Effective version of the Shannon theorem: Polar codes[1]

A. Barg

0. Introduction. We discuss a constructive sequence of codes that attains capacity of binary-input symmetric memoryless channels. The main result is due to E. Arikan [1].

Let W be a binary-input discrete memoryless channel $W : \mathcal{X} \to \mathcal{Y}$, $\mathcal{X} = \{0, 1\}$. Define

$$I(W) = \frac{1}{2} \sum_{x,y} W(y|x) \log \frac{W(y|x)}{\frac{1}{2}W(y|0) + \frac{1}{2}W(y|1)}.$$

If W is (weakly) symmetric, then $I(W)$ equals its capacity, otherwise it is less than capacity[2].

Definition: Let $\mathcal{M}$ be a finite set of cardinality $M = 2^{NR}$. A mapping $f : \mathcal{M} \to \mathcal{X}^N$ defines a binary error-correcting code $C = \{f(1), f(2), \ldots, f(M)\}$. The code is called linear if f is a linear map defined on $\{0, 1\}^{NR}$.

Let $g : \mathcal{Y}^n \to \mathcal{M}$ be a decoding map. Define the error probability

$$P_e(C) = \max_{1 \le i \le M} Pr\{g(f(i)) \neq i\}.$$

A sequence of codes $C_n = f_n(\mathcal{M}_n), n \ge 1$ is said to attain the transmission rate $I(W)$ on the channel W if for any $\varepsilon > 0$ there exists a sufficiently large n_0 such that for all $n \ge n_0$ both $R > I(W) - \varepsilon$ and $P_e(C_n) \le \varepsilon$.

Below we construct a sequence of linear binary codes $C_n, n \ge 1$ of length $N = 2^n$ that attains the rate $I(W)$ of the channel W. For symmetric binary-input channels these codes are capacity-achieving.

1. Data transformation. Consider transmitting binary digits u_1, u_2 in two uses of the channel W. The combined channel can be written as $W^2(y_1^2|u_1^2)$, where $y_1^2 = (y_1, y_2)$, $u_1^2 = (u_1, u_2)$. Of course,

$$W^2(y_1^2|u_1^2) = W(y_1|u_1)W(y_2|u_2),$$

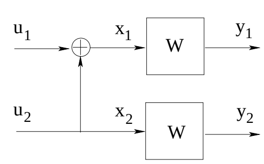
and the capacity is $I(W^2) = 2I(W)$, so nothing interesting happens. Let us transform the input bits so that the two uses of the channel, while still carrying $2I(W)$ bits, lead to outputs of unequal reliability. Toward this end, let us send

(1) $$x_1 = u_1 \oplus u_2 \text{ and } x_2 = u_2$$

in the first and the second channel uses, respectively. In other words, consider the channel

Data combining, $n = 1$

$$W_2(y_1^2|u_1^2) = W(y_1|u_1 \oplus u_2)W(y_2|u_2)$$



The capacity of W_2 is still $I(Y_1^2|U_1^2)$ because $(U_1, U_2) \leftrightarrow (X_1, X_2)$ is a one-to-one transformation. Using the chain rule we obtain

(2) $$2I(W) = I(U_1^2; Y_1^2) = I(U_1; Y_1^2) + I(U_2; Y_1^2|U_1) = I(U_1; Y_1^2) + I(U_2; Y_1^2, U_1),$$

[1]These notes were prepared for the course ENEE627 (Information Theory) in Spring semester 2012 and were taught in 3 class sessions.

[2]The above definition uses the prior distribution $P_X(0) = P_X(1) = 1/2$. This assumption is used throughout these notes.

the last step because U_1 and U_2 are independent (here $Y_1^2 = (Y_1, Y_2)$, same for U_1^2). Moreover,

$$I(U_2; Y_1^2, U_1) = H(U_2) - H(U_2|Y_1^2, U_1) \geq H(U_2) - H(U_2|Y_2) = I(W). \tag{3}$$

From (2), (3) we obtain

$$I(U_1; Y_1^2) \leq I(W) \leq I(U_2; Y_1^2, U_1). \tag{4}$$

2. Virtual channels. The mutual information quantities in (2) give rise to conditional distributions that we denote $W^+(y_1^2, u_1|u_2)$ and $W^-(y_1^2|u_1)$. They are well defined once P_U and $W(y|x)$ are defined. We will call W^+ and W^- "virtual channels" (or simply channels). Their input alphabet is $\mathcal{X} = \{0, 1\}$ and the output alphabets are $\mathcal{Y}^+ = \mathcal{Y}^2 \times \{0, 1\}$ and $\mathcal{Y}^- = \mathcal{Y} \times \{0, 1\}$, respectively.

