

Polar Codes with Higher-Order Memory¹

H. Afşer^{a,b,*} and H. Deliç^{a,**}

^aWireless Communications Laboratory, Department of Electrical and Electronics Engineering, Boğaziçi University, Istanbul, Turkey

^bDepartment of Electrical and Electronics Engineering, Adana Science and Technology University, Adana, Turkey
e-mail: *huseyin.afser@boun.edu.tr, **delic@boun.edu.tr

Received May 17, 2017; revised August 6, 2018; accepted August 7, 2018

Abstract—We introduce a construction of a set of code sequences $\{\mathcal{C}_n^{(m)} : n \geq 1, m \geq 1\}$ with memory order m and code length $N(n)$. $\{\mathcal{C}_n^{(m)}\}$ is a generalization of polar codes presented by Arıkan in [1], where the encoder mapping with length $N(n)$ is obtained recursively from the encoder mappings with lengths $N(n-1)$ and $N(n-m)$, and $\{\mathcal{C}_n^{(m)}\}$ coincides with the original polar codes when $m=1$. We show that $\{\mathcal{C}_n^{(m)}\}$ achieves the symmetric capacity $I(W)$ of an arbitrary binary-input, discrete-output memoryless channel W for any fixed m . We also obtain an upper bound on the probability of block-decoding error P_e of $\{\mathcal{C}_n^{(m)}\}$ and show that $P_e = O(2^{-N^\beta})$ is achievable for $\beta < 1/[1+m(\phi-1)]$, where $\phi \in (1, 2]$ is the largest real root of the polynomial $F(m, \rho) = \rho^m - \rho^{m-1} - 1$. The encoding and decoding complexities of $\{\mathcal{C}_n^{(m)}\}$ decrease with increasing m , which proves the existence of new polar coding schemes that have lower complexity than Arıkan’s construction.

DOI: 10.1134/S0032946018040014

1. INTRODUCTION AND OVERVIEW

Channel polarization [1] is a method to achieve the symmetric capacity $I(W)$ of an arbitrary binary-input, discrete-output memoryless channel (B-DMC) W . By applying channel combining and splitting operations [2], one transforms N uses of W into another set of synthesized binary-input channels. When N increases, the symmetric capacities of the synthesized binary-input channels polarize: $I(W)$ fraction of them gets close to 1, and $1 - I(W)$ fraction of them gets close to 0. The resulting code sequences, called polar codes, have encoding and decoding complexities $O(N \log N)$, and their block error probabilities scale as 2^{-N^β} where $\beta < 1/2$ is the exponent of the code [3].

Let $W: \mathcal{X} \rightarrow \mathcal{Y}$ denote a B-DMC with binary input $x \in \mathcal{X} = \{0, 1\}$ and arbitrary discrete output $y \in \mathcal{Y}$. Considering Arıkan’s polar codes, let W_n denote the vector channel $W_n: \mathcal{X}^N \rightarrow \mathcal{Y}^N$, $N = 2^n$, $n \geq 1$, obtained at channel combining level n . The vector channel W_n is obtained recursively from two independent realizations of W_{n-1} , where the recursion starts with $W_0 = W$. The duplication of W_{n-1} , in a way, creates $N/2$ diversity paths for the $N/2$ inputs of W_{n-1} , and this allows polarization, which one sees in the synthesized binary-input channels obtained by splitting W_n . Consequently, at each combining level the code length doubles with respect to the previous step, scaling as $N = 2^n$.

Considering higher-order memory in channel polarization, let $N = N(n)$ denote the code length at channel combining level n and memory parameter m , $m \geq 1$, which we assume to be fixed. The

¹ This work was supported by Boğaziçi University Research Fund under Project 11A02D10. H. Afşer was also supported by Aselsan Elektronik A.Ş.

vector channel W_n is obtained by combining the inputs of W_{n-1} with an independent realization of W_{n-m} , where one chooses $W_0 = W_{-1} = \dots = W_{1-m} = W$ to initiate the recursion. The number of binary-inputs in W_{n-1} and W_{n-m} are $N(n-1)$ and $N(n-m)$, respectively. In turn, with the controlled memory parameter m and at channel combining level n , one only injects $N(n-m)$ new diversity paths with W_{n-m} , for the $N(n-1)$ inputs of W_{n-1} , to obtain W_n . Because $N(n-m)$ becomes smaller compared with $N(n-1)$ as m increases, it is possible to slow down the speed with which one injects new channels to enable polarization. At first glance, it seems that increasing m will decrease the polarization effect obtained after each combining and splitting stage. However, it will also allow the code length to increase less rapidly in n . In order to see this, consider the code length obeying the recursion

$$N(m) = N(n-1) + N(n-m), \quad n \geq 1, \quad m \geq 1, \quad (1)$$

with initial conditions

$$N(0) = N(-1) = \dots = N(1-m) = 1, \quad m \geq 1. \quad (2)$$

As will be explained in the sequel, the code length takes the form

$$N = O(\phi^n), \quad n \geq 1, \quad (3)$$

where $\phi \in (1, 2]$ is the largest real root of the m th order polynomial

$$F(m, \rho) = \rho^m - \rho^{m-1} - 1, \quad (4)$$

and ϕ decreases with increasing m . Therefore, if we increase m , it will take more channel combining and splitting stages to reach a pre-defined code length, where the ratio of injected diversity paths to existing paths in each combining stage will also decrease. The aim of this paper is to understand the effects of this trade-off on the polarization performance that one can obtain at a fixed code length N .

The original construction of polar codes by Arikan is closely related to the recursive construction of Reed–Muller codes based on the 2×2 kernel $\mathbf{F}_2 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$. For these codes the encoding matrix \mathbf{G}_N is of the form $\mathbf{G}_N = \mathbf{F}_2^{\otimes n}$, where \otimes denotes the Kronecker power, suitably defined in [1]. In [4] Korada et al. generalize the channel polarization idea, where $\ell \geq 2$ independent uses of W_{n-1} are arbitrarily combined to obtain W_n and code length scales as $N = \ell^n$. Although the channel combining mechanism is generalized to combining arbitrary numbers of W_{n-1} to obtain W_n , this setup also possesses first-order memory. The authors express the combining mechanism by an $\ell \times \ell$ polarization kernel \mathbf{K}_ℓ . With an arbitrary \mathbf{K}_ℓ , the encoding matrix takes the form $\mathbf{G}_N = \mathbf{K}_\ell^{\otimes n}$. The asymptotic polarization performance is characterized by the distance properties of the rows of \mathbf{K}_ℓ . The encoding and decoding complexities of these polar codes increase with ℓ , scaling as $O(\ell N \log N)$ and $O((2^\ell N \log N)/\ell)$, respectively.

In this paper we introduce a novel polar code family $\{\mathcal{C}_n^{(m)} : n \geq 1, m \geq 1\}$ with code length $N = O(\phi^n)$, $\phi \in (1, 2]$, and arbitrary but fixed memory parameter m . We show that $\{\mathcal{C}_n^{(m)}\}$ achieves the symmetric capacity of arbitrary BDMCs for any choice of m , which complements Arikan's conjecture that channel polarization is in fact a general phenomenon. By developing a new mathematical framework, we obtain an asymptotic bound on the achievable exponent β of $\{\mathcal{C}_n^{(m)}\}$. The encoding and decoding complexities of $\{\mathcal{C}_n^{(m)}\}$ decrease with increasing m , and $\{\mathcal{C}_n^{(m)}\}$ is the first example of a polar code family that has lower complexity as compared with the original codes presented by Arikan.

