CS 37: INFORMATION THEORY
FINAL PROJECT WRITEUP: POLAR CODES

HENRY SCHEIBLE AND WARREN SHEPARD

1. Motivation and Background

Polar codes were introduced as the first channel codes to achieve the capacity of symmetric binary (in the sense that the input, but not necessarily output, is a binary alphabet) discrete memoryless channels (B-DMCs). It was previously known that capacity-achieving codes exist (by the achieveability component of Shannon's Theorem [2]) but not how to construct them. Note that we will only consider symmetric B-DMCs, so any channel mentioned here is assumed to be a symmetric B-DMCs unless otherwise stated.

The basic idea behind polar codes is as follows. Consider an arbitrary symmetric B-DMC W with capacity $C(W)$. If $C(W) \in \{0,1\}$, it is easy to construct optimal codes to send over W. Therefore, if we could "transform" any channel W into a collection of extreme channels (i.e. with $C(W) \in \{0,1\}$), then communication over the new channels will be trivial. Additionally, if the fraction of transformed channels with capacity 1 is equal to $C(W)$, then we can communicate over the transformed channels with the same capacity as W.

The main result of Arıkan's original paper [1] is a deterministic algorithm to construct polarized channels from an arbitrary channel W. Here, we focus on formalizing polarization and proving that the channels constructed by Arıkan's algorithm are polarized.

2. Polarizing two BEC's

2.1. **Construction of W_2.** Let BEC_α be the channel with $\mathscr{X} = \{0,1\}$, $\mathscr{Y} = \{0, \perp, 1\}$ and transition probabilities shown in Figure 1 for some $\alpha < 1$.

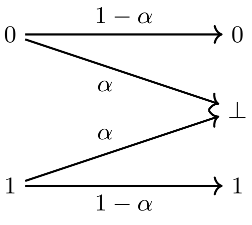


Figure 1. Transition Probabilities for BEC_α

For now, let $W = BEC_\alpha$ for some fixed α. Consider the compound channel W_2 sending a message in $\mathscr{X}^2$ to $\mathscr{Y}^2$ by applying the transformation $(U_1, U_2) \mapsto (U_1 + U_2, U_2)$, then sending each bit through W. This process is depicted in Figure 2. Denote the input by the random variables $X_1 = U_1 + U_2, X_2 = U_2$ and output by Y_1, Y_2.

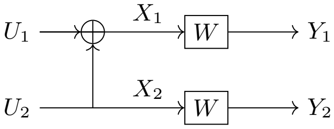


Figure 2. Compound Channel W_2

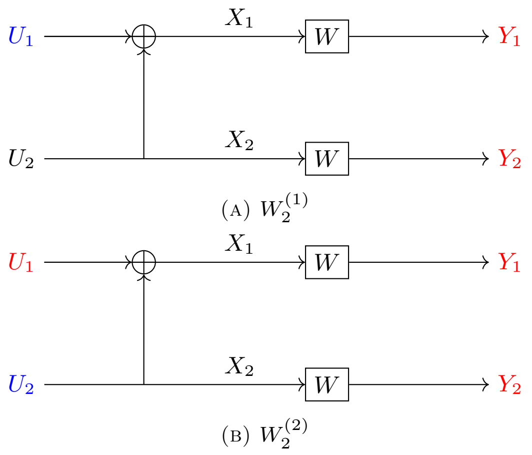
When sending a string $u = U_1U_2 \in \{0,1\}$ through W_2, are we more likely to successfully decode U_1 or U_2?