Lemma 1. *We have*

$$W^+(y_1^2, u_1|u_2) = \frac{1}{2} W(y_1|u_1 \oplus u_2) W(y_2|u_2)$$
$$W^-(y_1^2|u_1) = \frac{1}{2} \sum_{u_2=0}^{1} W(y_1|u_1 \oplus u_2) W(y_2|u_2).$$

Proof We have

$$W^+(y_1^2, u_1|u_2) = \frac{P_{Y_1 Y_2 U_1 U_2}(y_1^2, u_1^2)}{P_U(u_2)} = 2W_2(y_1^2|u_1 \oplus u_2, u_2) P_{U_1 U_2}(u_1^2) = \frac{1}{2} W(y_1|u_1 \oplus u_2) W(y_2|u_2)$$

and

$$W^-(y_1^2|u_1) = \frac{P_{Y_1 Y_2 U}(y_1^2, u_1)}{P_U(u_1)} = 2 \sum_{u_2} \frac{1}{2} P_{Y_1 Y_2 U_1|U_2}(y_1^2, u_1|u_2) = \frac{1}{2} \sum_{u_2=0}^{1} W(y_1|u_1 \oplus u_2) W(y_2|u_2).$$

Lemma 2. $I(W^+) \geq I(W) \geq I(W^-)$ *with equality iff* $I(W)$ *equals 0 or 1.*

Proof : The first part of the claim is established in (4). By (3), equality is attained if $I(U_2; Y_1, U_1|Y_2) = H(U_2|Y_2) - H(U_2|Y_1^2, U_1) = 0$. One can show that this equality is equivalent to

$$W(y_2|0)W(y_2|1)[W(y_1|0) - W(y_1|1)] = 0,$$

i.e., either $W(y_2|0)W(y_2|1) = 0$ for all $y_2 \in \mathcal{Y}$ (capacity 1) or $W(y_1|0) = W(y_1|1)$ for all $y_1 \in \mathcal{Y}$ (capacity 0). ∎

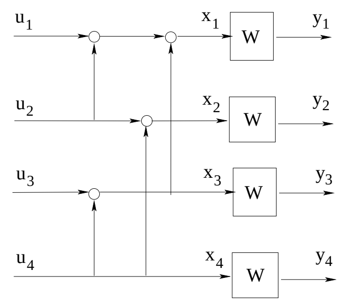
Note that $I(W^+) + I(W^-) = 2I(W)$. Therefore, if we iterate transformation (1), we can hope that some of the channels become very good, and potentially noiseless. Transformation (1) can be written as

$$(x_1, x_2) = (u_1, u_2) H_2, \quad \text{where } H_2 \triangleq \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \tag{5}$$

and the operations are modulo 2. Iterating this construction one more time, we obtain the following scheme (the circles mean addition mod 2):

Data combining, $n = 2$

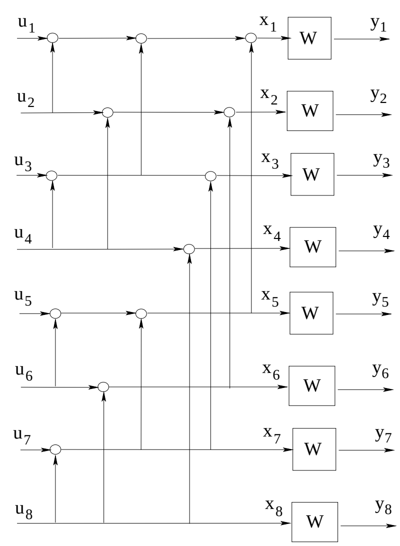
(6)
$$x_1^4 = u_1^4 \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$



One more step produces a scheme for sending u_1^8 as shown below.

Data combining, $n = 3$

(7)
$$x_1^8 = u_1^8 \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$



We denote the matrix of the linear transformation $u_1^N \to x_1^N$ by H_N and note that[3] $H_N = H_2^{\otimes n}$, where $N = 2^n$.

General setting: Let $W^N(y_1^N|x_1^N) = \prod_{i=1}^N W(y_i|x_i)$ be the N-th degree extension of the original channel W. After n iterations of the type (6)-(7), $N = 2^n$, we obtain a channel
$$W_N(y_1^N|u_1^N) \triangleq W^N(y_1^N|u_1^N H_N).$$
Now let us isolate *virtual channels* for the bits $u_1, \dots u_N$.

[3] $(H_2)^{\otimes 2} = H_2 \otimes H_2$ is defined as the 4×4 matrix of the form $\left(\begin{array}{c|c} h_{11}H_2 & h_{12}H_2 \\ \hline h_{21}H_2 & h_{22}H_2 \end{array}\right)$ where h_{ij} are the elements of H_2. Generally, $H_N = H_2 \otimes H_{N/2}$, where $N = 2^n, n \geq 2$.