Our work differs from [4] in the sense that by introducing higher-order memory we modify the channel combining process. Moreover, the encoding matrix of polar codes with memory $m > 1$

cannot be obtained by applying Kronecker power to an arbitrary polarization kernel. As a result, one needs new mathematical tools to investigate β .

The paper is organized as follows. Section 2 provides the necessary material for the analysis in the sequel. In Section 3 we explain the design, encoding, and decoding of $\{\mathcal{C}_n^{(m)}\}$. In Section 4 we develop a probabilistic framework to investigate $\{\mathcal{C}_n^{(m)}\}$. After showing that $\{\mathcal{C}_n^{(m)}\}$ achieves the symmetric capacity of arbitrary B-DMCs, we obtain an achievable bound on its block-decoding error probability. In Section 5 we analyze impact of higher-order memory on the encoding and decoding complexities of $\{\mathcal{C}_n^{(m)}\}$. Section 6 concludes the paper and provides some future research directions.

Notation: We use upper-case letters A, B for random variables and lower-case a, b for their realizations taking values from sets \mathcal{A}, \mathcal{B} , where the sets have cardinalities $|\mathcal{A}|$ and $|\mathcal{B}|$, respectively. $\Pr(a)$ denotes the probability of the event $A = a$. We write $\mathbf{a}_n = (a_1, a_2, \dots, a_n)$ to denote a vector and $(\mathbf{a}_n, \mathbf{b}_n)$ to denote the concatenation of \mathbf{a}_n and \mathbf{b}_n . We use standard Landau notation $o(n)$, $O(N)$ to denote the limiting values of functions. Unless stated otherwise, proofs are provided in the Appendix.

2. PRELIMINARIES

Let $W(y|x)$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$, denote the transition probabilities of W . Throughout the paper we assume that x is uniformly distributed in \mathcal{X} , and use base-2 logarithm. The symmetric capacity $I(W)$ of W is

$$I(W) \triangleq \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} \frac{1}{2} W(y|x) \log \frac{W(y|x)}{\frac{1}{2} W(y|0) + \frac{1}{2} W(y|1)}. \quad (5)$$

The Bhattacharyya parameter $Z(W)$ of W provides an upper bound on the probability of error for maximum likelihood (ML) decoding over W and is defined as

$$Z(W) \triangleq \sum_{y \in \mathcal{Y}} \sqrt{W(y|0)W(y|1)}. \quad (6)$$

The symmetric cut-off rate $J(W)$ of W is [1]

$$J(W) \triangleq \log \frac{2}{1 + Z(W)}. \quad (7)$$

As Arkan shows in [1, Proposition 1], $Z(W) = 1$ implies $I(W) = 0$, and $Z(W) = 0$ implies $I(W) = 1$. By using this fact and (7), we see that if $J(W) = 0$, then $I(W) = 0$, and if $J(W) = 1$, then $I(W) = 1$.

Let W' and W'' be two B-DMCs with inputs $x_1, x_2 \in \mathcal{X}$ and outputs $y_1 \in \mathcal{Y}_1$ and $y_2 \in \mathcal{Y}_2$, respectively. Channel polarization is based on a single-step channel transformation where one first combines the inputs of W' and W'' to obtain a vector channel

$$W(y_1, y_2 | x_1, x_2) = W'(y_1 | x_1 \oplus x_2) W''(y_2 | x_2). \quad (8)$$

Next, by choosing a channel ordering, one splits the vector channel to obtain two new binary-input channels, $W^-: \mathcal{X} \rightarrow \mathcal{Y}_1 \times \mathcal{Y}_2$ and $W^+: \mathcal{X} \rightarrow \mathcal{X} \times \mathcal{Y}_1 \times \mathcal{Y}_2$, with transition probabilities

$$W^-(y_1, y_2 | x_1) = \sum_{x_2} \frac{1}{2} W'(y_1 | x_1 \oplus x_2) W''(y_2 | x_2), \quad (9)$$

$$W^+(y_1, y_2, x_1 | x_2) = \frac{1}{2} W'(y_1 | x_1 \oplus x_2) W''(y_2 | x_2), \quad (10)$$

We use the following short-hand notation for the transforms in (9) and (10), respectively:

$$W^- = W' \boxplus W'', \quad (11)$$

$$W^+ = W' \boxminus W''. \quad (12)$$

The polarization transforms preserve the symmetric capacity:

$$I(W^-) + I(W^+) = I(W') + I(W''), \quad (13)$$

and they help polarization by creating disparities in $I(W^+)$ and $I(W^-)$ such that

$$I(W^+) \geq \max\{I(W'), I(W'')\}, \quad (14)$$

$$I(W^-) \leq \min\{I(W'), I(W'')\}, \quad (15)$$

where the above inequalities are strict as long as $I(W') \in (0, 1)$ and $I(W'') \in (0, 1)$. This polarization effect is quantitatively observed in the Bhattacharyya parameters: they take the form

$$Z(W^+) = Z(W')Z(W''), \quad (16)$$

$$Z(W^-) \leq Z(W') + Z(W'') - Z(W')Z(W''), \quad (17)$$

where the equality in (17) is achieved if $Z(W') \in \{0, 1\}$ or $Z(W'') \in \{0, 1\}$, or if W' and W'' are binary erasure channels (BECs).

Equations (13)–(17) are proved in [1] when W' is identical to W'' . Their generalizations to the case where W' and W'' are different channels are straightforward and omitted. The proposition below will be crucial in the sequel.

Proposition 1. *We have*

$$J(W^-) + J(W^+) \geq J(W') + J(W''),$$

where equality is achieved only if $J(W') \in \{0, 1\}$ or $J(W'') \in \{0, 1\}$.

The above proposition indicates that one can obtain coding gain by applying channel combining and splitting operations as long as the symmetric cut-off rate of W' and W'' is in $(0, 1)$, where the coding gain manifests itself as an increase in the sum cut-off rate of channels W^- and W^+ as compared with W' and W'' . In this paper we use the parameters $J(W)$ and $I(W)$ together to show that $\{\mathcal{C}_n^{(m)}\}$ achieves $I(W)$ of an arbitrary W , whereas the parameter $Z(W)$ will be used to characterize the polarization performance of $\{\mathcal{C}_n^{(m)}\}$.

3. POLARIZATION WITH HIGHER-ORDER MEMORY

We develop a method to design a family of code sequences $\{\mathcal{C}_n^{(m)} : n \geq 1, m \geq 1\}$ with code length $N = N(n, m) = O(\phi^n)$, $\phi \in (1, 2]$, and fixed memory order m . $\{\mathcal{C}_n^{(m)}\}$ is based on the channel polarization idea of Arikan in [1]. This section is devoted to explaining the design, encoding, and decoding of $\{\mathcal{C}_n^{(m)}\}$, while preparing some grounds for investigating its characteristics in the subsequent sections.