2.2. **Splitting into Virtual Channels.** To answer this question, we need to fix a decoding strategy. This strategy will later generalize to what we call *successive cancellation decoding*, see Section 5 for more details. For now, it suffices to assume that $U_1, U_2 \sim \text{Bern}(1/2)$. We will first attempt to decode U_1 knowing both Y_1 and Y_2, then if successful decode U_2 knowing Y_1, Y_2 and U_1. By considering what information the decoder has access to at each stage, we can describe effective "virtual" channels that the decoder must be able to decode. Call these virtual channels $W_2^{(1)}$ and $W_2^{(2)}$:

- $W_2^{(1)} : U_1 \mapsto Y_1Y_2$
- $W_2^{(2)} : U_2 \mapsto Y_1Y_2U_1$

The notation above indicates that $W_2^{(1)}$ is a channel which takes U_1 as an input and outputs both Y_1 and Y_2, so the decoder has access to both Y_1 and Y_2 and is trying to predict U_1. For $W_2^{(2)}$, the situation is similar. Although U_1 is obviously not actually affected by U_2, we consider a decoder which has access to Y_1, Y_2, and U_1 (because the decoder for $W_2^{(1)}$ already decoded it) and needs to choose U_2. See diagrams of $W_2^{(1)}$ and $W_2^{(2)}$ in Figure 3.

2.3. **Conservation of Capacity.** This construction has effectively split W_2 into two new channels, but this splitting operation conserves the total capacity: Using the Chain Rule for Mutual Information and the independence



(A) $W_2^{(1)}$

(B) $W_2^{(2)}$

FIGURE 3. Split Virtual Channels for W_2 (blue variables denote inputs and red variables denote outputs)

of U_1 and U_2, see that

$$\begin{aligned} C(W_2) &= I(U_1U_2 : Y_1Y_2) \\ &= I(U_1 : Y_1Y_2) + I(U_2 : Y_1Y_2|U_1) \\ &= I(U_1 : Y_1Y_2) + I(U_2 : Y_1Y_2U_1) \\ &= C\left(W_2^{(1)}\right) + C\left(W_2^{(2)}\right) \end{aligned}$$

Note that this argument applies to any symmetric B-DMC, not just to a BEC. For a BEC, however, we can explicitly calculate what both of these capacities are.

2.4. **Explicit capacities for BEC.** Now we will compute $C(W_2^{(1)})$ and $C(W_2^{(2)})$ explicitly when W is BEC_α. First, we compute the capacity of $W_2^{(1)}$. Note that $U_1U_2 \sim \text{Bern}(1/2)^{\otimes 2}$ implies that $X_1X_2 \sim \text{Bern}(1/2)^{\otimes 2}$, so $Y_1, Y_2 \sim p$, where $p(0) = 1/2(1-\alpha)$, $p(\perp) = \alpha$, and $p(1) = 1/2(1-\alpha)$.

Thus, to compute $H(U_1|Y_1Y_2)$, we need only consider two cases: If $Y_1, Y_2 \neq \perp$ (w.p. $(1-\alpha)^2$), then

$$H(U_1) = H_2(0) = 0$$

If $Y_1 = \perp$ or $Y_2 = \perp$ (w.p. $1-(1-\alpha)^2$), then

$$H(U_1) = H_2(1/2) = 1.$$

We can now use this result to compute $C(W_2^{(1)}) = I(U_1 : Y_1Y_2)$:

$$\begin{aligned} C(W_2^{(1)}) &= I(U_1 : Y_1Y_2) \\ &= H(U_1) - H(U_1|Y_1Y_2) \\ &= 1 - \sum_{(y_1,y_2)\in\{0,1,\perp\}^2} H(U_1|E_{y_1,y_2})\mathbb{P}\{E_{y_1,y_2}\} \\ &= 1 - (1-(1-\alpha)^2) = (1-\alpha)^2, \end{aligned}$$

where E_{y_1,y_2} denotes the event that $Y_1Y_2 = y_1y_2$. Similar to our analysis above, note that we can group $W_2^{(2)}$'s outputs into three events. The events $\overline{0}$ and $\overline{1}$ occur when an erasure occurs in at most one channel and the decoder knows that U_2 is equal to 0 or 1, respectively (we omit specific listings of outcomes for these events because their outcomes differ depending on the fixed value of U_1). The event $\overline{\perp}$ occurs when an erasure occurs in both channels.

