2$:

$$\begin{aligned} I\left(U,V;Y_1,Y_2\right) &\overset{(a)}{=} I\left(X_1,X_2;Y_1,Y_2\right) \\ &\overset{(b)}{=} I\left(X_1,X_2;Y_1\right) + I\left(X_1,X_2;Y_2|Y_1\right) \\ &\overset{(c)}{=} I\left(X_1;Y_1\right) + I\left(X_1,X_2;Y_2|Y_1\right) \\ &\overset{(d)}{=} I\left(X_1;Y_1\right) + I\left(X_2;Y_2\right) \\ &= 2C_{BEC}, \end{aligned} \tag{7}$$

where

(a) is due to the transformation from (U,V) to (X_1,X_2) being invertible and Lemma 1.

(b) follows from the chain rule for mutual information.

(c) follows from the memoryless of the channel and no feedback, i.e., $P(y_1|x_1,x_2) = P(y_1|x_1)$

(d) follows and the fact that Y_1 is independent of Y_2 since U is Bernouli($\frac{1}{2}$), hence $H(Y_2|Y_1) = H(Y_2)$, and from the memoryless of the channel that implies that $H\left(Y_2|Y_1,X_1,X_2\right) = H\left(Y_2|X_2\right)$.

Now, we analyse the channel's Mutual Information by utilizing the chain rule,

$$\begin{aligned} I\left(U,V;Y_1,Y_2\right) &\overset{(a)}{=} I\left(U;Y_1,Y_2\right) + I\left(V;Y_1,Y_2|U\right) \\ &\overset{(b)}{=} I\left(U;Y_1,Y_2\right) + I\left(V;Y_1,Y_2,U\right), \end{aligned} \tag{8}$$

where

(a) follows from the chain rule for mutual information.

(b) holds since U and V are independent.

Example 3 (Polar Code with two $BEC\left(p\right)$ Channels) Now lets us consider the case of polar codes $n = 2$ where the channel is BEC. The first term represents a channel $U \xrightarrow{W^-} Y_1, Y_2$ with mutual information

$$\begin{aligned} I\left(U; Y_1, Y_2\right) &= H\left(U\right) - H\left(U|Y_1, Y_2\right) \\ &\overset{(a)}{=} 1 - \sum_{y_1 \in \{0,1,?\}} \sum_{y_2 \in \{0,1,?\}} P_{Y_1,Y_2}\left(y_1, y_2\right) H\left(U|Y_1 = y_1, Y_2 = y_2\right) \\ &\overset{(b)}{=} 1 - P_{Y_1,Y_2}\left(y_1 = x_1, y_2 =?\right) - P_{Y_1,Y_2}\left(y_1 =?, y_2 = x_2\right) - P_{Y_1,Y_2}\left(y_1 =?, y_2 =?\right) \\ &\overset{(c)}{=} 1 - (1-p)p - p(1-p) - p^2 \\ &= (1-p)^2 \\ &< C_{BEC}, \end{aligned} \tag{9}$$

where

(a) follows from $H\left(V\right) = 1$ when $V \sim Ber\left(\frac{1}{2}\right)$ + definition of the conditional entropy.

(b) holds since the only case where U is known is when we know both Y_1 and Y_2, in that case $H\left(U|Y_1 = y_1, Y_2 = y_2\right) = 0$, in all other cases $H\left(U|Y_1 = y_1, Y_2 = y_2\right) = 1$.

(c) follows from $P_{Y_1,Y_2}\left(y_1, y_2\right) = P_{Y_1}(y_1)P_{Y_2}(y_2)$ since the channels are independent and we have shown that X_1, X_2 are independent, making Y_1, Y_2 independent as well.