3.1. Channel Combining

Consider an arbitrary B-DMC W , where its N independent uses take the form $W(\mathbf{y}_N | \mathbf{x}_N) = \prod_{i=1}^N W(y_i | x_i)$, $\mathbf{x}_N \in \mathcal{X}^N$, $\mathbf{y}_N \in \mathcal{Y}^N$. Let $\mathbf{u}_N \in \mathcal{X}^N$ be the binary information vector that needs to be

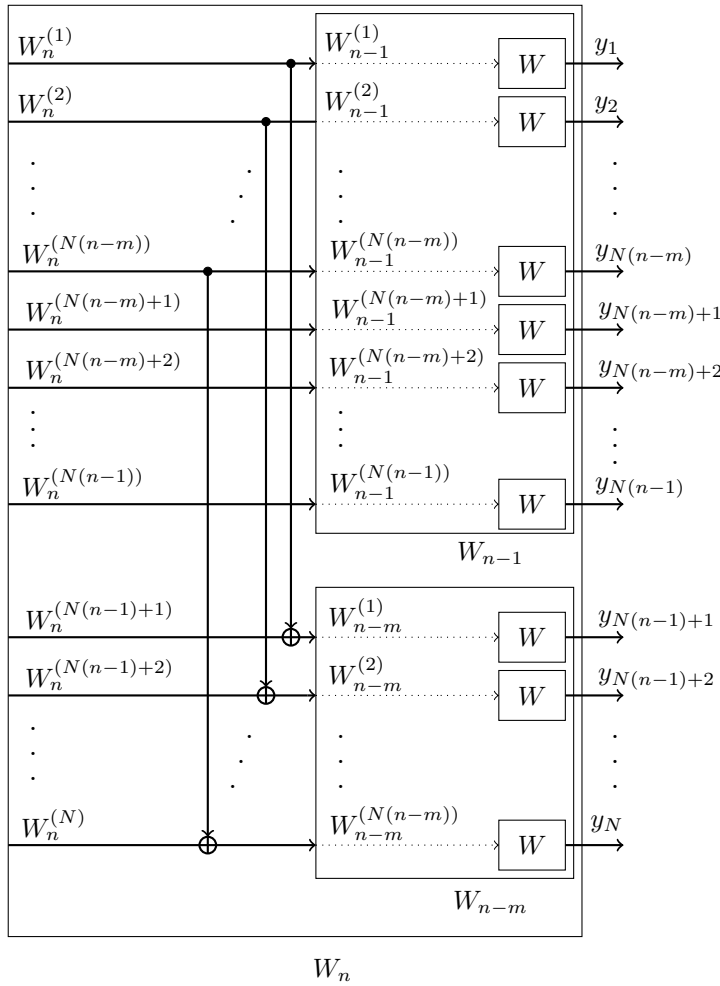


Fig. 1. Recursive construction of the vector channel W_n from W_{n-1} and W_{n-m} , where $W_n^{(i)}$, $i \in \mathbb{N}_n$, denotes the binary-input channels in W_n . The arrows on the left show the directions of flow for the binary-inputs of $W_n^{(i)}$, and \oplus is the XOR operation. The arrows on the right show the outputs of successive uses of W . The XOR operations that take place on the dotted arrows within W_{n-1} and W_{n-1} are not shown, since they obey the same recursion.

transmitted over N uses of W . Channel combining phase creates a vector channel $W_n: \mathcal{X}^N \rightarrow \mathcal{Y}^N$ of the form

$$W_n(\mathbf{y}_N | \mathbf{u}_N) = \prod_{i=1}^N W(y_i | x_i),$$

where $\mathbf{x}_N = \mathbf{u}_N \mathbf{G}_N$, with \mathbf{G}_N being an $N \times N$ encoding matrix and encoding taking place in $GF(2)$.

Let $\mathbb{N}_n = \{1, 2, \dots, N\}$, $N = O(\phi^n)$, denote the set of the indices at the channel combining level n . There are N binary-input channels in W_n to transmit information. We index those channels as $W_n^{(i)}$, $i \in \mathbb{N}_n$, and demonstrate the channel combining operations in Fig. 1. Observe that we index the topmost binary-input channel of W_n as $W_n^{(1)}$, and index i of $W_n^{(i)}$ increases as one moves downwards. The vector channel W_n is obtained by combining W_{n-1} with W_{n-m} . To accomplish this combining, we apply XOR operations on the binary inputs of W_n and transmit the resultant bits through the inputs of W_{n-1} and W_{n-m} . By continuing the same recursion within W_{n-1} and W_{n-m} , the encoded bits are transmitted through independent uses of W channels, because we start

the combining recursion by choosing $W_0 = W_{-1} = \dots = W_{1-m} = W$. If we use the binary-input channels $W_n^{(1)}, W_n^{(2)}, \dots, W_n^{(N)}$ to transmit the symbols u_1, u_2, \dots, u_N , respectively, the encoding matrix \mathbf{G}_N can be expressed as

$$\mathbf{G}_N = \left[\begin{array}{c|c} \mathbf{G}_{N(n-1)} & \mathbf{G}_{N(n-m)} \\ \hline \mathbf{0}_1 & \mathbf{0}_2 \\ \hline & \mathbf{G}_{N(n-m)} \end{array} \right], \quad n \geq 1, \tag{18}$$

where $\mathbf{G}_{N(0)} = \mathbf{G}_{N(-1)} = \dots = \mathbf{G}_{N(1-m)} = [1]$, and $\mathbf{0}_1$ and $\mathbf{0}_2$ are $N(n-m) \times N(n-1)$ and $(N(n-1) - N(n-m)) \times N(n-m)$ all-zero matrices, respectively. Observe that when $m = 1$, the $\mathbf{0}_2$ matrix vanishes and \mathbf{G}_N can be represented as $\mathbf{G}_n = (\mathbf{F}_2^\top)^{\otimes n}$, where $\mathbf{F}_2 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$ is the kernel used by Arikan in [1]. However, when $m > 1$, \mathbf{G}_N cannot be represented via Kronecker power.

3.2. Channel Ordering

After performing the channel combining operation, we have to define an order to split the vector channel $W_n: \mathcal{X}^N \rightarrow \mathcal{Y}^N$ and obtain N binary-input channels. This ordering is carried out with the help of a permutation $\pi_n: \mathbb{N}_n \rightarrow \mathbb{N}_n$. The $W_n^{(i)}$ channels in W_n are split in increasing $\pi_n(i)$ values (from 1 to N) so that each $W_n^{(i)}$ channel is of the form $W_n^{(i)}: \mathcal{X} \rightarrow \mathcal{Y}^N \times \mathcal{X}^{\pi(i)-1}$. In order to explain this operation, we associate a unique state vector $\mathbf{s}_n^{(i)}$ with each $W_n^{(i)}$ channel, which has the form

$$\mathbf{s}_n^{(i)} = (s_1^{(i)}, s_2^{(i)}, \dots, s_n^{(i)})$$

where

$$\mathbf{s}_k^{(i)} \in \{+, -, \star\}, \quad k = 1, 2, \dots, n.$$

The $\mathbf{s}_k^{(i)}$ terms will be referred to as a ‘‘state,’’ and we use $+, -, \star$ symbols to track down the channel transformations that $W_n^{(i)}$ channels undergo as $n = 1, 2, \dots$. States $+$ and $-$ will correspond to the polarization transforms \boxplus and \boxminus , as defined in (9) and (10), respectively, whereas state \star will correspond to a nonpolarizing transform. We let

$$\mathcal{S}_n = \{\mathbf{s}_n^{(i)} : i \in \mathbb{N}_n\} \tag{19}$$

to be the set of all possible state vectors at level n . Since each $\mathbf{s}_n^{(i)} \in \mathcal{S}_n$ is unique (as we will show shortly) we have $|\mathcal{S}_n| = N$ and $\mathcal{S}_n \subset \{+, -, \star\}^n$. The vectors $\mathbf{s}_n^{(i)} \in \mathcal{S}_n$ are assigned recursively from $\mathbf{s}_{n-1}^{(j)} \in \mathcal{S}_{n-1}$, with a state assigning procedure $\varphi_n: \mathcal{S}_{n-1} \rightarrow \mathcal{S}_n$. The operation of φ_n is explained in the following definition.

