Technical University of Crete

Department of Electronic and Computer Engineering

Design and Decoding of Polar Codes

By:

Nikolaos Tsagkarakis

Submitted on July 26th, 2011 in partial fulfillment of the requirements for the Electronic and Computer Engineering diploma degree.

Supervisor:	Assistant Professor George Karystinos
committee:	Professor Athanasios Liavas
	Assistant Professor Aggelos Bletsas

Abstract

In this work we explain and analyze a method, known as channel polarization, to construct block codes that achieve the symmetric capacity of any binary-input discrete memoryless channel (B-DMC). The symmetric capacity is the highest rate achievable subject to uniformity in input's distribution. By channel polarization, it is possible to synthesize, out of N independent copies of a given B-DMC W, a second set of N binary-input channels $\{W_N^{(i)} : 1 \leq i \leq N\}$. The codes generated that way are called Polar Codes because of the tendency of channels $W_N^{(i)}$ $\forall i$ to have polarized symmetric capacities, i.e. their symmetric capacities are either 1 (ideal channel) or 0 (useless channel). The proof of their capacity achieving property is also given. Finally, polar codes are compared with a family of codes that are very closely related; the Reed-Muller codes.

Περίληψη

Στο παρόν κείμενο εξηγείται και αναλύεται μία μέθοδος, γνωστή με το όνομα πόλωση καναλιού (channel polarization), για να κατασκευή κωδίκων block που επιτυγχάνουν τη συμμετρική χωρητικότητα οποιουδήποτε διακριτού καναλιού χωρίς μνήμη με δυαδική είσοδο (B-DMC). Η συμμετρική χωρητικότητα είναι ο μέγιστος δυνατός ρυθμός αν θέσουμε ομοιόμορφη κατανομή εισόδου. Εφαρμόζοντας τη μέθοδο πόλωσης καναλιού, μπορούμε να συνθέσουμε, με τη χρήση N ανεξάρτητων αντιγράφων ενός B-DMC W, ένα δεύτερο σύνολο από N κανάλια με δυαδική είσοδο $\{W_N^{(i)} : 1 \leq i \leq N\}$. Οι κώδικες που παράγονται έτσι καλούνται Πολικοί Κώδικες λόγω της τάσης των καναλιών $W_N^{(i)} \forall i$ να έχουν πολωμένη συμμετρική χωρητικότητα, δηλαδή, οι συμμετρικές χωρητικότητες τους είναι είτε 1 (ιδεατό κανάλι) ή 0 (άχρηστο κανάλι). Επίσης, δίνεται η απόδειξη ότι η χωρητικότητα επιτυγχάνεται. Τέλος, οι πολικοί κώδικες συγκρίνονται με μία άλλη οικογένεια κωδίκων πολύ στενά συνδεδεμένη, τους κώδικες Reed-Muller.

Contents

1 Introduction **9**

2 Polarization **9**
2.1 Symmetric Capacity 9
2.2 Bhattacharyya Parameter 10
2.3 The Polarization Step 11
2.4 Effective Channel 13

3 Capacity Achievability **14**

4 Information Set Selection **18**
4.1 Reed-Muller Codes 18
4.2 Polar Codes 18

5 Tx and Rx **19**
5.1 Encoder-Decoder 19
5.2 Code Construction 20

A Performance Results **22**

List of Figures

1 BEC(ϵ). 11
2 W_2. 11
3 W_N. 12
4 Linear Code. 12
5 $W_N^{(i)}$. 13
6 $Z(W_{32}^{(i)}) \forall i$ and for any BEC with erasure probability $\epsilon \in [0,1]$ 15
7 $Z(W_N^{(i)}) \forall i$ and for BSC with crossover probability in $[0,1]$ 15
8 The tree process for the recursive channel construction. 16
9 Distribution of random variable $Z = Z(W_{2^n}^{(i)}), i = 1, \ldots, 2^n$ when n = 15 (top), and 24 (bottom), and W BEC(0.4) 17
10 Decoder's scheme 20
11 Graphical representation of G_8. 21
12 The corresponding tanner tree graph for i^{th} bit. 22
13 Polar (SC), various block lengths. 23
14 Polar (SC) vs LDPC. 23
15 Polar (BP) vs LDPC. 24

1 Introduction

This thesis presents Arikan's polar codes published in 2007 [1] based on an idea that in essence offers to the community a way to build capacity achieving codes with relatively low cost.

To take the issue from the beginning, Claude Shannon introduced the concept of channel coding in 1948 through a breaking paper of the time [6]. In this context, he defined new quantities like mutual information and channel capacity which constitute the tools for proving his theorems. One of them is the following: "*If the correction channel has a capacity equal to C it is possible to so encode the correction data as to send it over this channel and correct all but an arbitrarily small fraction ϵ of the errors. This is not possible if the channel capacity is less than C.*" [6, Theorem 10]. Relying on this theorem, communication scientists have ever since been trying to actually code the data with such rate. LDPC [7] and Turbo [8] codes are the best codes that are close in achieving this goal; however they lack a mathematical proof of their properties (with the exception of LDPC when transmitting over a binary erasure channel (BEC)). Polar codes are the first provably capacity-achieving codes for any channel of the class of binary discrete memoryless channels (B-DMC's) with tractable complexity.

Of course, there exist Shannon's nonlinear and Elias' linear random block codes that, in the mean, achieve channel capacity. Unfortunately, the complexity of encoding and decoding operations is too high since they do not follow a certain pattern. More specifically, Shannon's codes consist of codewords with coefficients randomly selected. The distribution of the coefficients is uniform and each one is selected independently. As it is obvious, the complexity is exponential with respect to codeword length N. Elias proposed a method for constructing linear codes by building the generator matrix randomly. Accordingly, he chooses the coefficients of the matrix independently and from the uniform distribution. Again the complexity of the decoder and encoder is exponential and quadratic, respectively.

We begin with some necessary definitions and preliminaries. What is a code? Consider a countable alphabet $\mathcal{X}$. All the words that can be produced by k symbols in this alphabet are $|\mathcal{X}|^k$. Any of these k-tuples could be matched with another tuple of $n \geq k$ symbols of the same alphabet. The very set of the n-tuples is a code. Our study specializes of the binary alphabet ($\mathcal{X} \doteq \{0,1\}$).

If the code is well designed, then the initial information is protected in terms of symbol (bit) error probability between the transmitter ($\mathrm{T_x}$) and the receiver ($\mathrm{R_x}$). The ratio ${}^{k}/_{n}$ is called the rate of the code. The larger the rate is the faster Tx and Rx can communicate. The maximum achievable rate subject to arbitrarily low bit detection error probability is called Shannon Capacity. A typical example of a code that does not achieve the capacity is the repetition code. Suppose we want to sent a BPAM symbol ($\pm A$) over an AWGN(σ). The probability that the optimal receiver doesn't decides the bit correctly is $Q\left(\sqrt{\frac{A^2}{\mathrm{W}_0\sigma^2}}\right)$, where W_0 is the used bandwidth. We can "code" this bit with the repetition code by repeating it N times. This will reduce the probability of error to $Q\left(\sqrt{\frac{NA^2}{\mathrm{W}_0\sigma^2}}\right) \overbrace{=}^{R={}^{1}/_{N}} Q\left(\sqrt{\frac{A^2}{R\mathrm{W}_0\sigma^2}}\right)$. This type of coding is equivalent to multiplying the symbol period with N or using N times less bandwidth. The trade off between bandwidth and bit error probability is unprofitable. In order to make no mistakes ($Prob\{bit\ error\}=0$) we push the rate (R) to zero!

An important property of channel coding is that as N gets bigger the size of the code grows exponentially. Such a fact may make a simple mapping impossible for large values of N in terms of memory efficiency.

2 Polarization

2.1 Symmetric Capacity

We consider a discrete memoryless channel with input alphabet $\mathcal{X} = \{0,1\}$ and output alphabet $\mathcal{Y}$. Then, channel polarization is a strategy to achieve the symmetric capacity of the channel, defined as

$$I(W) \doteq \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} W(y|x) \log_2 \frac{W(y|x)}{\frac{1}{2}W(y|0) + \frac{1}{2}W(y|1)} \tag{1}$$

where W denotes a B-DMC. The symmetric capacity of the channel is equal to the Shannon capacity if we enforce the input distribution to be uniform ($P(X=0)=P(X=1)=\frac{1}{2}$ where X is the random variable modeling channel input). Therefore, if the channel is symmetric (for which we know that the optimal input distribution is the uniform), then polar codes achieve Shannon capacity.