Lemma 3. *Let* $u_{i+1}^N \triangleq (u_{i+1}, \dots, u_N)$, *then*

$$P(y_1^N, u_1^{i-1}|u_i) = \frac{1}{2^{N-1}} \sum_{u_{i+1}^N \in \mathcal{X}^{N-i}} W_N(y_1^N|u_1^N). \tag{8}$$

Proof : Similar to Lemma 1. We have

$$P(y_1^N, u_1^{i-1}|u_i) = \frac{P(y_1^N, u_1^i)}{P(u_i)} = 2 \sum_{u_{i+1}^N \in \mathcal{X}^{N-i}} P(y_1^N, u_1^i|u_{i+1}^N) 2^{-(N-i)}$$

(the last equality is the total probability formula)

$$= 2 \sum_{u_{i+1}^N \in \mathcal{X}^{N-i}} 2^{-(N-i)-i} P_{Y_1^N|X_1^N}(y_1^N|u_1^N) = \frac{1}{2^{N-1}} \sum_{u_{i+1}^N \in \mathcal{X}^{N-i}} W_N(y_1^N|u_1^N)$$

where on the last line we used the notation W_N introduced after (8). ∎

We also use alternative notation for these channels, which we shall now develop. Note that the channels W^+ and W^- above are binary-input, and so the operations $+$ and $-$ apply to them. In the next step of the iteration we obtain the following expressions.

Lemma 4.

$$W^{++}(y_1^4, u_1^3|u_4) = \frac{1}{2} W^+(y_1, y_3, u_1 \oplus u_2|u_3 \oplus u_4) W^+(y_2, y_4, u_2|u_4) \tag{9}$$

$$W^{+-}(y_1^4, u_1^2|u_3) = \frac{1}{2} \sum_{u_4} W^+(y_1, y_3, u_1 \oplus u_2|u_3 \oplus u_4) W^+(y_2, y_4, u_2|u_4) \tag{10}$$

$$W^{-+}(y_1^4, u_1|u_2) = \frac{1}{2} W^-(y_1, y_3|u_1 \oplus u_2) W^-(y_2, y_4|u_2) \tag{11}$$

$$W^{--}(y_1^4|u_1) = \frac{1}{2} \sum_{u_2} W^-(y_1, y_3|u_1 \oplus u_2) W^-(y_2, y_4|u_2). \tag{12}$$

Proof : For instance, let us derive the expression for $W^{++}(y_1^4, u_1^3|u_4)$. Using (6) we obtain

$$W^{++}(y_1^4, u_1^3|u_4) = \frac{1}{2} \Big(\frac{1}{2} W(y_1|u_1 \oplus u_2 \oplus u_3 \oplus u_4) W(y_3|u_3 \oplus u_4) \Big) \Big(\frac{1}{2} W(y_2|u_2 \oplus u_4) W(y_4|u_4) \Big)$$

By definition of W^+ the first of the two bracketed factors equals $W^+(y_1, y_3, u_1 \oplus u_2|u_3 \oplus u_4)$ (since u_1, u_2, u_3 are fixed) and the second gives $W^+(y_2, y_4, u_2|u_4)$. Therefore, we obtain the claimed expression (9). Likewise

$$W^{--}(y_1^4|u_1) = \frac{1}{8} \sum_{u_2} \sum_{u_4} W(y_2|u_2 \oplus u_4) W(y_4|u_4) \sum_{u_3} W(y_1|u_1 \oplus u_2 \oplus u_3 \oplus u_4) W(y_3|u_3 \oplus u_4).$$

Notice that for a fixed u_4, the sum on u_3 fits the definition of W^- in which $u_3 \oplus u_4$ is the bit value that is "averaged out." Then the probability $W^-(y_1, y_3|u_1 \oplus u_2)$ is taken outside the sum on u_4, which then becomes $W^-(y_2, y_4|u_2)$. ∎

Example: Let $W : \{0, 1\} \to \{0, 1, ?\}$ be a BEC(p) (the binary erasure channel). Then $I(W) = 1 - p$. In this case $I(W^+) = 1 - p^2$ (much better than $1 - p$) and $I(W^-) = (1 - p)^2$ (much worse). This can be computed directly and also follows from Lemma 6 below.

Suppose that we begin with BEC(0.5). Capacities of the channels evolve as follows:

	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8
$n=1$	0.25	0.75						
$n=2$	0.0625	0.4375	0.5625	0.9375				
$n=3$	0.00390625	0.121094	0.191406	**0.683594**	0.316406	**0.808594**	**0.878906**	**0.996094**

Here on the third line we list channels $W^{---}, W^{--+}, \dots, W^{+++}$. Capacities typset in bold corresponds to the channels that are "good" and that should be used to transmit 4 bits of data. Thus, in this scheme we transmit 4 bits as u_4, u_6, u_7, u_8. As $n \to \infty$, the proportion of good channels approaches $I(W) = 0.5$. Note that one shouldn't assume that the more $+$ signs the channel gets, the better it is. This may be misleading (except for the all-$(+)$ or all-$(-)$ channels). The question of locating the good indices among $N = 2^n$ indices is a separate issue.