Definition 1 (state vector assignment procedure). Let $\mathbf{s}_{n-1}^{(j)} \in \mathcal{S}_{n-1}$ be the state vector of $W_{n-1}^{(j)}$. The state vectors $\mathbf{s}_n^{(i)} \in \mathcal{S}_n$ associated with $W_n^{(i)}$ take the form

$$\mathbf{s}_n^{(j)} = (\mathbf{s}_{n-1}^{(j)}, +), \quad \mathbf{s}_n^{(j+N(n-1))} = (\mathbf{s}_{n-1}^{(j)}, -), \quad j \in \mathbb{N}_{n-m}, \tag{20}$$

$$\mathbf{s}_n^{(j)} = (\mathbf{s}_{n-1}^{(j)}, \star), \quad j \in \mathbb{N}_{n-1} \setminus \mathbb{N}_{n-m}. \tag{21}$$

Based on the above definition, as is also demonstrated in Fig. 2, we observe that φ_n appends a new state $\{+, -, \star\}$ to $\mathbf{s}_{n-1}^{(j)} \in \mathcal{S}_{n-1}$ in order to construct $\mathbf{s}_n^{(i)} \in \mathcal{S}_n$. For $j \in \mathbb{N}_{n-m}$, φ_n appends $+$ and $-$ to $\mathbf{s}_{n-1}^{(j)}$ to obtain $\mathbf{s}_n^{(j)}$ and $\mathbf{s}_n^{(j+N(n-1))}$, respectively. For $j \in \mathbb{N}_{n-1} \setminus \mathbb{N}_{n-m}$, φ_n appends \star to $\mathbf{s}_{n-1}^{(j)}$ in order to construct $\mathbf{s}_n^{(j)}$. Because of the inherent memory in the combining procedure,

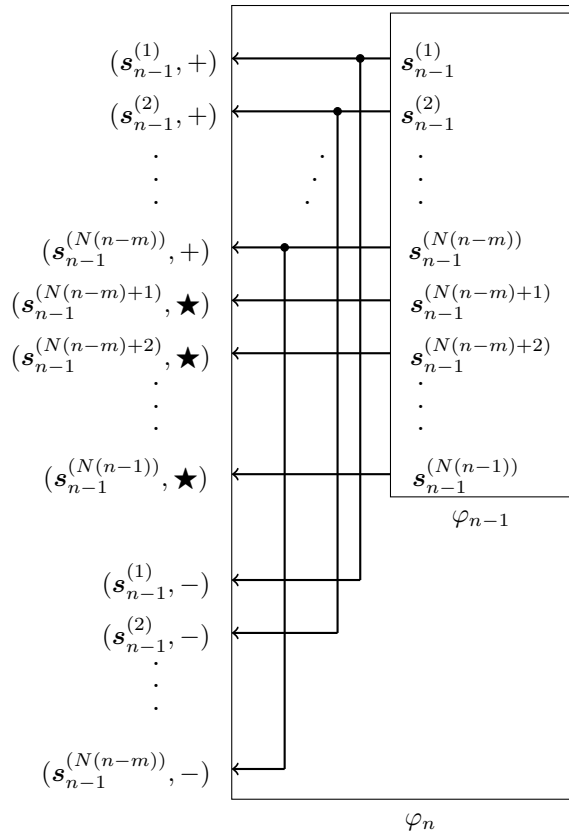


Fig. 2. State labeling procedure $\varphi_n: \mathcal{S}_{n-1} \rightarrow \mathcal{S}_n$. State vectors $\mathbf{s}_n^{(i)} \in \mathcal{S}_n$, are obtained by appending a new state $\{+, -, \star\}$, to the vectors $\mathbf{s}_{n-1}^{(j)} \in \mathcal{S}_{n-1}$.

it is difficult to obtain closed-form expressions for $\mathbf{s}_n^{(i)}$, for any i and m . Nevertheless, with the above definition one can recursively obtain $\mathbf{s}_n^{(i)}$ by applying $\varphi_1, \varphi_2, \dots, \varphi_n$. With the following proposition, we give the formal structure of the possible state vector $\mathbf{s}_n^{(i)}$ and thus the set \mathcal{S}_n .

Proposition 2. *Let $\mathbf{s}_n \in \mathcal{S}_n$ be a valid state vector one can obtain after applying $\varphi_1, \varphi_2, \dots, \varphi_n$. Only the transitions between s_k and s_{k+1} , $k = 1, 2, \dots, n$, that are shown in the state transition diagram of Fig. 3 are possible, where the imposed initial condition is $s_1 \in \{+, -\}$.*

The above proposition is a direct consequence of the channel combining and state vector assigning procedure φ_n , and it can be verified by induction through stages $\varphi_1, \varphi_2, \dots, \varphi_n$.

Proposition 3. *The state vector $\mathbf{s}_n^{(i)} \in \mathcal{S}_n$, $i \in \mathbb{N}_n$, assigned to each $W_n^{(i)} \in \mathcal{W}_n$ is unique.*

The above proposition will be crucial for the ongoing analysis as it states that each $W_n^{(i)}$ is uniquely addressable by $\mathbf{s}_n^{(i)}$. We will use this fact to obtain the ordering π_n . Before accomplishing this, we obtain binary vectors $\mathbf{b}_n^{(i)} = (b_1^{(i)}, b_2^{(i)}, \dots, b_n^{(i)})$, $b_k^{(i)} \in \mathcal{X}$, $k = 1, 2, \dots, n$, from $\mathbf{s}_n^{(i)}$, which will allow us to sort and provide an order. The mapping between $\mathbf{s}_n^{(i)}$ and $\mathbf{b}_n^{(i)}$ is obtained as

$$b_k^{(i)} = \begin{cases} 0 & \text{if } s_k^{(i)} \in \{-, \star\}, \\ 1 & \text{if } s_k^{(i)} = +, \end{cases} \quad k = 1, 2, \dots, n. \tag{22}$$

We note that although both $s_k^{(i)} = -$ and $s_k^{(i)} = \star$ are mapped as $b_k^{(i)} = 0$, the $\mathbf{b}_n^{(i)}$ vectors will also be unique for each i , because every state $-$ in $\mathbf{s}_n^{(i)}$ is followed by $m - 1$ occurrences of state \star , and

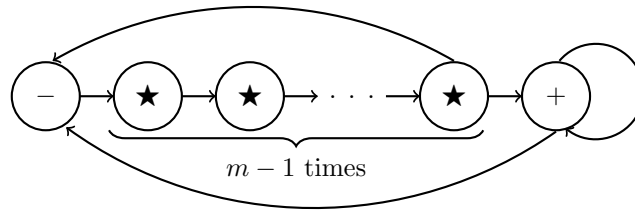


Fig. 3. Possible state transitions observed between s_k and s_{k+1} , $k = 1, 2, \dots, n$.

the distinction between different $\mathbf{s}_n^{(i)}$ is hidden in the location of + states in $\mathbf{s}_n^{(i)}$. The following definition uses this uniqueness property to obtain the ordering, π_n . It is an adaptation of the bit-reversed order of Arikan in [1] to the proposed coding scheme.