2.2 Bhattacharyya Parameter

In addition to symmetric capacity which is a measure of the maximum possible reliable transmission rate through channel with uniform input, we also introduce the Bhattacharyya parameter of a B-DMC W which is defined as

$$Z(W) \doteq \sum_{y \in \mathcal{Y}} \sqrt{W(y|0)W(y|1)}, \tag{2}$$

measures the reliability of the channel, equals the sum over all output combinations of those geometric means, and constitutes an upper bound on the error probability of an uncoded bit transmission. It will also prove to be useful in deriving an upper bound of the block error probability.

Symmetric capacity and Bhattacharyya parameter are related by the formulas

$$I(W) \geq \log_2 \frac{2}{1+Z(W)}, \tag{3}$$

$$I(W) \leq \sqrt{1-Z(W)^2}. \tag{4}$$

The proof that the above formulas hold true is omitted.
Proposition 1: The first corollary that can be driven by (3) and (4) is that for any W with $I(W)=1$ or 0 we have that $Z(W)=0$ or 1, respectively, and vice versa.

In order to see how Bhattacharyya parameter can pose an upper bound to bit error probability under maximum-likelihood (ML) detection, we define the error event

$$\begin{aligned}\mathcal{E} &\doteq \left\{(x, y_1^N) \in \mathcal{X} \times \mathcal{Y}^N : W(y_1^N|x \oplus 1) \geq W(y_1^N|x)\right\} \\ &= \left\{(x, y_1^N) \in \mathcal{X} \times \mathcal{Y}^N : \frac{W(y_1^N|x \oplus 1)}{W(y_1^N|x)} \geq 1\right\}\end{aligned}$$

for transitions over a discrete channel W with a single input and multiple outputs. The probability of event $\mathcal{E}$ is

$$\begin{aligned}P(\mathcal{E}) &\doteq \sum_{y_1^N \in \mathcal{Y}^N} \sum_{x \in \mathcal{X}} P_{X,Y_1^N}(x, y_1^N) \mathbf{1}_{\mathcal{E}}(x, y_1^N) \\ &= \sum_{y_1^N \in \mathcal{Y}^N} \sum_{x \in \mathcal{X}} P_X(x) P_{Y_1^N|X=x}(y_1^N) \mathbf{1}_{\mathcal{E}}(x, y_1^N) \\ &= \sum_{y_1^N \in \mathcal{Y}^N} \sum_{x \in \mathcal{X}} \frac{1}{2} W(y_1^N|x) \mathbf{1}_{\mathcal{E}}(x, y_1^N) \\ &\leq \sum_{y_1^N \in \mathcal{Y}^N} \sum_{x \in \mathcal{X}} \frac{1}{2} W(y_1^N|x) \sqrt{\frac{W(y_1^N|x \oplus 1)}{W(y_1^N|x)}} \\ &= \sum_{y_1^N \in \mathcal{Y}^N} \sum_{x \in \mathcal{X}} \frac{1}{2} \sqrt{W(y_1^N|x) W(y_1^N|x \oplus 1)} \\ &= \sum_{y_1^N \in \mathcal{Y}^N} \frac{1}{2} \sqrt{W(y_1^N|0) W(y_1^N|1)} + \frac{1}{2} \sqrt{W(y_1^N|1) W(y_1^N|0)} \\ &= \sum_{y_1^N \in \mathcal{Y}^N} \sqrt{W(y_1^N|0) W(y_1^N|1)} \\ &= Z(W).\end{aligned}$$

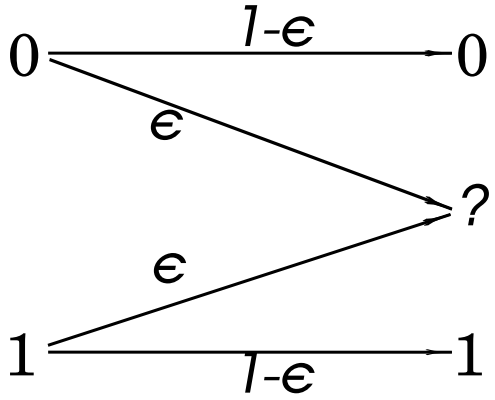

Figure 1: BEC(ϵ).

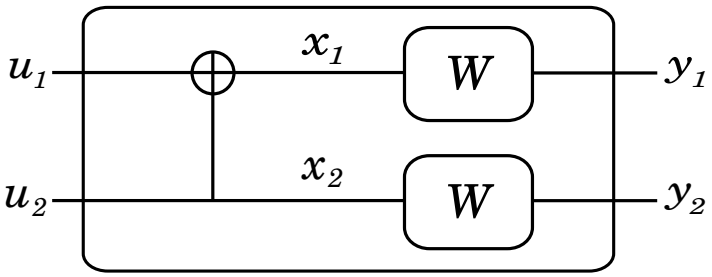

Figure 2: W_2.

Apparently, this is a result that applies to SISO (single-input single-output) channels as well. We will show how we can use Bhattacharyya parameter to set an upper bound on block error probability later on.

As an example, we consider the BEC (see Fig. 1). Once a bit is transmitted, the receiver will obtain either the bit correctly or a symbol that does not contain any information. The BEC does not introduce incorrect information. The mutual information between input and output is expressed as $I(W) = 1 - \epsilon$. This is a convenient form and linear to the channel's only parameter.

The Bhattacharyya parameter of a BEC(ϵ) turns out to be even simpler:

$$\begin{aligned} Z(W) &= \sum_{y \in \mathcal{Y}} \sqrt{W(y|0)W(y|1)} \\ &= \sqrt{W(0|0)W(0|1)} + \sqrt{W(?|0)W(?|1)} + \sqrt{W(1|0)W(1|1)} \\ &= \sqrt{(1-\epsilon) \times 0} + \sqrt{\epsilon\epsilon} + \sqrt{0 \times (1-\epsilon)} \\ &= \epsilon. \end{aligned}$$

2.3 The Polarization Step

Given N independent copies of one B-DMC W, the idea behind channel polarization is to synthesize a new set of B-DMC's $\{W_N^{(i)} : 1 \leq i \leq N\}$ where some of them have absolute reliability.

From the implementation's point of view the channel polarization process consists of many levels. The root of this multilevel recursive operation is realized by the following simple scheme. We set x_1 (the bit transmitted through the top copy of W) to be a function of both u_1 and u_2 (sum mod-2 a.k.a. XOR) while x_2 equals u_2. By combining a pair of copies of W, we construct a new composite channel denoted as W_2 with two bits as input and two bits as output, as shown in Fig. 2.

At the random i^{th} level of polarization we combine two channels $W_{\frac{N}{2}}$, as it is illustrated by Fig. 3. For an array of inputs $\{u_i\}_{i=1}^N$ we obtain a new one $\left(\{v_i\}_{i=1}^N\right)$ by replacing the odd indexed u's

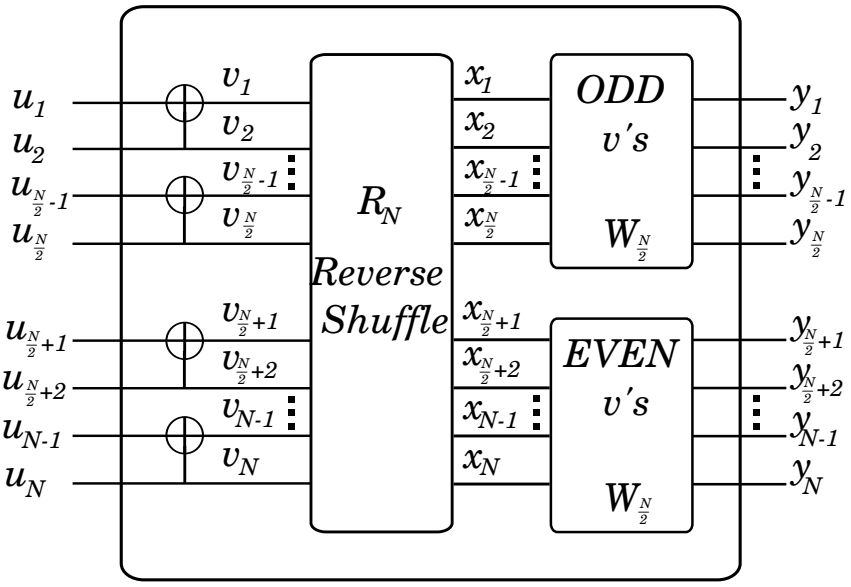

Figure 3: W_N.