3. Channel evolution. After n steps of the above recursion we obtain a collection of $N = 2^n$ channels

$$\mathscr{W}_n = \{W^B(\cdot\,|\,\cdot) \mid B \in \{+,-\}^n\}$$

Let us equip $\mathscr{W}_n$ with a uniform probability distribution, namely $\Pr(W^B) = \frac{1}{2^n}$ for any B. By sampling from $\mathscr{W}_n$ we obtain a "random channel" W_n. Denote its capacity by $I_n := I(W_n)$. In this part we establish convergence properties of the random process I_n.

Theorem 5. *The sequence of random variables I_n converges almost surely to a Bernoulli 0-1-valued random variable I_∞, and $P(I_\infty = 1) = I(W), P(I_\infty = 0) = 1 - I(W)$.*

This theorem implies that in the "polarization limit," the channels for bits $u_1, \dots, u_N$ become either *noiseless* (with probability $I(W)$) or fully random (with probability $1 - I(W)$). The polarization effect defines a subset $A_N(W) \subset \{1, 2, \dots, N\}$ of coordinates where the data is carried over the channel with no errors, and the number of these coordinates is $|A_N(W)| = I(W)N$. Thus, for $N \to \infty$ and any $R < I(W)$ we can transmit RN bits over W with no errors.

The previous paragraph describes the limiting behavior, i.e., the case $N = \infty$. In reality, we have $N = 2^n$, and for large n the capacity of each "good" virtual channel is close to 1, but not 1, so there will be some error rate. Nevertheless, we can transmit RN bits with low error probability, and by choosing sufficiently large n, we can make R to be arbitrarily close to $I(W)$. To prove Theorem 5 we introduce the *Bhattacharyya parameter* of the channel

$$Z(W) = \sum_{y \in \mathcal{Y}} \sqrt{W(y|0)W(y|1)}$$

For instance, if the channel is BEC(p), we obtain $Z(W) = p$ while for BSC(p) we get $Z(W) = 2\sqrt{p(1-p)}$.

We have $0 \le Z(W) \le 1$, where the left side is obvious, and the right follows by the Cauchy-Schwarz inequality[4]. If $Z(W) = 0$, then $W(y|0)W(y|1) = 0$ for all $y \in \mathcal{Y}$, so $I(W) = 1$, and if $Z(W) = 1$, then

[4]The Cauchy-Schwarz inequality states that any vectors $a, b \in \mathbb{R}^n$ satisfy

$$\sum_{i=1}^n a_i b_i \le \sqrt{\sum_i a_i^2}\sqrt{\sum_i b_i^2}$$

with equality iff $a = \alpha b$ for some $\alpha \in \mathbb{R}$. Now take $a_y = \sqrt{W(y|0)}, b_y = \sqrt{W(y|1)}, y \in \mathcal{Y}$ and use the fact that $\sum_y W(y|i) = 1, i = 0, 1$.

(from the equality condition in Cauchy-Schwarz) $W(y|0) = W(y|1), y \in \mathcal{Y}$, so $I(W) = 0$. Therefore, if $Z(W)$ is large then $I(W)$ is small, and vice versa. This is made formal in the following lemma.

Example for Theorem 5: BEC(0.5)

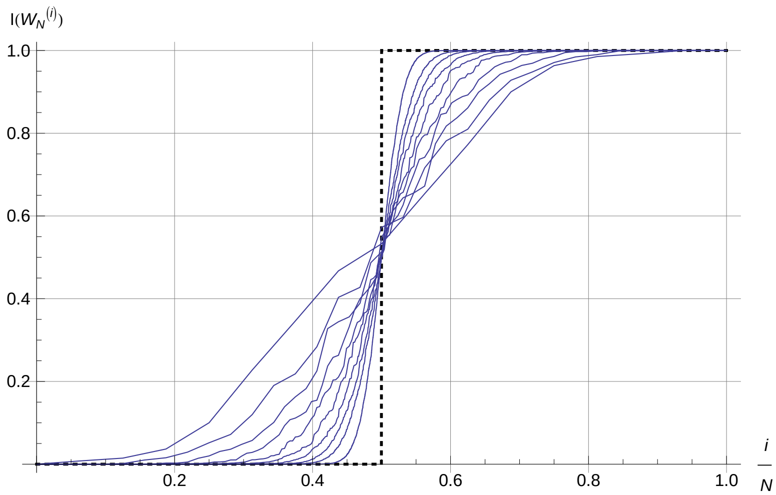


FIGURE 1. Polarization of virtual channels for the BEC case. On the x-axis we plot the relative channel index, on the y-axis the channel capacity. For a given n we compute capacities of the N channels W_N^B, sort them in increasing order, and joint the points by straight lines. The resulting curves are plotted in the figure. We show results of n iterations, $n = 4, 5, \ldots, 15$. The dotted (step) line represents the channel distribution for $n = \infty$.