Definition 2 (bit-reversed order). Let $(\mathbf{b}_n^{(i)})_2$ denote the value of $\mathbf{b}_n^{(i)}$ in the mod-2 notation $(b_1^{(i)}, b_2^{(i)}, \dots, b_n^{(i)})_2$ where $b_1^{(i)}$ is the most significant bit. The uniqueness of $\mathbf{b}_n^{(i)}$ for each i ensures the existence of a permutation $\pi_n: \mathbb{N}_n \rightarrow \mathbb{N}_n$ so that for some $i, j \in \mathbb{N}_n$ we have $\pi_n(i) < \pi_n(j)$ if $(\mathbf{b}_n^{(i)})_2 < (\mathbf{b}_n^{(j)})_2$.

The bit-reversed order π_n is obtained in terms of increasing $(\mathbf{b}_n^{(i)})_2$ values. Notice that the binary input channels $W_{n-m}^{(j)}$, $j \in \mathbb{N}_{n-m}$, of Fig. 1 have no effect in the recursive state assigning procedure φ_n , and thus in the bit-reversed order. Their sole purpose is to provide auxiliary channels for the combining process. In fact, the $N(n-m)$ inputs of W_{n-m} can be combined with the $N(n-1)$ inputs of W_{n-1} in $N(n-1)!/N(n-m)!$ different ways. However, we deliberately align the inputs of W_{n-1} and W_{n-m} so that the first $N(n-m)$ inputs of W_{n-1} are combined, respectively, with the first $N(n-m)$ inputs of W_{n-m} , as is shown in Fig. 1. This alignment in the combining process will be crucial in Section 4 when we investigate the evolution of binary-input channels in a probabilistic setting, because the channel pairs $W_{n-1}^{(j)}$ and $W_{n-m}^{(j)}$ share the same state history, as is explained in the following proposition.

Proposition 4. Let $\mathbf{s}_{n-1}^{(j)} = (s_1, s_2, \dots, s_{n-1}) \in \mathcal{S}_{n-1}$ be the state vector of $W_{n-1}^{(j)}$. Channel $W_{(n-m)}^{(j)}$ shares the same state history with $W_{(n-1)}^{(j)}$ through combining stages $1, 2, \dots, n-m$ in the sense that its state vector is $\mathbf{s}_{n-m}^{(j)} = (s_1, s_2, \dots, s_{n-m}) \in \mathcal{S}_{n-m}$.

3.3. Channel Splitting

We assume a genie-aided decoding mechanism where the $W_n^{(i)}$ channels are decoded successively in increasing $\pi_n(i)$ values, from 1 to N , and the genie provides the true values of already decoded bits. The decoder has no knowledge of the future bits that it will decode. With these assumptions, $W_n^{(i)}$ is the effective bit-channel that this genie-aided decoder faces while trying to decode its next bit. Let us define $u_n^{(i)} \in \mathcal{X}$ as

$$u_n^{(i)} = \text{binary input of the channel } W_n^{(i)},$$

and for $i, j \in \mathbb{N}_n$, let

$$\begin{aligned} \mathbf{u}_{n,b}^{(i)} &\triangleq (u_n^{(j)} : \pi_n(j) < \pi_n(i)), \\ \mathbf{u}_{n,a}^{(i)} &\triangleq (u_n^{(j)} : \pi_n(j) > \pi_n(i)). \end{aligned} \tag{23}$$

The $\mathbf{u}_{n,b}^{(i)}$ and $\mathbf{u}_{n,a}^{(i)}$ are the information vectors that are decoded by the genie-aided decoder before and after $u_n^{(i)}$, respectively. The length of $\mathbf{u}_{n,b}^{(i)}$ is $\pi_n(i) - 1$ and the length of $\mathbf{u}_{n,a}^{(i)}$ is $N - \pi_n(i)$,

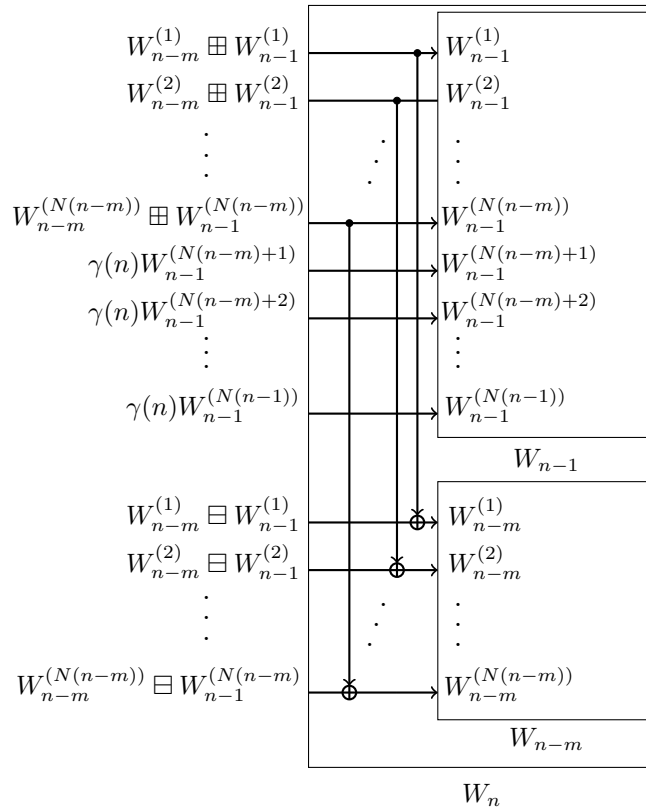


Fig. 4. Transition probabilities of $W_n^{(i)}$ channels after combining and splitting W_{n-1} and W_{n-m} .

so that $\mathbf{u}_{n,b}^{(i)} \in \mathcal{X}^{\pi_n(i)-1}$ and $\mathbf{u}_{n,a}^{(i)} \in \mathcal{X}^{N-\pi_n(i)}$. The following definition formalizes the transition probabilities of the $W_n^{(i)}$ channels:

$$W_n^{(i)} \triangleq \sum_{\mathbf{u}_{n,a}^{(i)}} \Pr(\mathbf{y}_N, \mathbf{u}_{n,a}^{(i)}, \mathbf{u}_{n,b}^{(i)} | u_n^{(i)}). \tag{24}$$

The above definition indicates that $W_n^{(i)}$ is the posterior probability of an arbitrary B-DMC obtained at channel combining and splitting level n . The genie-aided decoder has no knowledge of $\mathbf{u}_{n,a}^{(i)}$. It therefore averages the joint probability of all outputs and all inputs over $\mathbf{u}_{n,a}^{(i)}$, and takes \mathbf{y}_N and $\mathbf{u}_{n,b}^{(i)}$ as the effective output (observation) of the combined channels. Hence, each $W_n^{(i)}$ has input $u_n^N(n-1)(i) \in \mathcal{X}$ and output $(\mathbf{y}_N, \mathbf{u}_{n,b}^{(i)}) \in \mathcal{Y}^N \times \mathcal{X}^{\pi_n(i)-1}$.

Proposition 5. *The transition probabilities of $W_n^{(i)}$ channels take the following forms:*

$$W_n^{(j)} = W_{n-m}^{(j)} \boxplus W_{n-1}^{(j)}, \quad W_n^{(j+N(n-1))} = W_{n-m}^{(j)} \boxminus W_{n-1}^{(j)}, \quad j \in \mathbb{N}_{n-m}, \tag{25}$$

$$W_n^{(j)} = \gamma(n)W_{n-1}^{(j)}, \quad j \in \mathbb{N}_{n-1} \setminus \mathbb{N}_{n-m}, \tag{26}$$

where $\gamma(n) = \Pr(y_{N(n-1)+1}, y_{N(n-1)+2}, \dots, y_N)$ and $W_0 = W_{-1} = \dots = W_{1-m} = W$.

