Mathematical methods in communication

Lecture 12 Part C

Lecturer: Haim Permuter *Scribe: Omer Pilosof*

I. Background

In this lecture, we will discuss compression using polar code, shifting our focus from error correction to compression using a polarization matrix. Compression with polar codes is important in communication because it enables efficient data compression for transmission over a network. This is especially useful in practical distributed compression applications, such as wireless sensor networks and distributed video coding. Polar codes have emerged as new near-capacity channel codes with great potential and have been shown to achieve distributed compression rates close to theoretical limits. This lecture is based on [1].

II. Compression using polar code

Throughout this lecture, we will utilize the following notation:

- $P \in \mathbb{F}_2^{n\times n}$ - invertible matrix.
- $Z^n = (Z_1, Z_2, ..., Z_n) \sim Bern(p)^n$ - n size vector of i.i.d $Bern(p)$.
- $W^n \triangleq Z^n \cdot P$
- $W^i = (W_1, W_2, ..., W_i)$.
- We say that W_i is highly predictable given W^{i-1} if $H(W_i|W^{i-1}) \leq \tau$ for a small τ.
- $S = S_\tau = \{i \in [1, ..., n]|H(W_i|w^{i-1}) > \tau\}$ - the set of unpredictable elements.

Before we introduce the use of polar codes for compression, we start with several properties regarding the set S. First, in the following lemma, we introduce an upper bound on the entropy of W^n.

Lemma 1 (Upper bound on $H(W^n)$) Let S be the set of unpredictable elements of W^n. Then we have that $H(W^n) \leq |S| + n \cdot \tau$.

Proof:

$$\begin{aligned} H(W^n) &\overset{(a)}{=} \sum_{i=1}^{n} H(W_i|W^{i-1}) \\ &\overset{(b)}{=} \sum_{i\in S} H(W_i|W^{i-1}) + \sum_{i\notin S} H(W_i|W^{i-1}) \\ &\overset{(c)}{\leq} |S| + \tau(n-|S|) \\ &\leq |S| + n\cdot\tau, \end{aligned}$$

where (a) follows from the chain rule, (b) follows from sum property, and (c) follows from the fact that $H(X|Y) \leq H(X)$ and $H_b(W_i) \leq 1$ (W_i is binary). ■

Moving forward, using Lemma 1, we provide in the following lemma a lower bound on the set S of unpredictable elements of W^n.

Lemma 2 (Size of S) Let $Z^n = (Z_1, Z_2, ..., Z_n) \sim Bern(p)^n$ and $P \in \mathbb{F}_2^{n\times n}$ invertible matrix. Further, let $W^n \triangleq Z^n \cdot P$ and S be the set of unpredictable elements of W^n. Then, the size of the set S is lower bounded by $|S| \geq n\cdot(H(p)-\tau)$.

Proof: Consider the following:

$$\begin{aligned} |S| + n\cdot\tau &\overset{(a)}{\geq} H(W^n) \\ |S| + n\cdot\tau &\overset{(b)}{\geq} H(Z^n) \\ |S| + n\cdot\tau &\overset{(c)}{\geq} n\cdot H_b(p), \end{aligned}$$

where (a) follows from Lemma 1, (b) follows from the fact that P is invertible, and (c) follows from the fact that z^n is distributed i.i.d $Bern(p)$. Accordingly, we obtain that

$$|S| \geq n\cdot(H_b(p)-\tau).$$

■

In conclusion, Lemma 1 provides a lower bound on the size of the set S, which contains the indices of the unpredictable elements of the polarized vector W^n. This result is derived by using the chain rule of entropy and some properties of entropy to relate the entropy of W^n to the entropy of the original data symbols Z^n and the conditional entropy of each

bit W_i given the previous bits W^{i-1}. This lemma provides a useful tool for estimating the size of S and understanding the behavior of polar codes.

Definition 1 (Polarized matrix & unpredictable columns) We say that an invertible matrix $P \in \mathbb{F}_2^{n\times n}$ is (ϵ, τ)-polarizing for $Bern(p)^n$ if for $W^n = Z^n \cdot P$ (where $Z^n \sim Bern(p)^n$) and

$$S = S_\tau = \{i \in [1, ..., n] \mid H(W_i|W^{i-1}) > \tau\}$$

we have $|S| \leq n \cdot (H(p) + \epsilon)$.