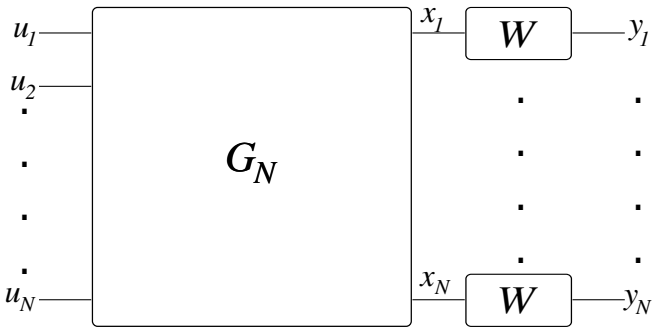

Figure 4: Linear Code.

with the XOR of consecutive pairs. That is, the odd indexed v_{2i-1} will be equal to $u_i \oplus u_{i+1}$ for each $i = 1, \ldots, \frac{N}{2}$. The rest u's (that are even indexed) are passed to the reverse shuffle operator as they are. Reverse shuffling distincts the odd and even indexed elements and passes them to separate channels $W_{\frac{N}{2}}$; one channel for the odd indexed u's (top) and one for the even indexed u's (bottom).

Regarding the reverse shuffle, it is interesting to note that grouping the odd and the even indexed inputs is a permutation that can be further comprehended as re-ordering the tuple of bits that is associated with each input. This tuple is the binary expression of $i - 1$. For instance, the first input (v_1) is associated with the all-zero tuple of length $n = \log_2 N$, that is $v_1 \rightarrow v_{00\ldots0}$. Accordingly, the last input is associated with the all-one tuple, that is $v_N \rightarrow v_{11\ldots1}$. In general, $v_{b_1 b_2 \ldots b_n} \leftarrow v_i$ where $i = 1 + \sum_{j=1}^{n} b_j 2^{n-j}$. Hence, reverse shuffle changes the order of this bit string as follows: $v_{b_1 b_2 \ldots b_n} = u_{b_2 b_3 \ldots b_n b_1}$. In other words, R_N cyclically right-shifts by one the bit-indeces of the elements of a left operand u_1^N.

The essential transformation after $\log_2(N)$ levels is linear. By this, we mean that before transmitting the information word u_1^N we first apply a linear transformation by multiplying it (from the right) with a matrix denoted as G_N. Each entry of the resulting codeword $x_1^N = u_1^N G_N$ is then transmitted independently though each copy of W.

As we can see, transmitting a word u_1^N through the composite channel W_N is equivalent to transmitting x_1^N through N independent copies of channel W. An analysis via linear algebra results in $G_N = B_N F^{\otimes n}$ where B_N is a permutation matrix and

$$F = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}.$$

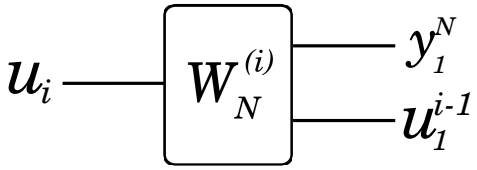

Figure 5: $W_N^{(i)}$.

The permutation matrix B_N inverts the bits of the tuple that each input is associated with. Specifically, if $p_1^N = q_1^N B_N$, then $p_{b_1 b_2 \dots b_n} = q_{b_n b_{n-1} \dots b_1}$ where $n = \log_2 N$.

2.4 Effective Channel

Further deep, we can study the transformation and correspond each information bit to its effective channel. Along these lines, we define $W_N^{(i)}$ with input u_i and outputs y_1^N and u_1^{i-1}. Its transition probabilities are given by

$$\begin{aligned}
W_N^{(i)}(y_1^N, u_1^{i-1}|u_i) &\doteq Prob\{y_1^n, u_1^{i-1}|u_i\} \\
&= \frac{Prob\{y_1^N, u_1^i\}}{Prob\{u_i\}} \\
&= \sum_{u_{i+1}^N} \frac{Prob\{y_1^N, u_1^N\}}{Prob\{u_i\}} \\
&= \sum_{u_{i+1}^N} \frac{Prob\{y_1^N|u_1^N\}Prob\{u_1^N\}}{Prob\{u_i\}} \\
&= \sum_{u_{i+1}^N} \frac{Prob\{y_1^N|u_1^N\}2^{-N}}{2^{-1}} \\
&= \sum_{u_{i+1}^N} \frac{Prob\{y_1^N|u_1^N\}}{2^{N-1}} \\
&= \sum_{u_{i+1}^N} \frac{1}{2^{N-1}} W_N(y_1^N|u_1^N)
\end{aligned}$$

where W_N is the composite channel for the entire information word.

Now that the channel $W_N^{(i)}$ is formally defined we can prove how Bhattacharyya parameter can help us derive an upper bound to the block error probability. The proof in Subsection 2.2 concerns channels with single input and multiple output which applies to $\{W_N^{(i)}\}_{i=1}^N$ as well. If we define $\mathcal{E}_i$ as the event that the i^{th} bit is detected incorrectly, then automatically we are driven to the inequality

$$P(\mathcal{E}_i) \leq Z(W_N^{(i)}).$$

We emphasize that $\mathcal{E}_i$ is independent of $\mathcal{E}_j$ for any $j < i$, because we assume the existence of a "genie" which provides us with the true values of previous information bits.

We define the block error event as

$$\mathcal{E} \doteq \left\{(u_1^N, y_1^N) \in \mathcal{X}^N \times \mathcal{Y}^N : \hat{U}_1^N \neq u_1^N\right\}.$$

We notice that $\mathcal{E} = \bigcup_{i=1}^N \mathcal{E}_i$ which implies the inequality (union bound)

$$P(\mathcal{E}) \leq \sum_{i=1}^N Z\left(W_N^{(i)}\right).$$

The channel polarization is a capacity preserving operation. Mathematically this is expressed as $\sum_{i=1}^{N} I(W_N^{(i)}) = NI(W)$. The above statement is true if the capacity is preserved after the basic polarization step; namely, $I(W_2^{(1)}) + I(W_2^{(2)}) = 2I(W)$. This is easily proven as shown below.

$$I(W_2^{(1)}) = I(U_2; Y_1, Y_2) \tag{5}$$

$$\begin{aligned} I(W_2^{(2)}) &= I(U_1, U_2; Y_1, Y_2) \\ &= I(U_2; Y_1, Y_2|U_1) + I(U_2; U_1) \\ &= I(U_2; Y_1, Y_2|U_1) \end{aligned} \tag{6}$$

$$\text{X}_1 \text{ \& } \text{X}_2 \text{ (symbols sent) } \textbf{are independent} \tag{7}$$

$$\left(\begin{aligned} I(X_1; X_2) &= H(X_1) - H(X_1|X_2) \\ &= H(U_1 \oplus U_2) - H(U_1 \oplus U_2|U_2) \\ &= H(U_1 \oplus U_2) - H(U_1) \\ &= 1 - 1 \\ &= 0 \end{aligned} \right)$$

$$\textbf{(5)+(6)} \Longrightarrow$$

$$\begin{aligned} I(W_2^{(1)}) + I(W_2^{(2)}) &= I(U_1; Y_1, Y_2) + I(U_2; Y_1, Y_2|U_1) \\ &= I(U_1, U_2; Y_1, Y_2) \\ &= I(X_1, X_2; Y_1; Y_2) \\ &= I(X_1; Y_1, Y_2) + I(X_2; Y_1, Y_2) \qquad \text{due to (7)} \\ &= I(X_1; Y_1) + I(X_2; Y_2) \\ &= I(W) + I(W) \\ &= 2I(W) \blacksquare \end{aligned}$$

Of course, channel polarization affects Bhattacharyya parameter.
Proposition 2: It can be proven that

$$\left. \begin{aligned} Z(W_2^{(2)}) &= Z(W)^2 \\ Z(W_2^{(1)}) &\leq 2Z(W) - Z(W)^2 \end{aligned} \right\} \Rightarrow Z(W_2^{(1)}) + Z(W_2^{(2)}) \leq 2Z(W). \ \blacksquare$$

The above inequality implies that $\sum_{i=1}^{N} Z(W_N^{(i)}) \leq NZ(W)$ and guarantees that by polarizing the channel the reliability will not decay.