Lemma 6. *For any binary-input DMC W*

$$I(W) \geq \log \frac{2}{1 + Z(W)} \tag{13}$$

$$I(W) + Z(W) \geq 1 \tag{14}$$

$$I(W)^2 + Z(W)^2 \leq 1 \tag{15}$$

Equality in (14) *holds if and only if W is a BEC*[5].

[5]A channel $W : \{0, 1\} \to \mathcal{Y}$ is called a BEC if for any output symbol $y \in \mathcal{Y}$ either $W(y|1) = W(y|0)$ or $W(y|1)W(y|0) = 0$. In particular, if $|\mathcal{Y}| = 2$, this gives the usual definition of a BEC.

Proof : Inequality (13) follows from (14), but the proof of (14) is not immediate. Let us prove (13).

$$\begin{aligned} I(X;Y) &= \sum_{y\in\mathcal{Y}} P_Y(y) \sum_{x\in\mathcal{X}} P_{X|Y}(x|y) \log \frac{P_{XY}(x,y)}{\frac{1}{2}P_Y(y)} \\ &= -2 \sum_{y\in\mathcal{Y}} P_Y(y) \sum_{x\in\mathcal{X}} P_{X|Y}(x|y) \log \sqrt{\frac{\frac{1}{2}P_Y(y)}{P_{XY}(x,y)}} \\ &\overset{\text{Jensen}}{\geq} -2 \sum_{y\in\mathcal{Y}} P_Y(y) \log \Big[\sum_{x\in\mathcal{X}} P_{X|Y}(x|y) \sqrt{\frac{\frac{1}{2}P_Y(y)}{P_{XY}(x,y)}} \Big] \\ &\overset{\text{Jensen}}{\geq} -\log \sum_{y\in\mathcal{Y}} P_Y(y) \Big[\sum_{x\in\mathcal{X}} P_{X|Y}(x|y) \sqrt{\frac{\frac{1}{2}P_Y(y)}{P_{XY}(x,y)}} \Big]^2 \end{aligned}$$

Taking the term $P_Y(y)$ inside, we obtain for the term in the brackets

$$\sum_{x\in\mathcal{X}} P_{X|Y}(x|y) \sqrt{\frac{\frac{1}{2}P_Y(y)}{P_{X|Y}(x|y)}} = \sum_x \sqrt{\frac{1}{2}P_{XY}(x,y)} = \sum_x \frac{1}{2}\sqrt{W(y|x)}$$

and

$$\sum_y \Big(\frac{1}{2} \sum_{x=0}^{1} \sqrt{W(y|x)} \Big)^2 = \frac{1}{4} \Big\{ \sum_x \sum_y W(y|x) + 2 \sum_y \sqrt{W(y|0)W(y|1)} \Big\} = \frac{1}{2}(1 + Z(W))$$

This proves (13). The proof of inequality (15) is similar, but more technically involved; see [1]. ∎

4. Martingales. We briefly recall the basic convergence results for martingales. The case that interests us is related to random processes that describe parameters of the random channels.

Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $\mathcal{F}_1$ be a σ-subalgebra of $\mathcal{F}$. Let X be $\mathcal{F}$-measurable random variable. The conditional expectation of X given $\mathcal{F}_1$ is an $\mathcal{F}_1$-measurable random variable Y such that for any $A \in \mathcal{F}_1$

$$\int_A X dP = \int_A Y dP.$$

A collection of σ-subalgebras $\mathcal{F}_n \subset \mathcal{F}, n = 1, 2, \dots$ is called a *filtration* if $\mathcal{F}_m \subseteq \mathcal{F}_n$ for all $m \leq n$. A family of random variables $X_n, n \geq 1$ is called adapted to a filtration $\mathcal{F}_n$ if X_n is $\mathcal{F}_n$-measurable for each $n \geq 1$. A family $(X_n, \mathcal{F}_n)_{n\geq 1}$ is called a *martingale* if the process X_n is adapted to the filtration $\mathcal{F}_n$, X_n is absolutely integrable for all n (i.e., $E|X_n| < \infty$), and

$$X_m \overset{\text{a.s.}}{=} E(X_n|\mathcal{F}_m) \quad \text{for } m \leq n.$$

If $=$ in this equation is replaced by $\geq$, then the sequence $(X_n, \mathcal{F}_n)_{n,\geq 1}$ is called a *supermartingale*. If you do not know what measurability and integration mean, do not worry because our use of these results will not go far beyond the following elementary example.