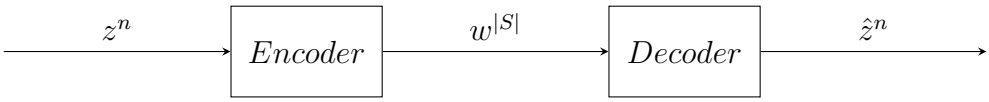
III. Polar compressor - source coding

Source coding is the process of efficiently representing information in a compact form. In the context of compression using polar codes, source coding refers to the use of polar codes to compress data symbols into a more concise representation.

The source coding process involves taking a sequence of data symbols as input and transforming them into a compressed representation using the polar code construction. The encoder identifies which bit channels are suitable for transmitting information and assigns the information bits to these good bit channels. The compressed representation generated by the encoder consists only of the information bits assigned to the good bit channels.

The goal of source coding is to create a compressed representation that is as concise as possible while still allowing for the original data to be accurately recovered by the decoder. In lossless compression, the decoder can recover the original data symbols exactly, with no loss of information. In lossy compression, some loss of information may occur, but the decoder still strives to recover the original data symbols as closely as possible.

In summary, source coding in compression using polar codes involves using polar codes to compress data symbols into a concise representation that allows for accurate recovery of the original data by the decoder.



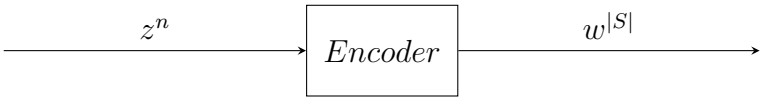
In the figure above, the source coding where z^n is an i.i.d vector $\sim Bern(p)^n$, $w^n \triangleq z^n \cdot P$, and S is equal to the set of unpredictable elements of w^n. The encoder receive z^n and outputs $w^{|S|}$ which consist of the w_i for which $i \in S$. The decoder receive $w^{|S|}$ to compute $\hat{z}^n$. Note that:

- We use $|S|$ bits over n bits message, and therefore the rate is $R = \frac{|S|}{n}$.
- At lecture 5 we have shown that $R \geq H(z^n)$.

A. Polar compressor - Encoder

In compression using polar codes, the encoder plays a crucial role. It takes a sequence of data symbols as input and transforms them into a compressed representation using the polar code construction. The encoder determines which bit channels are suitable for transmitting information and assigns the information bits to these good bit channels.

The compressed representation generated by the encoder consists solely of the information bits assigned to the good bit channels. The goal of the encoder is to create a compressed representation that is as concise as possible while still allowing for the original data to be recovered by the decoder.



In the figure above we have:

- Input: $z^n \in \mathbb{F}_2^n$, $S \subseteq [1, ..., n]$.
- Output: $w^{|S|}$ - compressed representation of z^n.

The operation of the encoder is just to multiply the input vector z^n by the matrix P and taking only w_i for which $i \in S$. Thus, the output of the encoder is $w^{|S|} = (z^n \cdot P)_{|S|}$ (where P is the polarization matrix). Our encoder reduces the number of bits from n bits to $|S|$ bits.

B. Polar compressor - Decoder

In compression using polar codes, the decoder is an essential component as well. It receives the compressed representation generated by the encoder and recovers the original data symbols using the polar code construction. The decoder identifies which bit channels are good for transmitting information and which are not. It then recovers the information bits from the good bit channels.

The decoder aims to recover the original data symbols from the compressed representation as accurately as possible. In lossless compression, the decoder can recover the original data symbols exactly, with no loss of information. In lossy compression, some loss of information may happen, but the decoder still tries to recover the original data symbols as closely as possible.

In summary, in compression using polar codes, the decoder's main role is to receive the compressed representation generated by the encoder and recover the original data symbols using the polar code construction.

The decoder operates according to the successive cancellation decompression (SCD) algorithm. The SCD algorithm works as follows:

- Input: $w^{|S|}$, $S \subseteq [1, ..., n]$, $P \in \mathbb{F}_2^{n \times n}$.
- Output: $\hat{z}^n \in \mathbb{F}_2^n$.
- Algorithm:

```
SCD(w^{|S|}, S, P)
    for i = 1 to n:
        if i ∈ S:
            ŵ_i = w_i
        else:
            ŵ_i = arg max_b {P_{W_i|W^{i-1}}(b|ŵ^{i-1})}
    return ẑ^n = ŵ^n · P^{-1}
```

Our decoder is taking the bits we sent $w^{|S|}$ and building new vector $\hat{w}^n$ of size n as follows. For indices $i \in S$ (that correspond to unpredictable bits) we set $\hat{w}_i = w_i$. Else, it means that $\hat{W}_i$ is predictable so we can predict him. After we got $\hat{w}^n$ we can compute $\hat{z}^n$ by the inverse matrix P, i.e., $\hat{z}^n = \hat{w}^n \cdot P^{-1}$

IV. THEOREM ERROR OF SOURCE CODING

In the following we introduce a theorem which provides a theoretical bound on the error rate of SCD algorithm when using polar codes for compression. The theorem states that if the polarizing matrix P is (ϵ, τ) polarizing with unpredictable columns S, then the failure probability of the SCD is $\tau \cdot n$, where n is the length of the code. This means that the probability that the original data symbols z^n cannot be recovered exactly from the compressed representation $\hat{z}^n$ is less than or equal to $\tau \cdot n$.