At this point it is interesting to show how the Bhattacharyya parameters of all the $\{W_N^{(i)}\}_{i=1}^{N}$ channels evolve if W is a BEC and its erasure probability ϵ changes. As we see in Fig. 6, $Z(W_{32}^{(i)}) = 0$ for every i if $\epsilon = 0$ and $Z(W_{32}^{(i)}) = 1$ for every i if $\epsilon = 1$. By looking at the curves that take place between those extreme cases, it becomes obvious that only a few of these curves cross and only once or twice. For N larger than 32 there are more curves that cross but still the fraction of them remains very small.

In the case that W is a binary symmetric channel (BSC), the Bhattacharyya parameter is maximized if the crossover probability is one half (which is the case where the BSC has zero capacity). A channel is explicitly better than the others if its $Z(W_N^{(i)})$ decays faster as the crossover probability moves away from $^1/_2$ (see Fig. 7).

3 Capacity Achievability

The theorem on which the whole study is based is the following: *For any B-DMC W, the channels* $\{W_N^{(i)}\}$ *polarize in the sense that, for any fixed* $\delta \in$ *(0,1), as N goes to infinity through powers of two, the fraction of indeces* $i \in \{1, ..., N\}$ *for which* $I(W_N^{(i)}) \in (1-\delta, 1]$ *goes to* $I(W)$ *and the fraction for which* $I(W_N^{(i)}) \in [0, \delta)$ *goes to* $1 - I(W)$.

The problem that we face as code designers is to determine N (the block length), A (the set of the indeces to the *good* channels), K (the number of the *good* channels), and u_{A^c} (the value of the frozen bits - A^c is the compliment of A over $\{1, ..., N\}$). Regarding to what values frozen bits take, the resulting codeword will be a coset of $u_A G_N(A)$: $x_1^N = u_A G_N(A) \oplus u_{A^c} G_N(A^c)$ - where

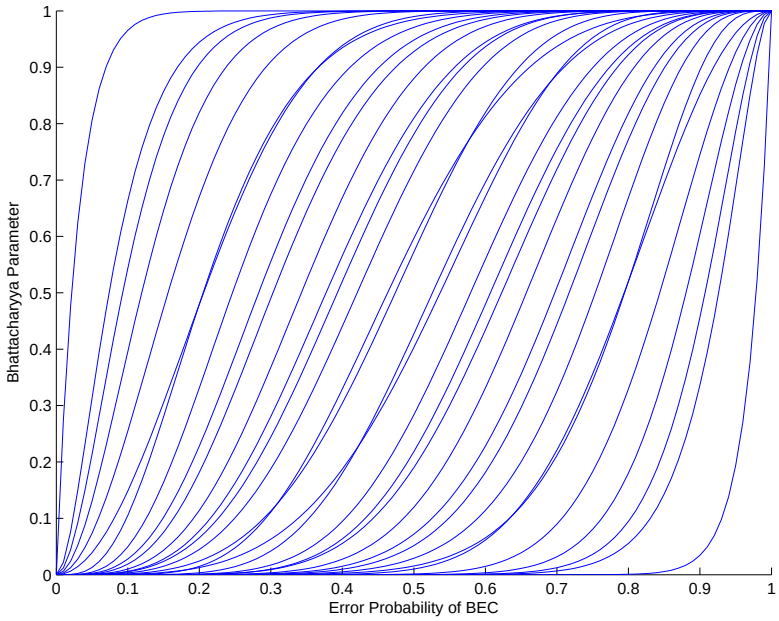

Figure 6: $Z(W_{32}^{(i)})\,\forall i$ and for any BEC with erasure probability $\epsilon \in [0,1]$

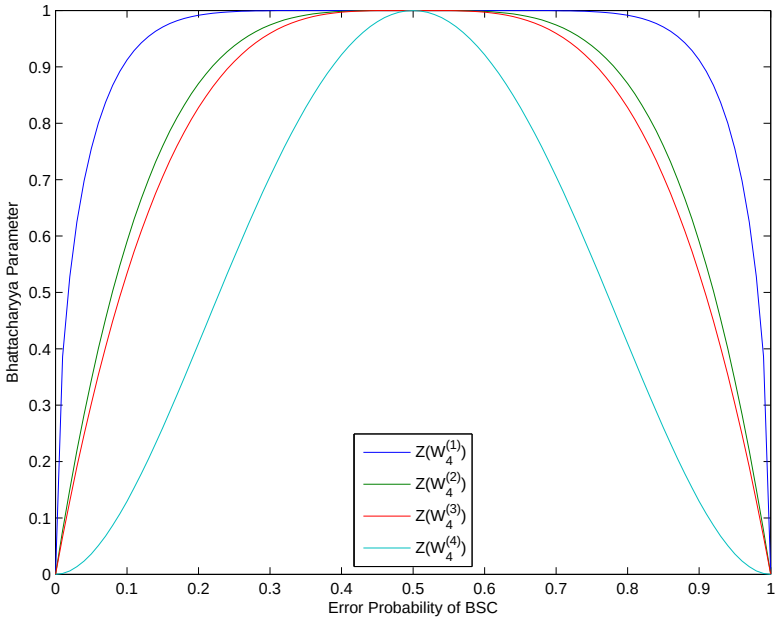

Figure 7: $Z(W_N^{(i)})\,\forall i$ and for BSC with crossover probability in $[0,1]$

by $G_N(A)$ and $G_N(A^c)$ we denote the matrices consisting only of the rows of G_N with index in the set A or A^c, respectively, e.g. by $G_N(\{4,6,7,8\})$ we refer to the fourth, sixth, seventh, and eighth rows of G_N.

The answers to the previous questions (N?, A?, K?, u_{A^c}?) are: the larger N is the better in terms of convergence. This is based on the fact that the fraction of channels that have not yet polarize is vanishing as N grows to infinity. Hence, it is more likely that we choose channels with symmetric information infinitely close to one. K is a function of N and the rate that we choose to transmit, namely $K = \lfloor RN \rfloor$. The information set A is chosen as the K-element subset of $\{1, ..., N\}$ such that $Z(W_N^{(i)}) \leq Z(W_N^{(j)})$ for all $i \in A, j \in A^c$ (this metric will prove to be capacity achieving). Finally, u_{A^c} can be chosen arbitrarily. We have this degree of freedom due to the fact that all the distances are preserved for any coset code. That makes those coset codes equivalent

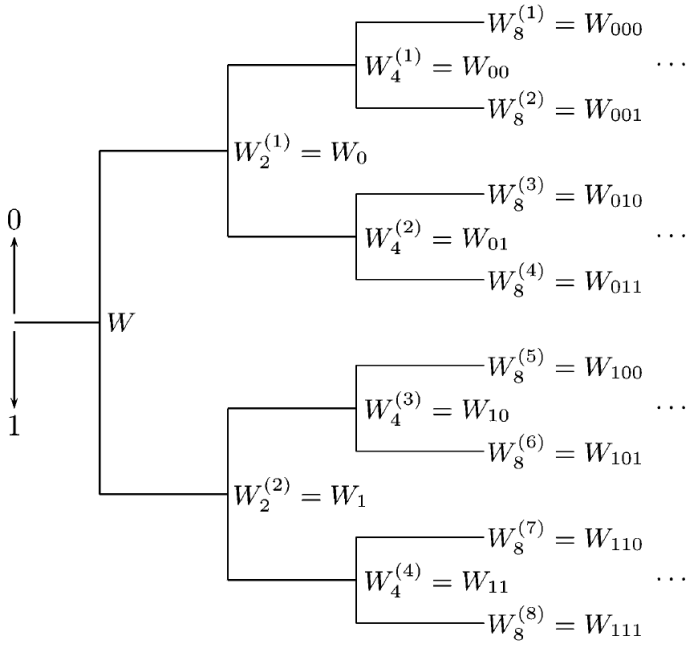

Figure 8: The tree process for the recursive channel construction.

to the original ($u_{A^c} = \mathbf{0}^{N-K}$). Notice that the receiver must have knowledge of u_{A^c} in order to decode appropriately.