To view this analysis a different way, note that you can group $W_2^{(1)}$'s outputs into three events: $\overline{0} = \{(0,0),(1,1)\}$, $\overline{1} = \{(0,1),(1,0)\}$, and $\overline{\perp} = \{(\perp,0),(\perp,1),(0,\perp),(1,\perp),(\perp,\perp)\}$. The decoder is able to correctly deduce U_1 deterministically if the output is in $\overline{0}$ or $\overline{1}$, and has gained no information about it if the output is in $\overline{\perp}$. This is very similar to the characteristics of a Binary Erasure Channel. Further, note that

$$\begin{aligned} \mathbb{P}\{\overline{\perp}\} &= 4\alpha((1-\alpha)/2) + \alpha^2 \\ &= 2\alpha - 2\alpha^2 + \alpha^2 \\ &= 2\alpha - \alpha^2 = 1-(1-\alpha)^2. \end{aligned}$$

Following the observations above, we see that $W_2^{(1)}$ is essentially $BEC_{1-(1-\alpha)^2}$, which agrees with our previous capacity calculation because

$$C(BEC_{1-(1-\alpha)^2}) = 1-(1-(1-\alpha)^2) = (1-\alpha)^2.$$

Next, we compute the capacity of $W_2^{(2)}$. Note that with U_1 fixed, X_1 and X_2 are both deterministic functions of U_2. We analyze $H(U_2)$ under two cases: If $Y_1 \neq \perp$ or $Y_2 \neq \perp$ (w.p. $1-\alpha^2$), then

$$H(U_2) = H_2(0) = 0$$

If $Y_1 = Y_2 = \perp$ (w.p. α^2), then

$$H(U_2) = H_2(1/2) = 1.$$

Thus,

$$\begin{aligned} C(W_2^{(2)}) &= I(U_2 : Y_1Y_2U_1) \\ &= H(U_2) - H(U_2|Y_1Y_2U_1) \\ &= 1 - \sum_{(y_1,y_2,u_1)\in\mathscr{A}} H(U_2|E_{y_1y_2u_1})\mathbb{P}\{E_{y_1y_2u_1}\} \\ &= 1-\alpha^2 \end{aligned}$$

Where $E_{y_1y_2u_1}$ denotes the event that $Y_1Y_2U_1 = y_1y_2u_1$ and $\mathscr{A} = \{0,1,\perp\}^2 \times \{0,1\}$.

Similar to our analysis of $W_2^{(1)}$, we can view $W_2^{(2)}$ as a BEC. We can group the outputs into the following events: $\overline{1}$ and $\overline{0}$ occur when at most one of the two copies of W's erases its input, and we know that $U_2 = 1$ or $U_2 = 0$, respectively. $\overline{\perp}$ occurs when both channels erase, meaning that we have no information about U_2. It is clear that $\mathbb{P}\{\overline{\perp}\} = \alpha^2$, so $W_2^{(2)}$ is essentially BEC_α.

To summarize, we have taken two copies of $W = BEC_\alpha$ with capacity $1-\alpha$, joined them to create W_2 with capacity $2-2\alpha$, then split W_2 back into two BEC's, this time given by $BEC_{1-(1-\alpha)^2}$ and BEC_{α^2}. These have capacities $(1-\alpha)^2$ and $1-\alpha^2$, respectively, so the conservation

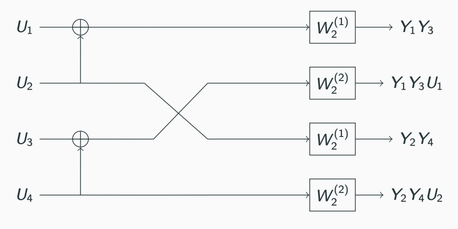


FIGURE 4. Four copies of W for which the 2nd stage transformation has been applied

property we proved earlier can be explicitly checked:

$$\begin{aligned} C(W_2^{(1)}) + C(W_2^{(2)}) &= (1-\alpha)^2 + 1 - \alpha^2 \\ &= 1 - 2\alpha + \alpha^2 + 1 - \alpha^2 \\ &= 2 - 2\alpha \\ &= C(W) + C(W). \end{aligned}$$

When generalizing this construction beyond two BEC's to two identical copies of any B-DMC, we lose the explicit formulas for capacity, but we do maintain the following two essential properties:

- $C(W_2^{(1)}) \leq C(W) \leq C(W_2^{(2)}$. This is the weakest possible meaning of the channels "polarizing". We will continue to formalize this and prove bounds in later sections
- $C(W_2^{(1)}) + C(W_2^{(2)}) = 2C(W)$. This is the conservation property we proved earlier, and our proof did not use the fact that the channels were BEC's

3. Recursive Construction

Next, we repeat the 2-channel construction devised above to four channels, a recursion which can be extended to construct W_N for any N which is a power of 2.