The theorem also states that if P is $(\epsilon, \epsilon/n)$ polarizing, then the error rate of the SCD decoder is at most ϵ. This means that by choosing an appropriate polarizing matrix P, it is possible to control the error rate of the SCD decoder and achieve a desired level of accuracy in recovering the original data symbols from the compressed representation.

In summary, this theorem provides a theoretical bound on the error rate of successive cancellation decompression when using polar codes for compression and shows how the choice of polarizing matrix can affect the accuracy of data recovery.

Theorem 1 If P is (ϵ, τ) polarizing with unpredictable columns S then the SCD has failure probability of $\tau \cdot n$, i.e.

$$\mathbb{P}_r(z^n \neq \hat{z}^n) \leq \tau \cdot n,$$

where $\hat{z}^n = SCD\Big((z^n \cdot P)_s, P, S\Big)$. Accordingly, for P that is $(\epsilon, \frac{\epsilon}{n})$ polarizing, we have that

$$\mathbb{P}_{err} \leq \epsilon.$$

Thus, if P is $(\epsilon, \frac{\epsilon}{n})$ polarizing, the error of SCD decoder is at most ϵ.

Before we can prove this theorem we will need to prove the following lemma.

Lemma 3 Let X be random variable with $H(X) \leq \alpha$. Then:

1) There exists $x \in X$ such that $\mathbb{P}_r(X \neq x) \leq \alpha$.
2) Let (X, Y) be jointly distributed random variable with $H(X|Y) \leq \alpha$, and $A(y) = \arg\max_{x \in X}(\mathbb{P}(X = x|Y = y))$. Then $\mathbb{P}_r(X \neq A(y)) \leq \alpha$.

Proof: [of part (1)]

Let $p_i \triangleq \mathbb{P}_r(X = i)$, $x \triangleq \arg\max(p_i)$, and $p_x = 1 - \gamma$ s.t. $\mathbb{P}_r(X \neq x) = \gamma$.

Test case 1 ($\gamma \leq \frac{1}{2}$)

$$\begin{aligned}
\alpha &\geq H(X) \\
&= \sum_{i=1}^{|X|} p_i \log\left(\frac{1}{p_i}\right) \\
&\overset{(a)}{\geq} \sum_{i \notin x} p_i \log\left(\frac{1}{p_i}\right) \\
&\overset{(b)}{\geq} \sum_{i \notin x} p_i \log\left(\frac{1}{\sum_{j \notin X} p_j}\right) \\
&\overset{(c)}{\geq} \gamma \log\left(\frac{1}{\gamma}\right) \\
&\geq \gamma,
\end{aligned}$$

where (a) follows from subtracting a positive number from the sum of positive numbers, (b) follows from the inequality $p_i \leq \sum_{j \notin X} p_j$ where $i \notin X$, and (c) follows from the equality $P_r(X \neq x) = \gamma$.

Test case 2 ($\gamma \geq \frac{1}{2}$)

$$\begin{aligned}
\alpha &\geq H(x) \\
&= \sum_{i=1}^{|X|} p_i \log(\frac{1}{p_i}) \\
&\overset{(a)}{\geq} \sum_{i=1}^{|X|} p_i \log(\frac{1}{p_x}) \\
&\overset{(b)}{=} \log(\frac{1}{p_x}) \geq 1 \\
&\overset{(c)}{\geq} \mathbb{P}_r(X \neq x),
\end{aligned}$$

where (a) follows from the inequality $\forall i \in [1, |X|] p_x \geq p_i$, (b) follows from the quality $\log(\frac{1}{p_x})$ is a constant and $\sum_{i=1}^{|X|} p_i = 1$, and (c) follows from the quality $\sum_{i=1}^{|X|} p_i = 1$. ■

Proof: [of part (2)]

$$\begin{aligned}
\mathbb{P}_r(X \neq A(Y)) &\overset{(a)}{=} \sum_y \mathbb{P}_r(X \neq A(Y), y) \\
&\overset{(b)}{=} \sum_y \mathbb{P}_r(y)\mathbb{P}_r(X \neq A(Y)|y) \\
&\overset{(c)}{\leq} \sum_y \mathbb{P}_r(y)H(X|Y=y) \\
&\overset{(d)}{=} H(X|Y) \\
&\leq \alpha,
\end{aligned}$$

where (a) follows from the law of total probability, (b) follows from the Lemma 2 in part (1), (c) follows from the chain rule, and (d) follows from conditional entropy law.