The proof of the theorem which claims capacity achievement is outlined with the help of the random sequences I_n and Z_n. Before we define them, we discuss bit indexing of the channels. Fig. 8 depicts how a string of bits $b_1 b_2 \dots b_n$ can be associated with a single channel in the process of polarization. The length of the string indicates the depth and the bits themselves index (naturally) the channel. Now, let us define a walk-through into this tree as a random tree process $\{K_n\}$. The root of the tree, or the value K_0, is the deterministic initial channel W which is indexed with the null string. Given that K_n is a certain channel $W_{b_1 b_2 \dots b_n}$ then K_{n+1} is either $W_{b_1 b_2 \dots b_n 0}$ or $W_{b_1 b_2 \dots b_n 1}$ with probability $\frac{1}{2}$ each. Now we are ready to define I_n as the symmetric capacity of the channels K_n: $I_n = I(K_n) = I(W_{b_1 b_2 \dots b_n}) = I\left(W_{2^n}^{\left(1+\sum_{i=1}^{n} b_i 2^{n-i}\right)}\right)$. In correspondence with the previous steps, Z_n is defined as the Bhattacharyya parameter of channels K_n: $Z_n = Z(K_n) = Z(W_{b_1 b_2 \dots b_n}) = Z\left(W_{2^n}^{\left(1+\sum_{i=1}^{n} b_i 2^{n-i}\right)}\right)$. The next step is to study the stochastic convergence of these sequences. According to Theorem 9.4.6 in [2], I_n converges a.e. to a random variable I_∞ such that $E\{I_\infty\} = I(K_0) = I(W)$. The only thing that remains is to prove that I_∞ is binary, i.e. $I_\infty \in \{0, 1\}$. This will be proven by showing that $Z_\infty \in \{0, 1\}$ which is equivalent because of *Proposition* 1. Theorem 9.4.5 in [2] shows that since Z_n is super-martingale[1] it converges a.e. to Z_∞ such that $E\{|Z_n - Z_\infty|\} \to 0$. By the Cauchy criterion for random series, it follows that $E\{|Z_{n+1} - Z_n|\} \to 0$. But by *Proposition* 2, $Z_{n+1} = Z_n^2$ with probability $\frac{1}{2}$, hence, $E\{|Z_{n+1} - Z_n|\} \geq \frac{1}{2} E\{|Z_n^2 - Z_n|\} = \frac{1}{2} E\{|Z_n(Z_n - 1)|\} = \frac{1}{2} E\{Z_n(1 - Z_n)\} \to 0$, which implies $E\{Z_\infty(1 - Z_\infty)\} = 0$. This, in turn, means that Z_∞ equals 0 or 1 a.e.. ■

We have proven that the Bhattacharyya parameters of the channels $W_N^{(i)}$ polarize a.e. and only

[1] Z_n is a super-martingale because the condition $Z_n \geq E\{Z_{n+1}\}$ holds true:

$$
\begin{aligned}
2Z(W_{b_1 b_2 \dots b_n}) &\geq Z(W_{b_1 b_2 \dots b_n 0}) + Z(W_{b_1 b_2 \dots b_n 1}) \Rightarrow \\
Z(W_{b_1 b_2 \dots b_n}) &\geq \tfrac{1}{2} Z(W_{b_1 b_2 \dots b_n 0}) + \tfrac{1}{2} Z(W_{b_1 b_2 \dots b_n 1}) \\
&= \tfrac{1}{2} Z(K_{n+1} | b_{n+1} = 0) + \tfrac{1}{2} Z(K_{n+1} | b_{n+1} = 1) \\
&= Prob\{b_{n+1} = 0\} Z(K_{n+1} | b_{n+1} = 0) + Prob\{b_{n+1} = 1\} Z(K_{n+1} | b_{n+1} = 1) \\
&= E\{Z(K_{n+1})\} \\
&= E\{Z_{n+1}\} \Rightarrow \\
Z_n &\geq E\{Z_{n+1}\}
\end{aligned}
$$

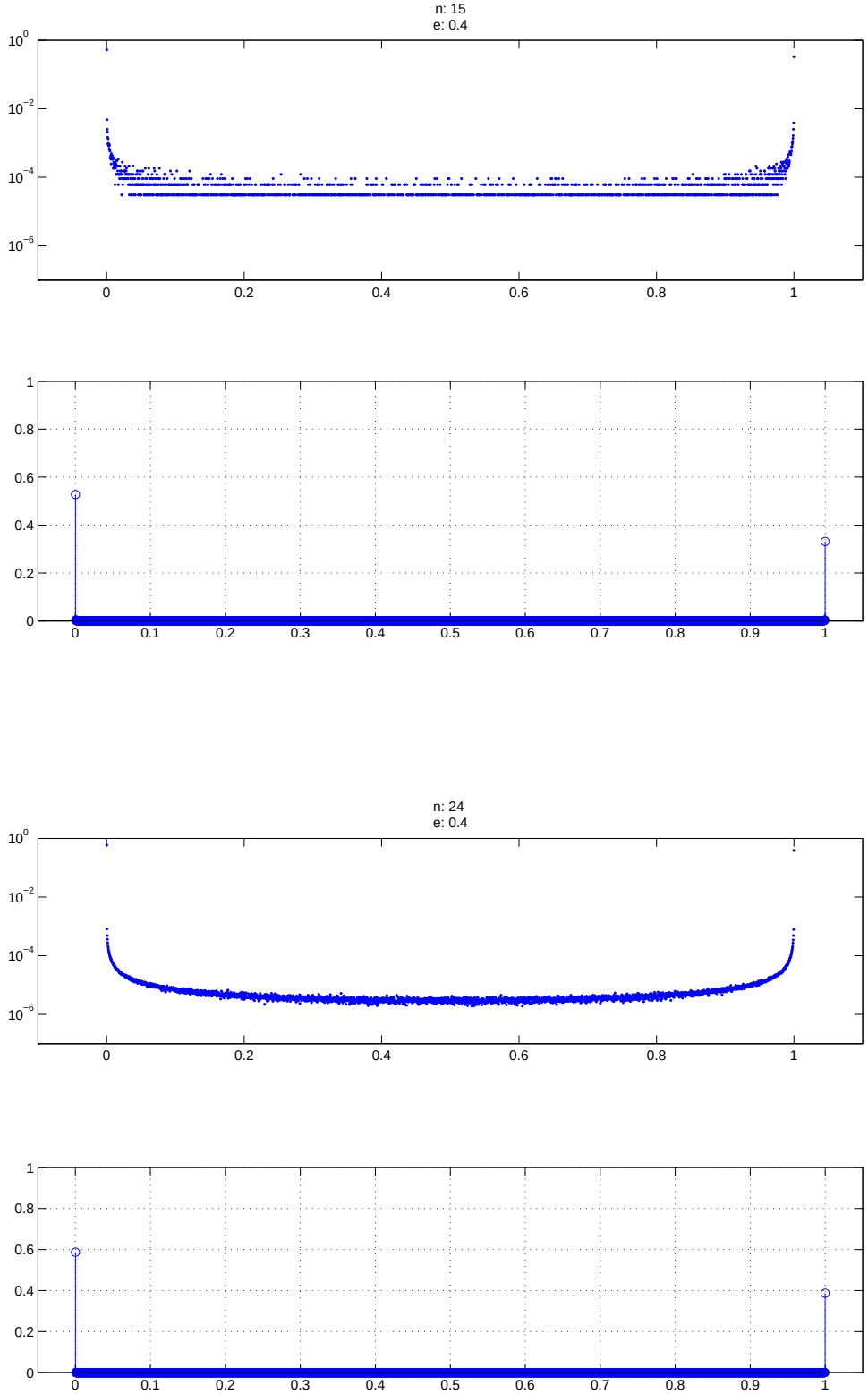

Figure 9: Distribution of random variable $Z = Z(W_{2^n}^{(i)}), i = 1, \ldots, 2^n$ when n = 15 (top), and 24 (bottom), and W BEC(0.4)

as N goes to infinity. Let us see how close theory is to reality. In Fig. 9, we plot the distribution of $Z(W_N^{(i)})$ assuming that W is the BEC with erasure probability 0.4 and $N = 2^n$ where $n = 15$ and 24.

As we can see, N needs to be very large if we expect the mass to gather near 0 and 1. Even 2^{15} is not enough.

4 Information Set Selection

4.1 Reed-Muller Codes

There is another family of codes that have much in common with polar codes. This family is called Reed-Muller (RM) codes and have the same structure. An explicit algebraic description can illustrate better their difference. Suppose that we want to construct the RM and polar codes with block length eight, for given rate and a given channel W, by determining their generator matrices.