The above proposition is illustrated in Fig. 4, where we see that the overall effect of the XOR operations after channel splitting is to provide diversity paths for the $N(n-m)$ inputs of W_{n-1} in the sense that for $j \in \mathbb{N}_{n-m}$ we have $W_n^{(j)} = W_{n-m}^{(j)} \boxplus W_{n-1}^{(j)}$. Therefore, the input of $W_n^{(j)}$ is transmitted through both $W_{n-m}^{(j)}$ and $W_{n-1}^{(j)}$. Note that in order to provide this diversity, the

inputs of $W_n^{(j+N(n-1))}$ must be decoded by the genie-aided decoder before the inputs of $W_n^{(j)}$, indicating that $\pi_n(j) > \pi_n(j + N(n - 1))$ must hold. Thanks to the bit-reversed order, as is explained in Definition 2, this requirement can easily be satisfied. To see this, consider the state vectors $\mathbf{s}_{n-1}^{(j)}$ of $W_{n-1}^{(j)}$ to which one appends + and - in order to construct $\mathbf{s}_n^{(j)}$ and $\mathbf{s}_n^{(j+N(n-1))}$, respectively. After this operation, the mapping between $\mathbf{s}_n^{(i)}$ and $\mathbf{b}_n^{(i)}$, as given by (22), indicates that $\mathbf{b}_n^{(j)} = (\mathbf{b}_{n-1}^{(j)}, 1)$ and $\mathbf{b}_n^{(j+N(n-1))} = (\mathbf{b}_{n-1}^{(j)}, 0)$ hold. Thus,

$$(\mathbf{b}_n^{(j)})_2 > (\mathbf{b}_n^{(j+N(n-1))})_2, \quad n = 1, 2, \dots,$$

and by Definition 2, $\pi_n(j) > \pi_n(j + N(n - 1))$ holds for all $n \geq 1$. On the other hand, in order to decode $W_n^{(j+N(n-1))}$ correctly, the inputs of $W_{n-1}^{(j)}$ and $W_{n-m}^{(j)}$ must be decoded correctly, which implies that we must have $W_n^{(j+N(n-1))} = W_{n-m}^{(j)} \boxplus W_{n-1}^{(j)}$. The above analysis proves (25) by induction through the combining and splitting stages $1, 2, \dots, n$. In order to prove (26), we note that for $j \in \mathbb{N}_{n-1} \setminus \mathbb{N}_{n-m}$ the channel $W_n^{(j)}$ is as good as $W_{n-1}^{(j)}$ in the sense that the genie-aided decoder can always decode $W_{n-1}^{(j)}$ instead of $W_n^{(j)}$. Inspecting Fig. 4, we observe that the binary input of $W_n^{(j)}$ is not transmitted through the inputs of W_{n-m} . Therefore, the combining of W_{n-m} with W_{n-1} does not provide any new information regarding the input of $W_n^{(j)}$. This, in turn, indicates that $W_n^{(j)}$ is the same as $W_{n-1}^{(j)}$ except for a scaling factor $\gamma(n)$, as in (26).

3.4. Effects of Channel Combining and Splitting on the Symmetric Capacity

Let us define $I_n^{(i)} = I(W_n^{(i)})$ and analyze the implications of Proposition 5. Equation (25) states that the channel pairs $W_{n-m}^{(j)}$ and $W_{n-1}^{(j)}$, $j \in \mathbb{N}_{n-m}$, undergo a polarization transform, \boxplus and \boxminus , from which two new channels, $W_n^{(j)}$ and $W_n^{(j+N(n-1))}$, emerge. In the light of (14) we have

$$I_n^{(j)} \geq \max\{I_{n-1}^{(j)}, I_{n-m}^{(j)}\}, \quad j \in \mathbb{N}_{n-m}. \tag{27}$$

The injection of $W_{n-m}^{(j)}$ enables $W_n^{(j)}$ to be a superior channel compared to $W_{n-m}^{(j)}$ and $W_{n-1}^{(j)}$. This comes with the expense that now $W_n^{(j+N(n-1))}$ is an inferior channel as compared to $W_{n-m}^{(j)}$ and $W_{n-1}^{(j)}$, because from (15) one has

$$I_n^{(j+N(n-1))} \leq \min\{I_{n-1}^{(j)}, I_{n-m}^{(j)}\}, \quad j \in \mathbb{N}_{n-m}. \tag{28}$$

Although $I_n^{(j)}$ and $I_n^{(j+N(n-1))}$ move away from $I_{n-1}^{(j)}$ and $I_{n-m}^{(j)}$, the transformations preserve the symmetric capacity, because, as is indicated by (13), we have

$$I_n^{(j)} + I_n^{(j+N(n-1))} = I_{n-1}^{(j)} + I_{n-m}^{(j)}, \quad j \in \mathbb{N}_{n-m}. \tag{29}$$

The remaining channels $W_n^{(j)}$, $j \in \mathbb{N}_{n-1} \setminus \mathbb{N}_{n-m}$, in (26) do not see any polarization transforms: their transition probabilities are scaled by $\Pr(y_{N(n-1)+1}, \dots, y_N)$ with respect to $W_{n-1}^{(j)}$. This scaling, in turn, results in

$$I_n^{(j)} = I_{n-1}^{(j)}, \quad j \in \mathbb{N}_{n-1} \setminus \mathbb{N}_{n-m}. \tag{30}$$

All in all, the combining and splitting of W_{n-1} and W_{n-m} preserves the sum symmetric capacity:

$$\sum_{i \in \mathbb{N}_n} I_n^{(i)} = \sum_{j \in \mathbb{N}_{n-1}} I_{n-1}^{(j)} + \sum_{k \in \mathbb{N}_{n-m}} I_{n-m}^{(k)}. \tag{31}$$

3.5. Decoding

We will take successive cancellation decoding (SCD) of [1] as the default decoding method for $\{\mathcal{C}_n^{(m)}\}$. The genie-aided decoder that we have explained in Section 3.3 and the definition of $W_n^{(i)}$ as given by (24) already provide us with a guideline for SCD. The only difference is that during the calculation of (24), SCD uses its own estimates for the vector $\mathbf{u}_{n,b}^{(i)}$, which we denote by $\hat{\mathbf{u}}_{n,b}^{(i)}$.

Likelihood ratios (LRs) should be preferred in SCD so that one can eliminate the term $P(y_{N(n-1)+1}, y_{N(n-1)+2}, \dots, y_{N_n})$ in (26). The LR for the channel $W_n^{(i)}$ is defined as

$$L_n^{(i)} \triangleq \frac{\sum_{\mathbf{u}_{n,a}^{(i)}} \Pr(\mathbf{y}_N, \mathbf{u}_{n,a}^{(i)}, \hat{\mathbf{u}}_{n,b}^{(i)} | 0)}{\sum_{\mathbf{u}_{n,a}^{(i)}} \Pr(\mathbf{y}_N, \mathbf{u}_{n,a}^{(i)}, \hat{\mathbf{u}}_{n,b}^{(i)} | 1)}.$$

By using the LR relationships given in [1] for the \boxplus and \boxminus transformations, from Proposition 5 we obtain

$$L_n^{(j)} = L_{n-1}^{(j)} (L_{n-m}^{(j)})^{1-2\tilde{u}_n^{(j+N(n-1))}}, \tag{32}$$

$$L_n^{(j+N(n-1))} = \frac{L_{n-1}^{(j)} L_{n-m}^{(j)} + 1}{L_{n-1}^{(j)} + L_{n-m}^{(j)}}, \quad j \in \mathbb{N}_{n-m},$$

$$L_n^{(j)} = L_{n-1}^{(j)}, \quad j \in \mathbb{N}_{n-1} \setminus \mathbb{N}_{n-m}. \tag{33}$$

Therefore, while decoding $W_n^{(i)}$ one only needs to calculate $2N(n - m)$ LRs as given by (32), while the remaining $N - N(n - m)$ LRs for (33) are the same as on the previous level. This fact can be exploited to avoid unnecessary decoding complexity in hardware implementation.