Example: Suppose that a fair coin is tossed 3 times, and let $\Omega = \{HHH, HHT, \dots, TTT\}$ be the set of the outcomes. Consider the set of successively refined partitions of Ω :

$$S_1 = \{H**, T**\}, \quad S_2 = \{HH*, HT*, TH*, TT*\}, \quad S_3 = \{\text{all one-element subsets}\}.$$

Here $H**$ refers to the four outcomes that start with an H, etc. These partitions define σ-algebras of subsets $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3$, and we define $\mathcal{F}_0 = \{\emptyset, \Omega\}$. Define random variables X_1, X_2, X_3, where X_i bets \$1 on the outcome of the ith toss, $i = 1, 2, 3$ (i.e., $P_{X_i}(+1) = P_{X_i}(-1) = 1/2$).

Now let $Y_i = \sum_{j=1}^{i} X_j, i = 1, 2, 3$. This sequence is adapted to the filtration $\mathcal{F}_0 = \{\emptyset, \Omega\} \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3$ (for instance, Y_2 is $\mathcal{F}_2$-measurable because it is constant on the blocks of the partition S_2). Furthermore, after the first toss, Y_1 is known, and $E(Y_2|\mathcal{F}_1) = Y_1 + (1/2(+1) + 1/2(-1)) = Y_1$. Thus, the sequence $(Y_i, \mathcal{F}_i), i = 1, 2, 3$ forms a martingale.

This example highlights the idea behind the notion of the martingale which was conceived as an abstraction of a fair game. Suppose that we gamble by making bets on the outcomes of an experiment. Knowing the outcome of the first toss does not improve or hinder our chances to win, namely, $E(X_2 - X_1|\mathcal{F}_1) = 0$. Supermartingales model games that are apriori unfavorable since in this case $E(X_2 - X_1|\mathcal{F}_1) \le 0$.

The main result about martingales is given by the following theorem.

Theorem 7. *(Doob) Let $(X_n, \mathcal{F}_n)_{n\ge 1}$ be a bounded (super)martingale (i.e., $|X_n| < c$ for some constant c and all n). Then*

$$\lim_{n\to\infty} X_n = Y$$

almost surely, where Y is a random variable. Moreover, EY exists, and $E|X_n - Y| \to 0$.

5. Completing the proof of Theorem 5.

Lemma 8.

$$I(W^+) + I(W^-) = 2I(W) \tag{16}$$

$$Z(W^+) = Z(W)^2 \tag{17}$$

$$Z(W) \le Z(W^-) \le 2Z(W) - Z(W)^2 \tag{18}$$

Equality on the right-hand side of (18) *is attained if and only if W is a BEC.*

Proof: (16) was proved in (2). Relations (17) and (18) are proved by a direct calculation. For instance, let us prove (17). We have

$$\begin{aligned}
Z(W^+) &= \sum_{y_1^2, u_1} \sqrt{W^+(y_1^2, u_1|0)W^+(y_1^2, u_1|1)} \\
&= \sum_{y_1^2, u_1} \frac{1}{2}\sqrt{W(y_1|u_1)W(y_2|0)W(y_1|u_1 \oplus 1)W(y_2|1)} \\
&= \sum_{u_1} \frac{1}{2} \sum_{y_2} \sqrt{W(y_1|u_1)W(y_2|0)} \sum_{y_1} \sqrt{W(y_1|u_1 \oplus 1)W(y_2|1)} \\
&= Z(W)^2. \qquad \blacksquare
\end{aligned}$$

Remark: As a consequence of this lemma, we also have

$$\begin{gathered}
Z(W^+) + Z(W^-) \le 2Z(W) \\
I(W) \le I(W^+) \le 2I(W) - I(W)^2 \\
I(W)^2 \le I(W^-) \le I(W)
\end{gathered}$$

These relations are not used below.

Let us tie the evolution of channels to the context of the previous section. Let $\Omega = \{\omega | \omega \in \{+, -\}^*\}$ be the set of semi-infinite binary sequences. We may view Ω as a rooted binary tree where the nodes of the nth level corresponds to the channels of the form $W^b, b = (b_1, \dots, b_n)$, where $b_i \in \{+, -\}$ for all i. Define a set of increasingly refined partitions of Ω into subsets of the form $S(b_1, \dots, b_n) = \{\omega \in \Omega \mid \omega_1 = b_1, \dots, \omega_n =$

$b_n\}, n \geq 0$. Put $P(S(b_1, \dots, b_n)) = 2^{-n}$. Put $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and define a filtration $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}$, where $\mathcal{F}_n, n \geq 1$ is generated by the sets $S(b_1, \dots, b_n)$.