First, we note that we again have two pairs of identical channels (2 copies of $W_2^{(1)}$ and 2 copies of $W_2^{(2)}$) we can repeat the combining process, getting the construction in Figure 4. To view the full literal effect of this transformation, see 5, which shows that the resulting transformation is linear and given as the 2nd Kronecker power of a 2x2 matrix. In general W_N is given by the nth Kronecker power if $N = 2^n$.

4. Polarization Theorem

To formalize the idea of polarization and prove that the channels created by Arıkan's recursive construction in the previous section are polarized, consider the following theorem:

Theorem 1. [Polarization Theorem] For any channel W and $\delta \in [0,1]$, consider the channels $\left\{W_N^{(i)}\right\}$ where $N = 2^n$ is the number of channels for an integer n. As $n \to \infty$,

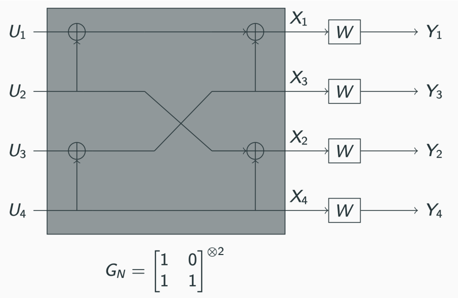


FIGURE 5. Full literal construction of W_4

(1) the fraction of indices $i \in [N]$ such that $C(W_N^{(i)}) \in (1-\delta, 1]$ approaches $C(W)$
(2) the fraction of indices i such that $C(W_N^{(i)}) \in [0, \delta)$ approaches $1 - C(W)$

where $C(W)$ denotes the *symmetric capacity* of W (same as standard capacity for BEC).

The symmetric capacity, $C(W)$, of a channel W is defined as follows:

$$C(W) = \sum_{x \in \mathscr{X}} \sum_{y \in \mathscr{Y}} \frac{1}{2} W(y \mid x) \log \frac{W(y \mid x)}{\frac{1}{2}W(y \mid 0) + \frac{1}{2}W(y \mid 1)}.$$

We held back on giving a precise equation for $C(W)$ thus far because in the case of a BEC (which is the only example we have considered), the symmetric capacity is equal to the standard capacity. Note $C(W) \in [0,1]$.

Additionally, let the *reliability*, denoted $Z(W)$, of W be defined as

$$Z(W) = \sum_{y \in \mathscr{Y}} \sqrt{W(y \mid 0) W(y \mid 1)}.$$

It is easy to see that $Z(W)$ takes on values in $[0,1]$. Additionally, a *lower* value of $Z(W)$ indicates that W is 'more reliable'. For example, in the BEC in Figure 2, if $\alpha = 0$, then $Z(W) = 0$. But, if $\alpha = 0.5$, then $Z(W) = 0.5$, and in the extreme case where $\alpha = 1$, $Z(W) = 1$. Additionally, the following upper and lower bounds on $C(W)$ in relation to $Z(W)$ can be shown:

$$(*) \qquad C(W) \leq \sqrt{1 - Z(W)^2},$$

and

$$(**) \qquad C(W) \geq \log \frac{2}{1 + Z(W)}.$$

The proof of Theorem 1 relies on two Lemmas, for which we need the following definitions. Additionally, note that the proof will be just a *sketch* where we will use several facts from the paper and about martingales without formal proof.

Definition 1. A martingale is a discrete sequence of random variables $\{X_n \mid n \geq 0\}$ such that at any time, the conditional expectation of the next value is equal to the current value), i.e.