■

In summary, Lemma 2 provides two results for bounding probabilities of events involving random variables with bounded entropy. The first result states that if the entropy of a random variable X is less than or equal to α, then there exists a value x such that the probability that X is not equal to x is also less than or equal to α. The second result states that if X and Y are jointly distributed random variables with conditional entropy $H(X|Y)$ less than or equal to α, then the probability that X is not equal to the value that maximizes the conditional probability of X given Y is also less than or equal to α.

Proof: [of Theorem 1]

Let $W^n \triangleq Z^n \cdot P$, and assume that for every $i \in S$ we set $\hat{W}_i = W_i$. For any $i \notin S$, by applying Lemma 2 part (2) with $X = W_i$, $Y = W^{i-1}$ and $\alpha = \tau$ we get that

$$\mathbb{P}_r(W_i \neq A_i(w^{i-1})) \leq \tau. \tag{1}$$

where $A_i(\cdot)$ is

$$A_i(w^{i-1}) = \arg\max_{w_i}(\mathbb{P}(W_i = w_i | W^{i-1} = w^{i-1})).$$

By union the bound, we get that

$$\sum_{i=1}^{n} \mathbb{P}_r(W_i \neq A_i(w^{i-1})) \leq \tau \cdot n.$$

We can note that the SCD algorithm used $A_i(\cdot)$ that is equal to the $\arg\max$. If $w^n \neq \hat{w}^n$ there exists i s.t. $\hat{w}_i \neq w_i$, which gives us $\mathbb{P}_r(w^n \neq \hat{w}^n) \leq \tau \cdot n$. Accordingly, we get

$$\mathbb{P}_r(z^n \neq \hat{z}^n (= SCD(z^n \cdot P)_s, P, S)) \leq \tau \cdot n.$$

■

Example 1 (Polarized compression with two input) Let $P_2 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$ and $Z_1, Z_2 \sim Bern(p)$ where $0 < p < \frac{1}{2}$ and $w^2 \triangleq z^2 \cdot P_2$. Then, we get that

$$\begin{aligned} H(W_1, W_2) &\overset{(a)}{=} H(Z_1, Z_2) \\ &\overset{(b)}{=} H(Z_1) + H(Z_2) \\ &= 2H_b(p). \end{aligned} \tag{2}$$

$$\begin{aligned} H(W_1) &\overset{(c)}{>} H(Z_1) \\ &= H_b(p). \end{aligned} \tag{3}$$

$$\begin{aligned} H(W_2|W_1) &= H(W_1, W_2) - H(W_1) \\ &\overset{(d)}{=} 2H_b(p) - H(W_1) \\ &\overset{(e)}{<} H_b(p), \end{aligned}$$

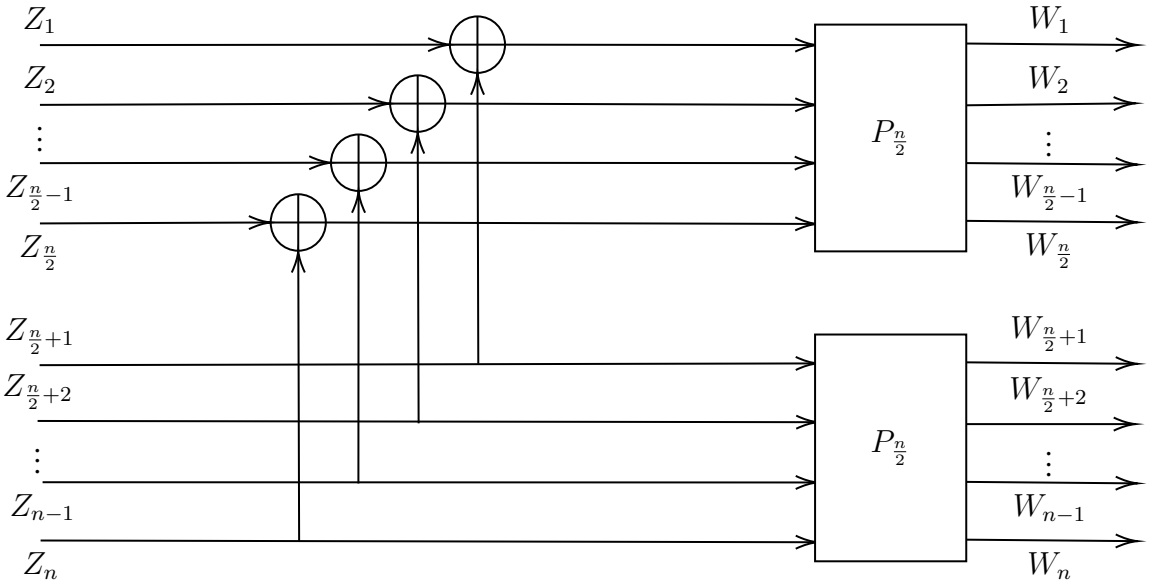
where (a) follows from $H(w^n) = H(z^n)$, (b) follows since z^2 is i.i.d, (c) follows from the inequality $\mathbb{P}_r(W_1 = 1) = 2p(1-p) > p$, (d) follows from (2), and (e) follows from (3).