Similarly, the second term in (8) represents a channel $V \xrightarrow{W^+} Y_1, Y_2, U$ with mutual information

$$\begin{aligned} I\left(V; Y_1, Y_2, U\right) &\overset{(a)}{=} H\left(V\right) - H\left(V|Y_1, Y_2, U\right) \\ &\overset{(b)}{=} 1 - \sum_{y_1 \in \{0,1,?\}} \sum_{y_2 \in \{0,1,?\}} P_{Y_1,Y_2}\left(y_1, y_2\right) H\left(V|Y_1 = y_1, Y_2 = y_2, U\right) \\ &\overset{(c)}{=} 1 - P_{Y_1,Y_2}\left(y_1 =?, y_2 =?\right) \\ &\overset{(d)}{=} 1 - p^2 \\ &> C_{BEC}, \end{aligned} \tag{10}$$

where

(a) follows from the chain rule for mutual information.

(b) follows from $H\left(V\right)=1$ when $V\sim Ber\left(\frac{1}{2}\right)$ + definition of the conditional entropy.

(c) holds since the only case where V is unknown is when we don't know know both Y_1 and Y_2, in that case $H\left(V|Y_1=y_1,Y_2=y_2,U\right) = 1$, in all other cases $H\left(V|Y_1=y_1,Y_2=y_2,U\right)=0$.

(d) follows from $P_{Y_1,Y_2}\left(y_1,y_2\right)=P_{Y_1}(y_1)P_{y_2}(y_2)$.

Note that the sum of $I\left(U;Y_1,Y_2\right)+I\left(V;Y_1,Y_2,U\right)$ equals, as expected from eq. (7), to $2C_{BEC}$

In total, looking at equation (8), we can interpret the two channel scheme depicted in Fig. 3 as an equivalent scheme composed of two parallel "virtual" channels:

- The W^- Channel, which corresponds to a BEC with parameter $1-(1-p)^2$.
- The W^+ Channel, which corresponds to a BEC with parameter p^2.

The Equivalent Scheme illustrated in Fig. 4 demonstrates the processing of independent variables U and V to generate the output sequences Y_1 and Y_2.

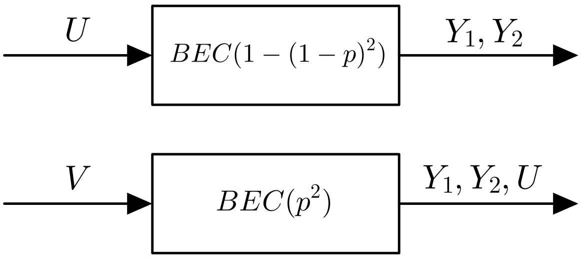

Fig. 4: Equivalent Scheme using Virtual Channels W^- and W^+.

The comparison yields four significant conclusions:

- The capacity of Channel W^- is lower than C_{BEC}.
- The capacity of Channel W^+ is greater than C_{BEC}.
- Decoding of Channel W^- requires knowledge of Y_1 and Y_2.
- Decoding of Channel W^+ requires knowledge of Y_1, Y_2, and U.

Based on our analysis of the scheme depicted in Fig. 3, we conclude that it can be effectively represented as two parallel channels, namely W^- and W^+. This finding is noteworthy as it allows us to create a channel with a higher capacity compared to the original $BEC(p)$ channel, while maintaining the same overall scheme capacity.

IV. Polar Coding of size n

In the previous section, we demonstrated the polar coding scheme for the case where $n = 2$ and using the $BEC(p)$ channel. In this section, we will generalize this approach to all channels and for all $n = 2^N$ with $N \geq 1$. By replicating the process outlined in the aforementioned example, we can further enhance the capacities of individual channels while simultaneously diminishing others.

Definition 1 (Kronecker Product) The Kronecker Product of two matrices $A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{p\times q}$ is $A \otimes B \in \mathbb{R}^{mp\times nq}$ is defined by

$$A \otimes B = \begin{bmatrix} [A]_{1,1}B & [A]_{1,2}B & \cdots & [A]_{1,n}B \\ [A]_{2,1}B & [A]_{2,2}B & \cdots & [A]_{2,n}B \\ \vdots & \vdots & \ddots & \vdots \\ [A]_{m,1}B & [A]_{m,2}B & \cdots & [A]_{m,n}B \end{bmatrix}. \tag{11}$$