(1) $\mathbf{Exp}[|X_n|] < \infty$
(2) $x_n = \mathbf{Exp}[X_{n+1} \mid X_1 X_2 \ldots X_n]$

Definition 2. A super martingale is a discrete sequence of random variables $\{X_n \mid n \geq 0\}$ such that at any time, the conditional expectation of the next value is less than or equal to the current value), i.e.,

(1) $\mathbf{Exp}[|X_n|] < \infty$
(2) $x_n \geq \mathbf{Exp}[X_{n+1} \mid X_1 X_2 \ldots X_n]$.

Now, consider the binary tree representation (Figure 6) of the recursive channel construction. Consider any path starting at W in the tree given by the binary sequence $b_1, b_2, b_3, \ldots$ (where a 0 represents going "up" and 1 represents going "down" in the tree) and corresponding channels $K_1, K_2, K_3, \ldots$. For any positive integer n, $W_{b_1 b_2 b_3 \ldots b_n}$ is a node in the nth layer of the tree. Additionally, let the sequences $\{C_n\} = \{C(K_n) \mid n \geq 0\}$ and $\{Z_n\} = \{Z(K_n) \mid n \geq 0\}$ be the corresponding capacity and reliability sequences.

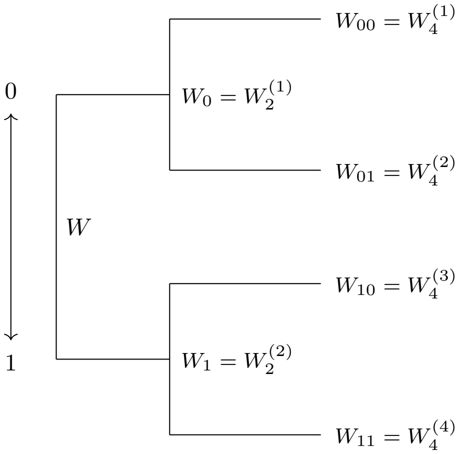


FIGURE 6. Binary Tree Representation of Channel Construction for $n = 2$

Lemma 1. $\{C_n\}$ is a martingale. Additionally, $\{C_n\}$ converges almost always to a random variable C_∞ with $\mathbf{Exp}[C_\infty] = C_0$.

Proof. Consider an arbitrary binary sequence $b_1, b_2, b_3, \ldots b_n$ and the corresponding channel $W_{b_1 b_2 b_3 \ldots b_n}$. Condition (1) is trivially true because $C(W) \in [0, 1]$. To show condition (2), consider

$$\begin{aligned}\mathbf{Exp}[C_{n+1} \mid (b_1, \ldots, b_n)] &= \frac{1}{2} C(W_{b_1 \ldots b_n 0}) + \frac{1}{2} C(W_{b_1 \ldots b_n 1}) \\ &= C(W_{b_1 \ldots b_n})\end{aligned}$$

which follows from conservation of capacity. Finally, since $\{C_n\}$ is a martingale, it follows directly as a fact about martingales that $\{C_n\}$ converges almost always to some random variable C_∞ such that $\mathbf{Exp}[C_\infty] = C_0$. □

Lemma 2. $\{Z_n\}$ is a super-martingale. Additionally, $\{Z_n\}$ converges to a random variable Z_∞ which almost always takes on value in $\{0, 1\}$.

Proof. Consider an arbitrary binary sequence $b_1, b_2, b_3, \ldots b_n$ and the corresponding channel $W_{b_1 b_2 b_3 \ldots b_n}$. Condition (1) is trivially true because $Z(W) \in [0, 1]$. To show condition (2), consider

$$\begin{aligned}\mathbf{Exp}[Z_{n+1} \mid (b_1, \ldots, b_n)] &= \frac{1}{2} Z(W_{b_1 \ldots b_n 0}) + \frac{1}{2} Z(W_{b_1 \ldots b_n 1}) \\ &\leq Z(W_{b_1 \ldots b_n})\end{aligned}$$

where the "$\leq$" follows from a fact shown in the paper. Therefore $\{Z_n\}$ is a super-martingale. It follows that Z_n converges to a random variable Z_∞ such that $\mathbf{Exp}[|Z_n - Z_\infty|] \to 0$. Note $Z_{n+1} = Z_n^2$ with probability $1/2$ (this is shown in a proposition in the paper), so

$$\mathbf{Exp}[|Z_{n+1} - Z_n|] \geq \frac{1}{2}\mathbf{Exp}[Z_n^2 - Z_n] = \frac{1}{2}\mathbf{Exp}[Z_n(1 - Z_n)] \geq 0.$$