Before that, let us recall RM codes and their properties. A RM code can be defined by a pair of integer parameters (m, r). The block length is 2^m and r is related to the rate which equals $\frac{\sum_{i=0}^{r}\binom{m}{i}}{2^m}$. The generator matrix rows is a subset of the row of the Kronecker power $\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}^{\otimes m}$. These rows have Hamming weight greater or equal to 2^{m-r}. For block length eight, m equals three.

$$\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}^{\otimes 3} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} \end{bmatrix} \Longrightarrow$$

$$G_{RM}(1,3) = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

4.2 Polar Codes

As for the polar codes, we construct the code by choosing the rows that correspond to the channels $\{W_8^{(i)} : 1 \leq i \leq 8\}$ with the lower Bhattacharyya parameter. Another difference is that the rows of the Kronecker power are permuted. This is due to the reverse shuffle of the encoder scheme.

$$G_8 = B_8 \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}^{\otimes 3} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

We cannot decide which rows the generator matrix will consist of, because we need the description of the channel W to calculate the Bhattacharyya parameters of channels $\{W_N^{(i)}\}$. Once we have the parameters at hand, we choose the rows of the above matrix that correspond to the channels with the lowest ones.

Two results that are derived from this are the following.

1. RM and polar codes share a common universal pair of encoder and decoder.

2. The polar codes and channel specific designs, for different transition probabilities of W, result in different transition probabilities of $W_N^{(i)}$. Thus, we are driven to different values of $Z(W_N^{(i)})$ that give us different generator matrices.

An example follows that shows that RM and polar codes could be equivalent. Let us study the case of transmissions the take place over a BEC with erasure probability 0.3. The rate is chosen to be 0.5 and the codeword length is 8. Under these conditions the values $Z(W_N^{(i)})$, for $i \in 1, ..., 8$, are

$$\begin{bmatrix} 0.9424 \\ 0.5774 \\ 0.4525 \\ \mathbf{0.0677} \\ 0.3143 \\ \mathbf{0.0295} \\ \mathbf{0.0161} \\ \mathbf{0.0001} \end{bmatrix}.$$

If we drop rows 1, 2, 3, and 5 of G_N, we end up with the generator matrix

$$G_8(A) = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

Apparently, this generator matrix contains the same set of rows with the generator matrix of the RM code. For a relatively low block length the two codes are equivalent. The difference will start to appear for larger codes with block length 32 or 64.

5 Tx and Rx

5.1 Encoder-Decoder

Let's now talk about something more practical. The encoder is implemented as shown in Fig. 3. The complexity is derived with the help of Master Theorem

$$\begin{aligned} T(N) = &\quad \frac{N}{2} + \Theta(N) + 2T\left(\frac{N}{2}\right) \Rightarrow \\ T(N) = &\quad \Theta(N \log(N)). \end{aligned}$$

The decoder proposed in [1] is called successive cancellation (SC) decoder and its role is to decide with the rule of closest neighbor on i^{th} bit ($1 \leq i \leq N$) that transmitted over $W_N^{(i)}$. However ML detection on the channel $W_N^{(i)}$ practically is impossible because the outputs are $N + i - 1$ bits and, hence, the complexity with the straightforward way is $\mathcal{O}(2^N 2^i)$.

Fortunately, we can calculate the likelihood ratios $L_N^{(i)}(y_1^N, u_1^{i-1}) \doteq \dfrac{W_N^{(i)}(y_1^N, u_1^{i-1}|0)}{W_N^{(i)}(y_1^N, u_1^{i-1}|1)}$ by an efficient recursive formula with time and space complexity $\mathcal{O}(N^2)$ and $\mathcal{O}(N)$, respectively. However, an efficient implementation reduces more the asymptotic complexity to $\mathcal{O}(N \log(N))$ for both time and space.

$$L_N^{(2i-1)}(y_1^N, \hat{u}_1^{2i-2}) = \frac{L_{\frac{N}{2}}^{(i)}(y_1^{\frac{N}{2}}, \hat{u}_{1,o}^{2i-2} \oplus \hat{u}_{1,e}^{2i-2}) L_{\frac{N}{2}}^{(i)}(y_{\frac{N}{2}+1}^N, \hat{u}_{1,e}^{2i-2}) + 1}{L_{\frac{N}{2}}^{(i)}(y_1^{\frac{N}{2}}, \hat{u}_{1,o}^{2i-2} \oplus \hat{u}_{1,e}^{2i-2}) + L_{\frac{N}{2}}^{(i)}(y_{\frac{N}{2}+1}^N, \hat{u}_{1,e}^{2i-2})} \tag{8}$$

$$L_N^{(2i)}(y_1^N, \hat{u}_1^{2i-1}) = \left[L_{\frac{N}{2}}^{(i)}(y_1^{\frac{N}{2}}, \hat{u}_{1,o}^{2i-2} \oplus \hat{u}_{1,e}^{2i-2})\right]^{1-2\hat{u}_{2i-1}} L_{\frac{N}{2}}^{(i)}(y_{\frac{N}{2}+1}^N, \hat{u}_{1,e}^{2i-2}) \tag{9}$$

$\mathcal{O}(N \log(N))$ space is required to store the LR's that can be reused (in fact, every LR's will be reused). To see where the computational saving will come from, we inspect (8) and (9) and note that each LR value in the pair

$$\left(L_N^{(2i-1)}(y_1^N, \hat{u}_1^{2i-2}), L_N^{(2i)}(y_1^N, \hat{u}_1^{2i-1})\right)$$

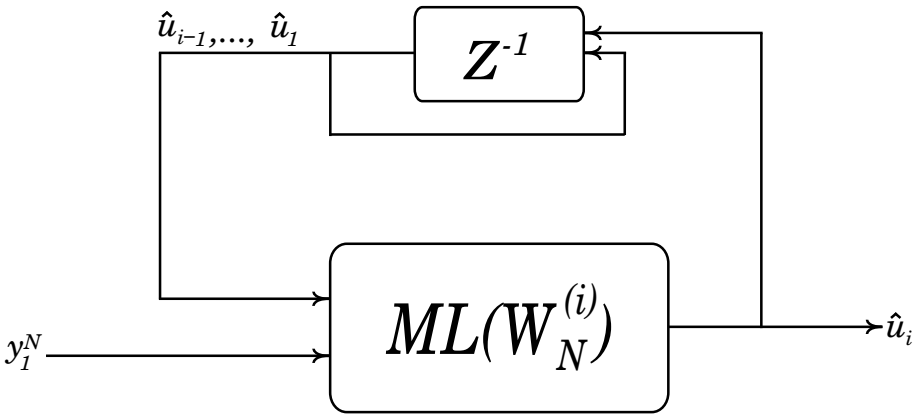

Figure 10: Decoder's scheme

is assembled from the same pair of LRs

$$\left(L^{(i)}_{\frac{N}{2}}(y_1^{\frac{N}{2}}, \hat{u}_{1,o}^{2i-2} \oplus \hat{u}_{1,e}^{2i-2}), L^{(i)}_{\frac{N}{2}}(y_{\frac{N}{2}+1}^{N}, \hat{u}_{1,e}^{2i-2})\right).$$

The size of the matrix in which the LR's will be stored is $N \times (\log_2(N) + 1)$. The number of the elements are exactly as many as the distinct LR's that will be eventually calculated. Each cell is filled after $\Theta(1)$ calculations which implies that the complexity to decode each word is $\mathcal{O}(N \log(N))$.

5.2 Code Construction

Besides encoding and decoding, it is interesting to examine how to construct the code. Even though we have an efficient algorithm to calculate the Bhattacharyya parameters for the special case of BEC, no such algorithm exists for the general B-DMC. So, we present an approximate method that was proposed in [1]. It is easy to grasp the main idea by noticing that the Bhattacharyya parameter of a channel is the mean value of the square root of the likelihood ratios

$$Z(W_N^{(i)}) = E\left\{\sqrt{\frac{W_N^{(i)}(Y_1^N, U_1^{i-1}|U_i \oplus 1)}{W_N^{(i)}(Y_1^N, U_1^{i-1}|U_i)}}\right\}. \tag{10}$$

Thus, by sampling as many samples (U_1^N, Y_1^N) as we can out of the the joint pmf $P_{Y_1^N, U_1^N}(y_1^N, u_1^N) = 2^{-N} W_N(y_1^N|u_1^N)$ and applying an algorithm to calculate the square roots of LR's, we can get a good approximation of the $Z(W_N^{(i)})$ parameters. This algorithm may be a SC decoder since the statistics of its decision elements are exactly the likelihood ratios.