3.6. Code Length

Recall that the code length $N = N(n)$ obeys the recursion in (1) with initial conditions of (2). It is easy to show that N can be calculated as

$$N = \sum_{i=1}^m c_i (\rho_i)^n, \tag{34}$$

where each $\rho_i, i = 1, 2, \dots, m$, is a root of the m th order polynomial

$$F(m, \rho) = \rho^m - \rho^{m-1} - 1, \tag{35}$$

and the constants c_i are calculated by using the initial conditions in (2) together with (34).

Proposition 6. For $m \geq 1$, let $\phi \in (1, 2]$ be a real root of $F(m, \rho)$. Then

1. ϕ is unique, i.e., there is only one real root in $(1, 2]$;
2. If $\rho_i \neq \phi$, we have $\sqrt{\rho_i \rho_i^*} / \phi < 1$, indicating that ϕ is the the largest magnitude root of $F(m, \rho)$;
3. ϕ is decreasing with increasing m .

Part 2 of the above proposition says that, as n gets large, the summation in (34) will be dominated by the ϕ^n term, and thus the code length will scale as $N = \varkappa \phi^n = O(\phi^n)$ where $\varkappa > 0$ is the constant scaler of ϕ^n in (34). Part 3 of Proposition 6 implies that as m increases the code length increases less rapidly in n , which we mentioned in the beginning of the paper.

3.7. Code Construction

The following proposition is a generalization of [1, Proposition 5], and its proof is omitted.

Proposition 7. *If W is a BEC, then $W_n^{(i)}$ channels obeying the transition probabilities as given by Proposition 5 are also BECs.*

In order to use $\{\mathcal{C}_n^{(m)}\}$ one has to fix a code parameter vector (W, N, K, \mathcal{A}) , where W is the underlying B-DMC, N is the code length, K is the dimensionality of the code, and $\mathcal{A} \subseteq \mathbb{N}_n$ is the set of information carrying symbols. We have $|\mathcal{A}| = K$ and $K/N = R$, where $R \in [0, 1]$ is the rate of the code.

Let $P_{e,n}^{(i)}$, $i \in \mathbb{N}_n$, denote the bit-error probability of $W_n^{(i)}$ with SCD. Code construction is choosing the set \mathcal{A} so that $\sum_{i \in \mathcal{A}} P_{e,n}^{(i)}$ is minimum. This problem can be analytically solved only when W is a BEC [1], since for this case the $W_n^{(i)}$ channels are also BECs (Proposition 7) and the Bhattacharyya parameters of $W_n^{(i)}$, which we denote by $Z_n^{(i)}$, obey $P_{e,n}^{(i)} = Z_n^{(i)}$. In the light of (16), (17), and Proposition 5, the $Z_n^{(i)}$ terms can then be recursively calculated as

$$\begin{aligned} Z_n^{(j)} &= Z_{n-1}^{(j)} Z_{n-m}^{(j)}, & Z_n^{(j+N(n-1))} &= Z_{n-1}^{(j)} + Z_{n-m}^{(j)} - Z_{n-1}^{(j)} Z_{n-m}^{(j)}, & j &\in \mathbb{N}_{n-m}, \\ Z_n^{(j)} &= Z_{n-1}^{(j)}, & j &\in \mathbb{N}_{n-1} \setminus \mathbb{N}_{n-m}. \end{aligned}$$

The case where W is not a BEC is a well-studied problem, where one approximates a suitable reliability measure for $W_n^{(i)}$ channels and uses this measure to choose the set \mathcal{A} . We refer the reader to [5] for an overview.

4. CHANNEL POLARIZATION

Channel polarization should be investigated by observing the evolution of the set $\{W_n^{(i)} : i \in \mathbb{N}_n\}$ as n increases. To track this evolution, we use the state vectors $\mathbf{s}_n^{(i)} \in \mathcal{S}_n$ assigned to $W_n^{(i)}$, because each $W_n^{(i)}$ is uniquely addressable by its $\mathbf{s}_n^{(i)}$.

4.1. Probabilistic Model for Channel Evolution

We define a random process $\{S_n\}$ and a random vector $\mathbf{S}_n = (S_1, S_2, \dots, S_n)$ obtained from the process $\{S_n\}$ where the state vectors $\mathbf{s}_n = (s_1, s_2, \dots, s_n)$, $\mathbf{s}_n \in \mathcal{S}_n$, of Section 3 are the realizations of \mathbf{S}_n . The process $\{S_n\}$ can be regarded as a tree process where \mathbf{s}_n form the branches of the tree, as is illustrated in Fig. 5 for $m = 2$. Because $|\mathcal{S}_n| = N = N(n)$, there are $N(n)$ different branches at tree level n . The process $\{S_n\}$ starts with the initial conditions $S_1 \in \{+, -\}$. At tree level n , $N(n)$ new branches emerge from $N(n-1)$ branches of level $n-1$. We assume that each branch is observed with identical probability

$$\Pr(\mathbf{S}_n = \mathbf{s}_n) = \frac{1}{N(n)}. \tag{36}$$

This, in turn, implies that each valid state transition of Fig. 3 between \mathbf{s}_{n-1} and \mathbf{s}_n has probability $N(n-1)/N(n)$. Consider the case $m = 1$, which coincides with Arıkan’s setup in [1], where there are two possible states: $S_n \in \{+, -\}$ and $|\mathcal{S}_n| = N(n) = 2^n$. Since transitions between \mathbf{S}_{n-1} and \mathbf{S}_n are valid if $S_n \in \{+, -\}$ and $S_{n-1} \in \{+, -\}$, each possible transition has probability $N(n-1)/N(n) = 1/2$. Consequently, the process $\{S_n\}$ is composed of independent realizations of Bernoulli(1/2) random variables: $\Pr(S_n = +) = \Pr(S_n = -) = 1/2$. On the other hand, when $m > 1$, there is memory in the state transition model, as is depicted in Fig. 3. Therefore, $\{S_n\}$ can be modeled as a Markov process with order $m-1$, i.e.,

$$\Pr(S_n | \mathbf{S}_{n-1}) = \Pr(S_n | S_{n-1}, S_{n-2}, \dots, S_{n-(m-1)}).$$

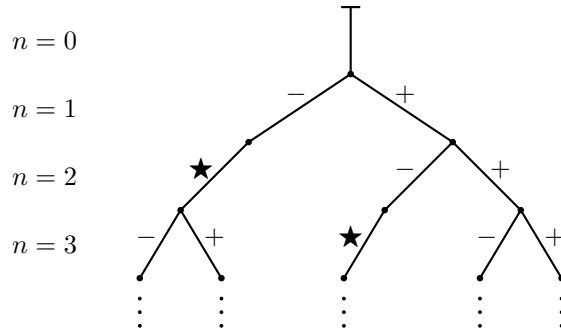


Fig. 5. Illustration of the evolution of $\{S_n\}$ as a tree for the case $m = 2$, where each branch is a state vector $s_n \in \mathcal{S}_n$.