Let $B_i, i = 1, 2, \dots$ be i.i.d. $\{+, -\}$-valued random variables with $P(B_1 = +) = P(B_1 = -) = 1/2$. The random channel emerging at time n will be denoted by W^B, where $B = (B_1, B_2, \dots, B_n)$, Thus, $P(W^B) = 2^{-n}$ for all realizations of B. Let $W_n = W^B$ and let $I_n = I(W^B), Z_n = Z(W^B)$ be random processes. The sequences I_n, Z_n are adapted to the above filtration. Note that this setting formalizes the discussion that preceded Theorem 5.

Proposition 9. *The sequence $(I_n, \mathcal{F}_n)_{n \geq 0}$ forms a bounded martingale. The sequence $(Z_n, \mathcal{F}_n)_{n \geq 0}$ forms a bounded supermartingale.*

Proof : We have

$$E(I_{n+1}|\mathcal{F}_n) = \frac{1}{2}(I(W^{B_1,\dots,B_n,+}) + I(W^{B_1,\dots,B_n,-})) = I_n,$$

where the second equality follows from (16). Therefore $(I_n, \mathcal{F}_n)_{n \geq 1}$ is a martingale. It is bounded because $I_n \in [0, 1]$ for all n.

Similarly,

$$E(Z_{n+1}|\mathcal{F}_n) = \frac{1}{2}(Z(W^{B_1,\dots,B_n,+}) + Z(W^{B_1,\dots,B_n,-})) \leq Z_n,$$

where the inequality follows from (17)-(18). Therefore $(I_n, \mathcal{F}_n)_{n \geq 1}$ is a supermartingale. It is bounded because $Z_n \in [0, 1]$ for all n. ∎

Now let us complete the proof of Theorem 5. By Doob's theorem, the sequence Z_n converges a.s. to a random variable Z_∞. A refinement of this theorem implies that $E|Z_n - Z_\infty| \to 0$, and therefore, $E|Z_n - Z_{n+1}| \to 0$. However, $Z_{n+1} = Z_n^2$ with probability $1/2$, so $E|Z_{n+1} - Z_n| \geq 1/2 E(Z_n(1 - Z_n)) \geq 0$. Thus, $\lim_{n \to \infty} E(Z_n(1 - Z_n)) = 0$, and so $E(Z_\infty(1 - Z_\infty)) = 0$. This implies that $Z_\infty = 0$ or 1 a.s.[6]

Again using Doob's theorem, we claim that $I_n \overset{\text{a.s.}}{\to} I_\infty$ and $EI_\infty = I_0$. But (13) and (15) imply that $I_\infty \in \{0, 1\}$ a.s., and so $P(I_\infty = 1) = I(W)$. This completes the proof of Theorem 5.

6. Code construction. We have shown that by iterating the basic data transformation we can transmit close to an $NI(W)$ proportion of bits almost noiselessly. Let us turn this observation into a code construction.

Let $H_N = H_2^{\otimes n}$ be the $N \times N$ matrix of the form (6), (7). Let A_N be the set of indices that correspond to channels of capacity close to 1. Let G_N be an $NI(W) \times N$ formed of the rows with indices in A_N.

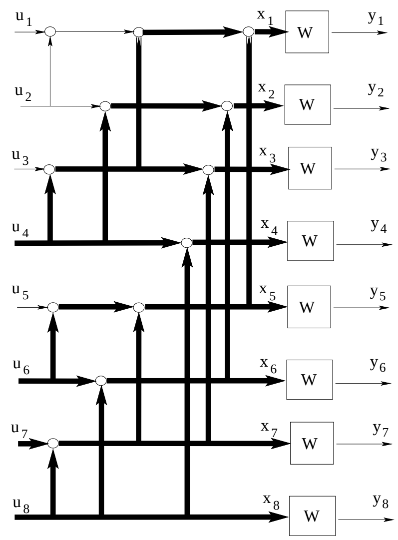
Definition: A *polar code* is a linear map $f : \{0, 1\}^{NR} \to \{0, 1\}^N$ given by $u \mapsto x = uG_N$.

We represent messages as binary strings of nR bits. A message u_1^{RN} is encoded as x_1^N and transmitted in N uses of the (physical) channel W.

Example: To continue with the example of BEC with $p = 0.5$ (see p.4), we encode 4 bits, call them u_4, u_6, u_7, u_8 using the scheme in (7). Namely, in the following figure

---

[6]This means that there exist disjoint subsets $\Omega_0, \Omega_1 \subset \Omega$ such that $P(\Omega_0 \cup \Omega_1) = 1$ and $\lim_{n \to \infty} Z_n(\omega) = i$ for $\omega \in \Omega_i, i = 0, 1$.



we set the bits u_1, u_2, u_3, u_5 to zero (or any other set of values known to both the sender and the receiver) and send the bits $x_1, \dots, x_8$ over the channel.