The proof of (10) is as follows.

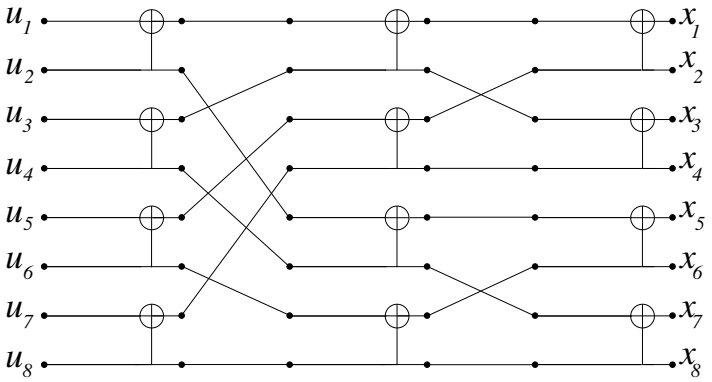

Figure 11: Graphical representation of G_8.

$$
\begin{aligned}
Z(W_N^{(i)}) &= E\left\{\sqrt{\frac{W_N^{(i)}(Y_1^N, U_1^{i-1}|U_i \oplus 1)}{W_N^{(i)}(Y_1^N, U_1^{i-1}|U_i)}}\right\} \\
&= \sum_{u_1^N, y_1^N} \frac{1}{2^N} W_N(y_1^N|u_1^N) \sqrt{\frac{W_N^{(i)}(y_1^N, u_1^{i-1}|u_i \oplus 1)}{W_N^{(i)}(y_1^N, u_1^{i-1}|u_i)}} \\
&= \sum_{u_1^i, y_1^N} \left[\frac{1}{2}\sqrt{\frac{W_N^{(i)}(y_1^N, u_1^{i-1}|u_i \oplus 1)}{W_N^{(i)}(y_1^N, u_1^{i-1}|u_i)}} \sum_{u_{i+1}^N, y_1^N} \left[\frac{1}{2^{N-1}} W_N(y_1^N|u_1^N)\right]\right] \\
&= \frac{1}{2} \sum_{u_1^i, y_1^N} \sqrt{\frac{W_N^{(i)}(y_1^N, u_1^{i-1}|u_i \oplus 1)}{W_N^{(i)}(y_1^N, u_1^{i-1}|u_i)}} \times W_N^{(i)}(y_1^N, u_1^{i-1}|u_i) \\
&= \frac{1}{2} \sum_{u_1^i, y_1^N} \sqrt{W_N^{(i)}(y_1^N, u_1^{i-1}|u_i \oplus 1) W_N^{(i)}(y_1^N, u_1^{i-1}|u_i)} \\
&= \frac{1}{2} \sum_{u_1^i, y_1^N} \sqrt{W_N^{(i)}(y_1^N, u_1^{i-1}|1) W_N^{(i)}(y_1^N, u_1^{i-1}|0)} \\
&\qquad + \frac{1}{2} \sum_{u_1^i, y_1^N} \sqrt{W_N^{(i)}(y_1^N, u_1^{i-1}|0) W_N^{(i)}(y_1^N, u_1^{i-1}|1)} \\
&= \sum_{u_1^{i-1}, y_1^N} \sqrt{W_N^{(i)}(y_1^N, u_1^{i-1}|1) W_N^{(i)}(y_1^N, u_1^{i-1}|0)} \\
&= Z(W_N^{(i)}).
\end{aligned}
$$

Another method for constructing polar codes for any B-DMC was proposed in [3]. The goal of this method is to minimize the probability of error between transmitter and receiver instead of its upper bound. This is achievable by the usage of a modern tool known as density evolution. Via density evolution we can derive the probability of bit error under the belief propagation (BP) decoding. In this case, we can apply the BP decoder to the graphical representation of matrix G_N (example for $N = 8$, Fig. 11).

For BEC's, SC decoding is only an instance of BP decoding because decoding the i^{th} bit with the SC decoder is equivalent to applying BP with the knowledge of the previous bits $(U_1, \ldots, U_{i-1})$. Normally, a usual BP decoder has access also to the next frozen bits (the frozen bit belonging to $U_{i+1}, \ldots, U_N$) [4].

We conclude that is it possible to apply a BP decoder on the sub-graph that corresponds to the i^{th} bit instead of applying the SC decoder. This, in conjunction with the fact that the decoding sub-graph is always a tree (see example in Fig. 12) allows us to use density evolution to extract the probability of error in decoding each bit. On the basis of the availability of $P(\mathcal{E}_i)$'s, one can construct the polar code by choosing the information set A which minimizes

$$\sum_{i \in A} P(\mathcal{E}_i).$$

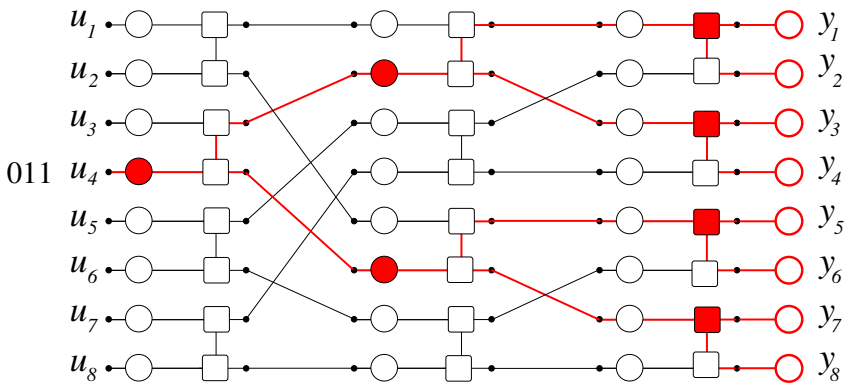

Figure 12: The corresponding tanner tree graph for i^{th} bit.

The only thing that remains is to derive those probabilities of error. Denote a_W the log-likelihood ratio of a channel W. Then $P(\mathcal{E}_i) = \mathfrak{C}(\mathrm{a}_N^i)$. Where:

$$\mathfrak{C}(\mathrm{a}) \doteq \lim_{\epsilon \to +0} \left(\int_{-\infty}^{-\epsilon} \mathrm{a}(x)dx + \frac{1}{2} \int_{-\epsilon}^{+\epsilon} \mathrm{a}(x)dx \right),$$

$$\mathrm{a}_{2N}^{2i} = \mathrm{a}_N^i \circledast \mathrm{a}_N^i, \qquad \mathrm{a}_{2N}^{2i-1} = \mathrm{a}_N^i \boxtimes \mathrm{a}_N^i, \qquad \mathrm{a}_1^1 = \mathrm{a}_W \tag{11}$$

Where $\circledast$ and $\boxtimes$ denote the convolutions of LLR density functions, which are defined in [5], corresponding to variable and check nodes respectively.

Now, (11) is so, due to the fact that, if we assume that the binary expansion of $(i-1)$ is $b_n, \ldots, b_1$, then the nodes at depth t of the tree tanner graph are check nodes and variable node if $b_t = 0$ and $b_t = 1$, respectively (in counting the depth we omit nodes in the tree with degree 2, because messages of BP are passed through such nodes unprocessed). Note that the leaf nodes carry the messages of the channel $Y_i \, \forall i$.

The advantage of this method is its low complexity. If we consider the convolution of the densities in the L as the G domain to be an operation of constant complexity and done with infinitely precise calculations, then the total complexity of calculating all the densities is linear.

$$\chi(N) = N + \chi\left(\frac{N}{2}\right) = N + \frac{N}{2} + \cdots + 1 = \mathcal{O}(N)$$

A Performance Results

The following three figures are selected as the most representative of the performance of polar codes in comparison with regular LDPC (with progressive-edge-growth construction) and under studied decoding algorithms (SC and BP). All of them contain results taken by simulations for codes with half rate (rate $= \frac{1}{2}$) on a BEC. Hence, the horizontal axis is marked with the erasure probability of the BEC and the vertical axis is marked with the magnitude of the probability of bit error between transmitter and receiver. The block length is 2^n.

Fig. 13 shows the BER curves of polar codes under SC for a large variety of block lengths. In Fig.14, we present the performance of polar codes under SC versus regular LDPC(3,6). Finally, the performance of polar codes under BP on G_N versus regular LDPC(3,6) is shown in Fig.15.