Therefore,

$$\mathbf{Exp}[Z_n(1 - Z_n)] \to 0 \implies \mathbf{Exp}[Z_\infty(1 - Z_\infty)] \to 0.$$

For the above to hold, Z_∞ must almost always be 0 or 1. □

Proof. Using the bounds given by * and **, it follows that $C_\infty = 1 - Z_\infty$. Since $\mathbf{Exp}[C_\infty] = C_0$, we get that C_∞ converges to 1 with probability C_0 and converges to 0 with probability $1 - C_0$. The formal statement of Theorem 1 follows from the definition of convergence. □

5. Code Construction

Given a symmetric DMC W^N which is split into $W_n^{(1)}, \ldots, W_n^{(n)}$ polarized channels, we define a polar code with rate k/n by picking k channels with $I(W_n^{(i)})$ maximized. Alice and Bob agree in advance on the values to be sent on the $n - k$ remaining channels, then Alice sends her message on those k channels.

Following the theorem, as $n \to \infty$ Alice will be able to pick a capacity-achieving rate with polarized channel capacity approaching 1.

To *actually* split W^N into $W_N^{(1)}, \ldots, W_N^{(N)}$, we must use a particular decoding strategy: *successive cancellation decoding*: For each $i \in 1, \ldots, N$:

- If i is in a meaningful channel: Decode $\hat{U_i}$ based on maximum likelihood from $Y_1, \ldots, Y_N, \hat{U_1}, \ldots, \widehat{U_{i-1}}$
  - Compare $W_N^{(i)}(Y_1 \cdots Y_N U_{1\,i-1}|0)$ with $W_N^{(i)}(Y_1 \cdots Y_N U_{1\,i-1}|1)$
- If i is in a meaningless channel: set $\hat{U_i} = U_i$

6. Further Results

In addition to the main Polarization Theorem, Arıkan also made a claim regarding the *rate* of polarization:

Theorem 2. *For any Binary DMC W with $C(W) > 0$ and any fixed $R < C(W)$, there exists a sequence of sets $\mathscr{A}_N \subset \{1, \ldots, N\}, N \in \{1, 2, \ldots, 2^n, \ldots\}$, such that $|\mathscr{A}_N| \geq NR$ and $Z(W_N^{(i)}) \leq O(N^{-5/4})$ for all $i \in \mathscr{A}_n s$.*

Finally, note an interesting modern result (published in 2015). As noted earlier, our current channel polarization result relies on *successive cancellation decoding*, which is significantly weaker than *maximum likelihood decoding.* A recent advance in [3] proposes keeping L candidate strings at all times then picking the most likely one at the end of the decoding process, a decoding strategy which is between the other two options. This makes Polar Codes actually useful for small to medium block lengths, allowing them to be the highest capacity code for relevant lengths, not just asymptotically.

References

[1] Erdal Arikan. "Channel Polarization: A Method for Constructing Capacity-Achieving Codes for Symmetric Binary-Input Memoryless Channels". In: *IEEE Transactions on Information Theory* 55.7 (July 2009), pp. 3051–3073. ISSN: 1557-9654. DOI: `10.1109/tit.2009.2021379`. URL: `http://dx.doi.org/10.1109/TIT.2009.2021379`.

[2] Amit Chakrabarti. *Information Theory in Computer Science: Lecture Notes*. Latest Update: March 3, 2025. 2025. URL: `https://www.cs.dartmouth.edu/~ac/Teach/CS37-Winter25/SlidesAndNotes/lecnotes.pdf`.

[3] Ido Tal and Alexander Vardy. "List Decoding of Polar Codes". In: *IEEE Transactions on Information Theory* 61.5 (2015), pp. 2213–2226. DOI: `10.1109/TIT.2015.2410251`.