Notice that the Kronecker product of the transformation matrix with itself is

$$G_2 := G^{\otimes 2} = G \otimes G = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}. \tag{12}$$

This matrix defines the manipulation on $n = 4$ inputs, U_1, U_2, U_3, U_4, generating

$$\begin{bmatrix} X_1, X_2, X_3, X_4 \end{bmatrix} = \begin{bmatrix} U_1, U_2, U_3, U_4 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} U_1 \oplus U_2 \oplus U_3 \oplus U_4 \\ U_2 \oplus U_4 \\ U_3 \oplus U_4 \\ U_4 \end{bmatrix}^T. \tag{13}$$

The polar coding scheme for $n = 4$ is depicted in Fig. 5. This scheme is equivalent to 4 virtual channels W^{++}, W^{+-}, W^{-+} and W^{--}, where the signs $+$ and $-$ relate to the increase or decrease in the capacity of the channels.

Example 4 (Polar Code with $n = 4$ $BEC(p)$ Channels) If we take the channels W to be $BEC(p)$, as we add more recursion layers we will get a polarization of virtual channel capacities as depicted in Fig. 6.

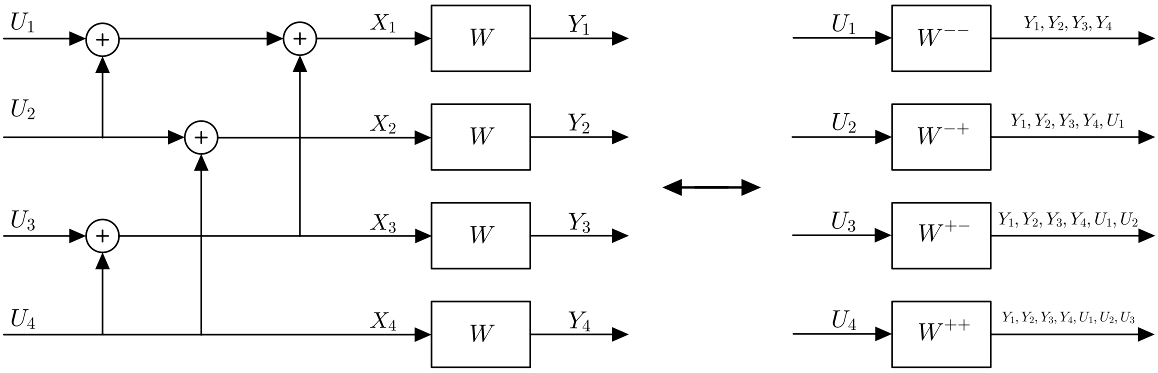

Fig. 5: Polar Coding Scheme with $n = 4$ Channels.

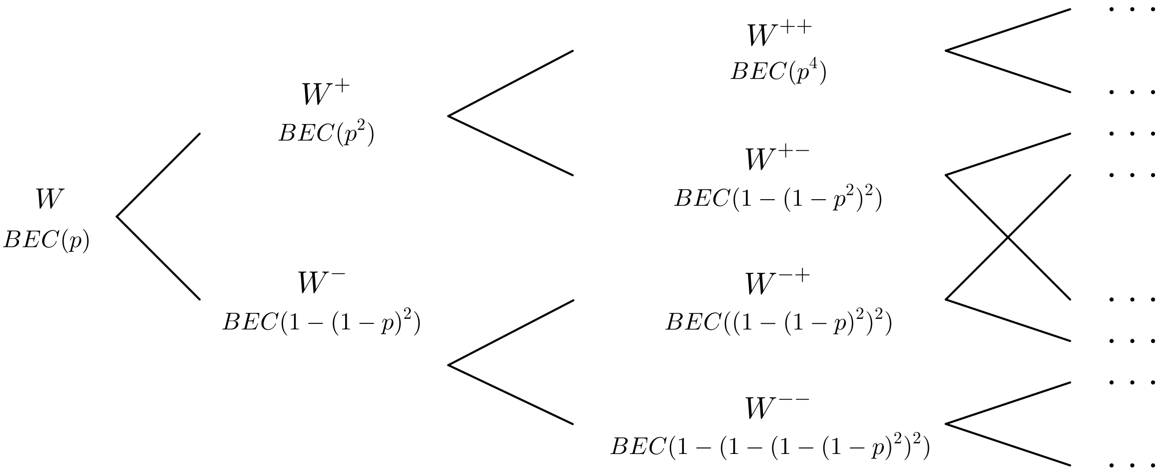

Fig. 6: Capacity of Virtual Channels as we add layers of $BEC\,(p)$ channels.