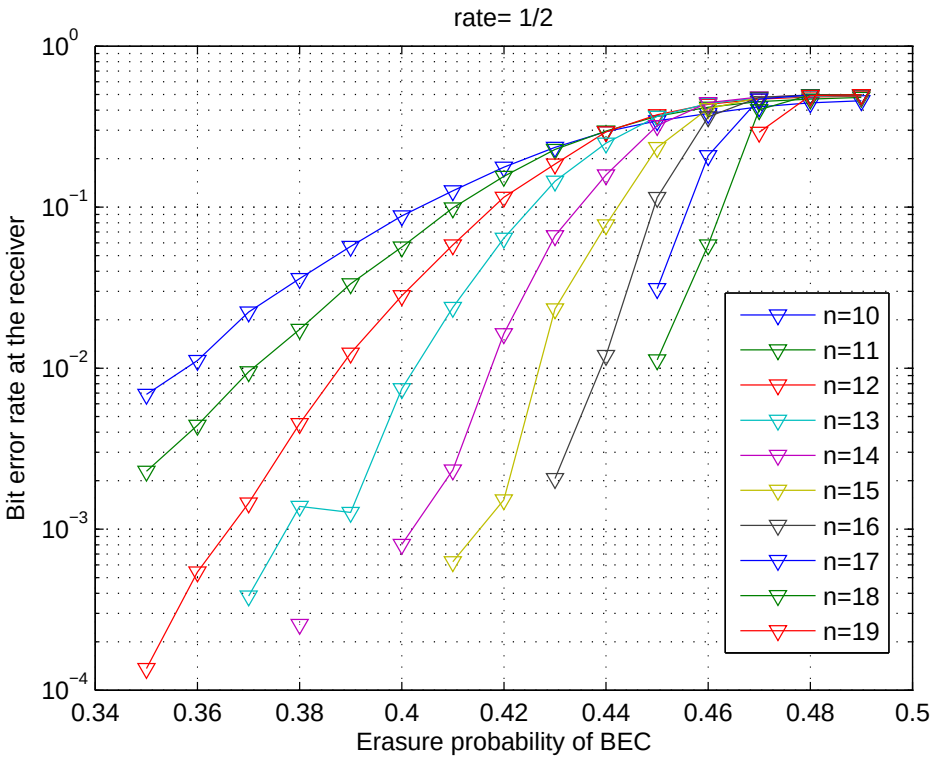

Figure 13: Polar (SC), various block lengths.

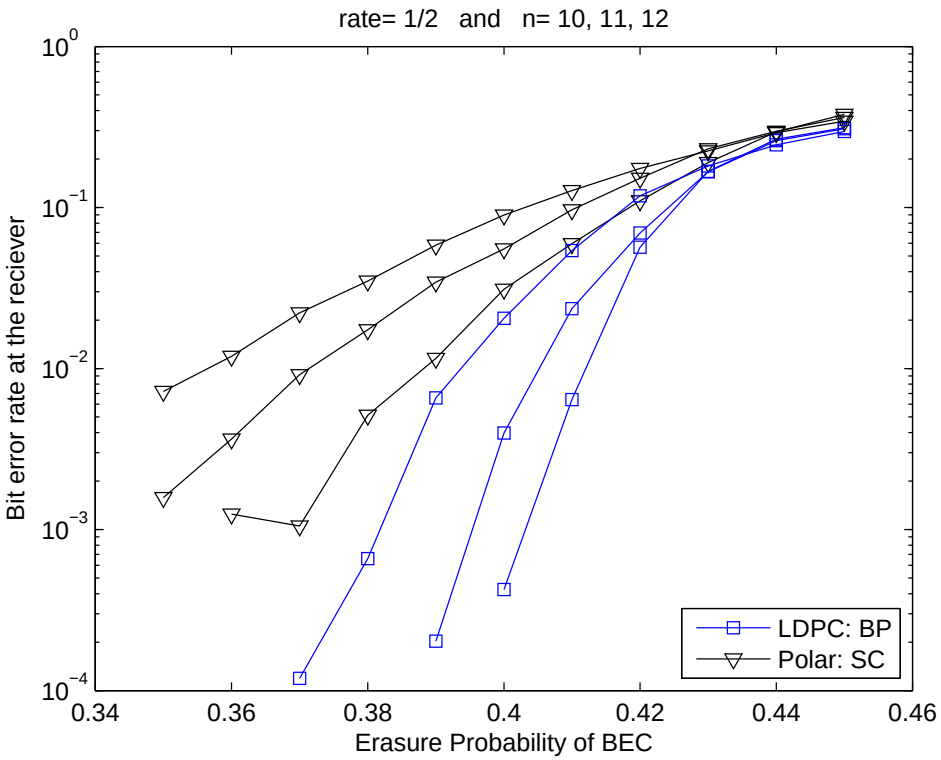

Figure 14: Polar (SC) vs LDPC.

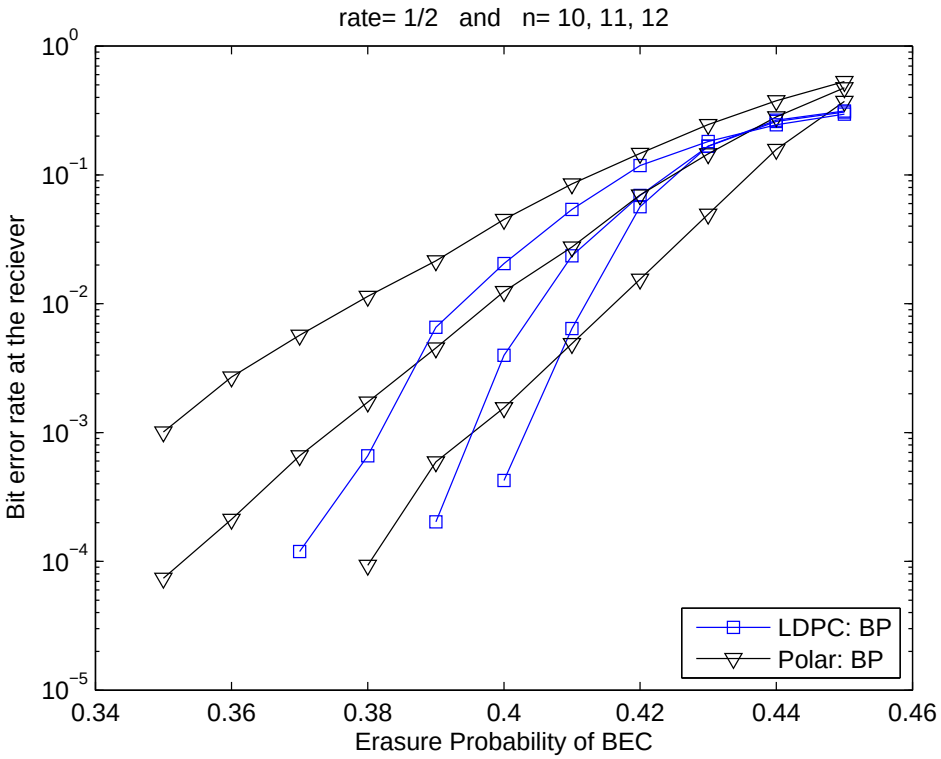

Figure 15: Polar (BP) vs LDPC.

References

[1] E. Arikan, "Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," *IEEE Transactions on Information Theory*, vol. 55, no. 7, July 2009.

[2] K. L. Chung, "A course in probability theory," 3rd edition, *Stanford University, Academic Press*.

[3] R. Mori and T. Tanaka, "Performance and construction of polar codes on symmetric binary-input memoryless channels," in *Proc. of the IEEE Int. Symposium on Inform. Theory*, Seoul, South Korea, July 2009, pp. 1496–1500.

[4] N. Hussami, S. B. Korada, and R. Urbanke "Performance of polar codes for channel and source coding," in *Proc. of the IEEE Int. Symposium on Inform. Theory*, Seoul, South Korea, July 2009, pp. 1488–1492.

[5] T. Richardson and R. Urbanke, "Modern coding theory". *Cambridge University Press*, 2008.

[6] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, Jul.–Oct. 1948.

[7] R. G. Gallager, "Low-density parity-check codes," *IRE Transactions on Inform. Theory*, vol. 8, pp. 21–28, Jan. 1962.

[8] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: Turbo codes," in *Proc. 1993 IEEE Int. Conf. Communications*, Geneva, Switzerland, May 1993, pp. 1064–1070.