Therefore, we can conclude that by multiplying z^2 by P_2, we have achieved polarization of two equally entropic bits into one bit with higher entropy and one bit with lower

entropy, as shown by the calculations of $H(W_1)$ and $H(W_2|W_1)$. This property is essential for the construction of polar codes, as explained in [1][pp.217].

Corollary 1 Polar codes are a type of error-correcting code used in communication systems. In this context, z^n represents a vector of two equally entropic bits, and P_2 is a $2x2$ matrix known as the polarization matrix. When z^n is multiplied by P_2, the resulting vector has one bit with higher entropy and one bit with lower entropy. This process is called polarization because it separates the entropy of the two input bits into one bit that is more random (higher entropy) and one bit that is more deterministic (lower entropy). This property is useful in the construction of polar codes because it allows for efficient encoding and decoding of information.

Example 2 (Polarized compression with n input) Let $P_{2^n} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}^{\otimes n}$ and $z^n = \{u, v\}$ where $u = \{Z_1, ..., Z_{\frac{n}{2}}\}$ and $v = \{Z_{\frac{n}{2}+1}, ..., Z_n\}$.



where $P_n(z^n) = P_n(u, v) = (P_{\frac{n}{2}}(u), P_{\frac{n}{2}}(v))$.

This example demonstrates how polarized compression can be applied to n input bits. This example shows how a larger polarization matrix P_2^n can be constructed by taking the Kronecker power of the 2x2 polarization matrix P_2. The input vector Z is then partitioned into two subvectors u and v, and the polarization matrix is applied to each subvector separately. The resulting polarized vector W has some bits that are more entropic and some bits that are less entropic. This example illustrates how the concept of polarization can be extended to larger input sizes and how it can be used to achieve efficient compression of data.

Theorem 2 (Strong polarization) Fix $p \in (0, \frac{1}{2})$ and constant c. There exists a polynomial function n_0 such that for every $\epsilon > 0$, there exists $n = 2^t$ with $\frac{1}{\epsilon} \leq n \leq \frac{n_0}{\epsilon}$, and a set $E \subseteq \{1, \ldots, n\}$ with $|E| \leq \frac{\epsilon}{2} \cdot n$ such that for every $i \notin E$, the conditional entropy $H(W_i|W^{i-1})$ is either less then n^{-c}, or greater then $1 - n^{-c}$. Furthermore, if we let $S = \{i \in [n]|H(W_i|W^{i-1}) \geq n^{-c}\}$ then $|S| \leq (H(p) + \epsilon)n$ and the matrix P_n is $(\epsilon, 1/n^c)$ polarizing for $BER(p)^n$ with unpredictable columns S.

The strong polarization (Theorem 2) allows us to specify how close to zero the conditional entropy of the polarized bits should be.

In conclusion, we have presented the topic of compression using polar codes, which are a novel type of error-correcting codes that have great potential for communication and compression applications. We have explained the concept of polarization, which is the key idea behind polar codes, and how it separates the entropy of equally entropic bits into more entropic and less entropic bits by using a polarization matrix. We have also shown how polar compression operates by encoding only the less entropic bits and using the SCD algorithm to decode them with low failure probability. Further, we have provided some theoretical results and proofs to demonstrate the properties and performance of polar codes.

For more information or more detailed proofs on polar codes and compression, please refer to [1].

REFERENCES

[1] V. Guruswami, A. Rudra and M. Sudan, *LaTeX: Essential Coding Theory*, 2nd ed. Cambridge, MA: MIT Press, 2022.