The decoder mapping utilizes the distributions (8) associated with the virtual channels $W(y_1^N, u_1^{i-1}|u_i), i = 1, \dots, N$ which assume that we decode N bits one by one. Since in reality we only have RN bits, the remaining $N(1-R)$ bits are set to 0 by the decoder.

The *successive cancellation decoder* is defined inductively as follows: for $i = 1, 2, \dots, N$ put

$$\hat{u}_i = \begin{cases} \arg\max_{z\in\{0,1\}} W(y_1^N, \hat{u}_1^{i-1}|z) & \text{if } i \in A_N \\ 0 & \text{if } i \in A_N^c. \end{cases}$$

The decoder takes the best decision for the ith bit based on the estimate of the bits 1 to $i-1$. Note that some of the "future" indices may be in the set A_N^c so the corresponding bits are known to the decoder which could use this additional information. The above definition disregards this knowledge. resulting in an easily implementable procedure.

Theorem 10. *Let $P_N(W) = \Pr(\hat{u}_1^N \neq u_1^N)$ be the error probability of decoding for the length-N polar code. Then*

$$P_N(W) \le \sum_{i\in A_N} Z(W_M^{(i)}).$$

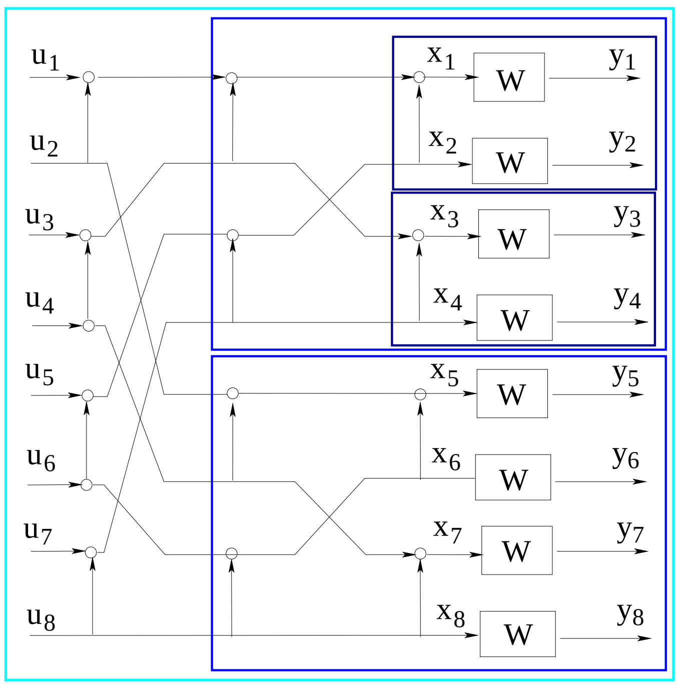
Proof: see homework 5.

An additional argument shows that $P_N(W) \le O(2^{-N^\beta})$, where β is any number in $(0, 1/2)$.

Remark: One issue about the overall proof remains unresolved. Namely, we have been assuming that all the rv's $U_i, 1 \le i \le N$ are uniform $\{0,1\}$-valued. However, above the bits that correspond to very noisy channels have been set to 0, violating this assumption. An averaging argument in [1] shows that there exists an assignment of bits such that P_N is bounded above as in the last theorem. Moreover, for symmetric channels any assignment of bits is as good as any other asignment. This shows that setting $U_i = 0, i \in A_n^c$ does not interfere with the proof.

7. Conclusion. This concludes the proof of the fact that polar codes with successive cancellation decoding achieve capacity of binary-input symmetric channels. The complexity of encoding and decoding with polar codes can be shown to grow as $O(N \log N)$, where N is the code length. Other known results for polar codes include estimates of the error probability of decoding, extension of the above arguments to nonbinary alphabets $\mathcal{X}$, using polarization to achieve limits of noisy data compression (source coding), algorithms for finding the set of good indices A_N, and replacing the basic transform H_2 with matrices of larger dimensions, which gives faster decline of the error rate as a function of N.

Finally, we remark that the code construction of [1] relies on matrices that are slightly different from $H_N = H_2^{\otimes n}$. The additional layer results in easier implementation of the codes, while the approach in these notes is more convenient for classroom use. The basic convergence results are not affected by this change. At the same time, some of the formulas in these notes, notably (9)-(12), do not match the corresponding results in [1]. The data combining proposed there (shown below for our running example) permits a recursive implementation similar to the Cooley-Tuckey algorithm for fast DFT and leads to the complexity estimates mentioned in the previous paragraph.



References

[1] E. Arıkan, Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels. IEEE Transactions on Information Theory, vol. 55, number 7, 2009, pp.3051–3073. Download from http://arxiv.org/abs/0807.3917 .