Generalizing this to $n = 2^N$ inputs $\{U_i \mid 1 \leq i \leq n\}$ we get

$$\underline{X} = \underline{U}G_N \iff X_j = \sum_{i=1}^{n} \left[G^{\otimes N}\right]_{i,j} U_i, \tag{14}$$

creating a collection of n virtual channels $\{W_i \mid 1 \leq i \leq n\}$.

V. Main theorem of Polar codes

Theorem 1 As the number of channels n grows to infinity, the capacity of the virtual channels converges to either 0 or 1. The ratio between the amount of "clean" $(C \approx 1)$

and ”dirty” $(C \approx 0)$ channels depends on the capacity $C\left(W\right)$ of the original channels:

$$\begin{aligned} &\forall \delta > 0 : \\ &\lim_{n \to \infty} \frac{\left| C\left(W_i\right) > 1 - \delta \right|}{n} = C\left(W\right) \\ &\lim_{n \to \infty} \frac{\left| C\left(W_i\right) < \delta \right|}{n} = 1 - C\left(W\right). \end{aligned} \tag{15}$$

The proof is out of scope for this course. Readers can refer to the original paper [1].

Theorem 1 states that we can approach the capacity $C = 1$ in certain channels as closely as desired, while in other channels we approach the capacity $C = 0$, provided that we choose a sufficiently large value for N. Due to the mathematical complexity associated with calculating the capacity of each channel, particularly for high orders of n, where the expressions become exponentially intricate, resorting to Monte Carlo experiments is an acceptable approach to estimate the capacity or reliability of each channel.

We can utilize the high reliability ”clean” channels with $C \approx 1$ to send data.
Question: What do we do with the low reliability ”dirty” channels with $C \approx 0$?
Answer: We ”freeze” them, constantly transmitting ’0’. It doesn’t matter what we send in particular, as long as the decoder knows their value, as they play a crucial role in the decoding process of the data channels.

VI. Decoding Polar codes

Once the data passes through the channel, we want to to decode the channel outputs $\underline{Y} = [Y_1, Y_2, \ldots, Y_n]$ in the most optimal manner to obtain the best estimate vector $\underline{\hat{U}} = \left[\hat{U}_1, ..., \hat{U}_n\right]$ for the inputs $\underline{U} = [U_1, ..., U_n]$. In the case of $n = 2$, depicted in Fig. 3, we saw that decoding W^+ to obtain $\hat{V}$ required knowledge of Y_1, Y_2 and U. This means we must first decode W^- and obtain $\hat{U}$ before we decode W^+. For larger n we also have to procedurally decode the channels. We will demonstrate the decoding process for $n = 8$ channels $W = BEC\left(\frac{1}{2}\right)$.

The polar coding scheme for $n = 8$ is depicted in Fig. 7. As we said, the less reliable channels will be frozen, and ’0’ will be transmitted on them constantly.

Decoding Process:

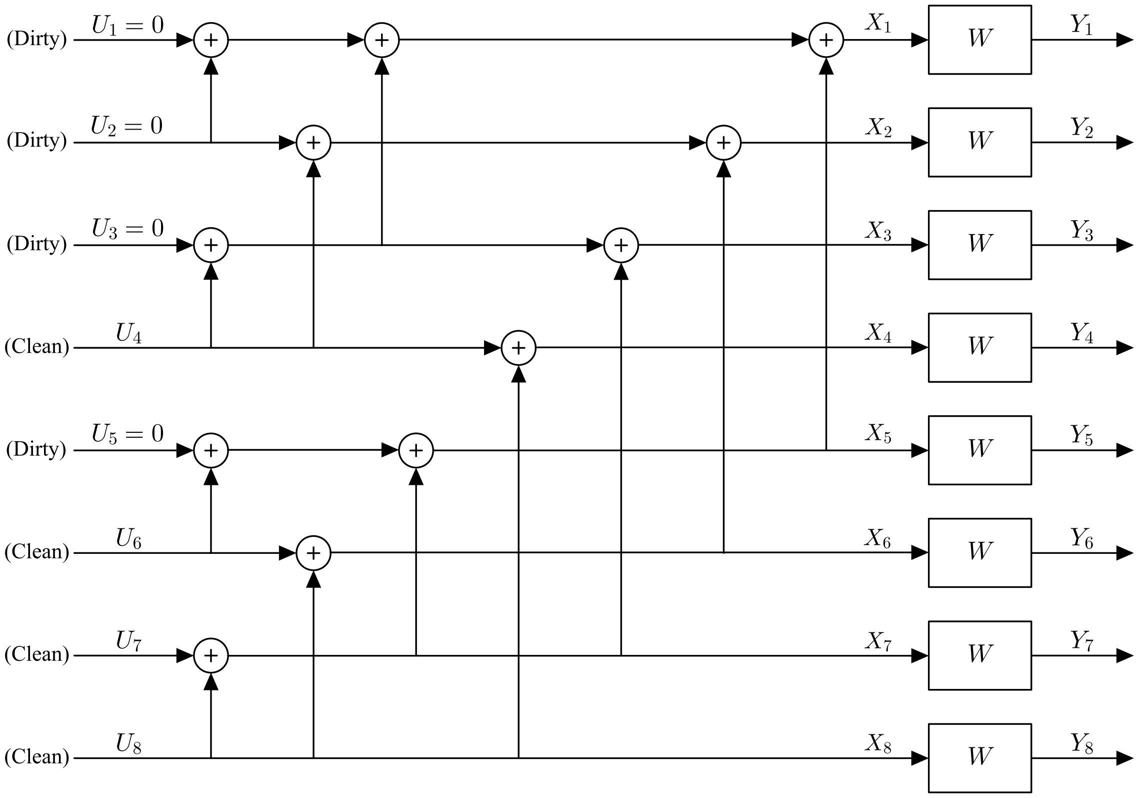

Fig. 7: Polar Coding Scheme with $n = 8$ Channels.

1) Decode U_1 (Frozen) → $U_1 = 0$.
2) Decode U_2 (Frozen) → $U_2 = 0$.
3) Decode U_3 (Frozen) → $U_3 = 0$.
4) Decode U_4 (Data!) → We use $Y_1, \ldots, Y_8$ and $U_1, \ldots, U_3$ to decode $\hat{U}_4$.
5) Decode U_5 (Frozen) → $U_5 = 0$.
6) Decode U_6 (Data!) → We use $Y_1, \ldots, Y_8$ and $U_1, \ldots, U_3, \hat{U}_4, U_5$ to decode $\hat{U}_6$.
7) Decode U_7 (Data!) → We use $Y_1, \ldots, Y_8$ and $U_1, \ldots, U_3, \hat{U}_4, U_5, \hat{U}_6$ to decode $\hat{U}_7$.
8) Decode U_8 (Data!) → We use $Y_1, \ldots, Y_8$ and $U_1, \ldots, U_3, \hat{U}_4, U_5, \hat{U}_6, \hat{U}_7$ to decode $\hat{U}_7$.

In general, we need to know the previous bits $\hat{U}_1, \ldots, \hat{U}_{k-1}$ to decode $\hat{U}_k$, so the decoding must be done sequentially.

REFERENCES

[1] Erdal Arikan. Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels. *IEEE Transactions on Information Theory*, 55(7):3051–3073, July 2009.

[2] David Tse. Ee/stats 376a: Information theory, lecture 12, 2017. Lecture notes.

[3] David Tse. Ee/stats 376a: Information theory, lecture 13, 2017. Lecture notes.