Throughout the paper we find it easier to work with the random vector S_n , keeping in mind the Markovian property of the process $\{S_n\}$.

We define a random channel process $\{K_n\}$, driven by $\{S_n\}$, as $K_n = W_{s_1, s_2, \dots, s_n}$. The realizations of K_n are $k_n = W_{s_1, s_2, \dots, s_n}$, and they correspond to the binary-input channels $W_n^{(i)}$ with state vectors $s_n = (s_1, s_2, \dots, s_n) \in \mathcal{S}_n$.

In order to obtain a characterization for the process $\{K_n\}$, we fix $(s_1, s_2, \dots, s_{n-1})$ to be the state vector associated with $W_{n-1}^{(j)}$, $j \in \mathbb{N}_{n-m}$, and let $k_{n-1} = W_{n-1}^{(j)}$. We know by Proposition 4 that the state vector of $W_{n-m}^{(j)}$ is $(s_1, s_2, \dots, s_{n-m})$, so that $k_{n-m} = W_{n-m}^{(j)}$. From the operation of $\varphi_n: \mathcal{S}_{n-1} \rightarrow \mathcal{S}_n$ in Fig. 2, we observe that the state vectors of $W_n^{(j)}$ and $W_n^{(j+N(n-1))}$ are $(s_1, s_2, \dots, s_{n-1}, +)$ and $(s_1, s_2, \dots, s_{n-1}, -)$, respectively. From Proposition 5 we notice that $W_n^{(j)} = W_{n-m}^{(j)} \boxplus W_{n-1}^{(j)}$ and $W_n^{(j+N(n-1))} = W_{n-m}^{(j)} \boxminus W_{n-1}^{(j)}$. These, in turn, ensure that $k_n = k_{n-1} \boxplus k_{n-m}$ when $s_n = +$, and $k_n = k_{n-1} \boxminus k_{n-m}$ when $s_n = -$. Next, we fix $(s_1, s_2, \dots, s_{n-1})$ to be the state vector associated with $W_{n-1}^{(j)}$, $j \in \mathbb{N}_{n-1} \setminus \mathbb{N}_{n-m}$, and hence $k_{n-1} = W_{n-1}^{(j)}$. From the operation of $\varphi_n: \mathcal{S}_{n-1} \rightarrow \mathcal{S}_n$ we now know that the state vector of $W_n^{(j)}$ is $(s_1, s, \dots, s_{n-1}, \star)$, and Proposition 5 tells us that $W_n^{(j)} = \gamma(n)W_{n-1}^{(j)}$. Combining these facts, we have $k_n = \gamma(n)k_{n-1}$ if $s_n = \star$. The above analysis relates k_n to k_{n-1} and k_{n-m} for all $s_n \in \{+, -, \star\}$, which we formally present with the following recursion:

$$K_n = \begin{cases} K_{n-m} \boxplus K_{n-1} & \text{if } S_n = +, \\ K_{n-m} \boxminus K_{n-1} & \text{if } S_n = -, \\ \gamma(n)K_{n-1} & \text{otherwise,} \end{cases} \tag{37}$$

where $K_n = W$ for $n < 1$.

4.2. Polarization

We define the processes $\{I_n : n \geq 1\}$ and $\{J_n : n \geq 1\}$ where $I_n = I(K_n) \in [0, 1]$ and $J_n = J(K_n) \in [0, 1]$. In [1] Arıkan shows that I_n converges to a random variable I_∞ with $\Pr(I_\infty = 1) = I(W)$ and $\Pr(I_\infty = 0) = 1 - I(W)$. This result indicates that the synthesized binary-input channels $W_n^{(i)}$ become either error-free or useless. We show in the following theorem that the same is true for polar codes with higher-order memory as well.

Theorem 1. *For any fixed $m \geq 1$ and for some $\delta \in (0, 1)$ as n tends to infinity, the probability of $I_n \in (1 - \delta, 1]$ goes to $I(W)$ and the probability of having $I_n \in [0, \delta)$ goes to $1 - I(W)$.*

Proof. We investigate the polarization of $\{J_n\}$ towards 0 and 1, since it will imply the polarization of $\{I_n\}$ as well. We write $E[J_n] = \sum_{\mathbf{s}_n} \Pr(\mathbf{S}_n = \mathbf{s}_n) J_n = \left(\sum_{\mathbf{s}_n} J_n \right) / N(n)$ to denote the expected value of J_n and $\{E[J_n] : n \geq 1\}$ to denote the deterministic sequences obtained from $E[J_n]$. The following lemma will be crucial for the proof.

Lemma 1. *We have*

$$E[J_n] \geq \mu E[J_{n-1}] + (1 - \mu) E[J_{n-m}], \tag{38}$$

where $\mu = N(n-1)/N(n)$, and the above equality is achieved only if $J_{n-1} \in \{0, 1\}$ or $J_{n-m} \in \{0, 1\}$ holds for all $S_n \in \{+, -\}$.

We apply the following decimation operation on the sequence $\{E[J_n]\}$ to generate the subsequence $\{E[\widehat{J}_k] : k = 1, 2, \dots, \lfloor n/m \rfloor\}$:

$$E[\widehat{J}_k] = \min_{i \in \{0, 1, \dots, m-1\}} \{E[J_{km-i}]\}. \tag{39}$$

The elements of $\{E[\widehat{J}_k]\}$ are obtained by choosing the minimum of m consecutive and nonoverlapping elements of $\{E[J_n]\}$.

Lemma 2. *The sequence $\{E[\widehat{J}_k]\}$ is monotonically increasing in the sense that*

$$E[\widehat{J}_k] \geq E[\widehat{J}_{k-1}].$$

We know that $E[\widehat{J}_k]$ is bounded in $[0, 1]$, and, since $\{E[\widehat{J}_k]\}$ is monotonically increasing, from the monotone convergence theorem [6, p. 21] we conclude that there exists a unique limit for $\{E[\widehat{J}_k]\}$ such that

$$\lim_{k \rightarrow \infty} E[\widehat{J}_k] = \sup\{E[\widehat{J}_k]\}. \tag{40}$$

Next, we let $n = km - i$ in Lemma 1 to obtain

$$E[J_{km-i}] \geq \mu E[J_{km-(i+1)}] + (1 - \mu) E[J_{(k-1)m-i}]. \tag{41}$$

We fix i such that $E[J_{km-i}] = E[\widehat{J}_k]$ is satisfied. For any choice of i observe that $E[J_{(k-1)m-i}] \geq E[\widehat{J}_{k-1}]$ and $E[J_{km-(i+1)}] \geq \min\{E[\widehat{J}_k], E[\widehat{J}_{k-1}]\} \geq E[\widehat{J}_{k-1}]$ hold. Using these results in (41) gives

$$E[\widehat{J}_k] \geq \mu E[\widehat{J}_{k-1}] + (1 - \mu) E[\widehat{J}_{k-1}] \geq E[\widehat{J}_{k-1}]. \tag{42}$$

Therefore, the monotonic increase in $E[\widehat{J}_k]$ will continue until the inequality in Lemma 1 is achieved with equality. This fact, together with the convergence of $E[\widehat{J}_k]$, indicates that, conditioned on the event $\{S_n : S_n \in \{+, -\}\}$, either $\lim_{n \rightarrow \infty} J_{n-1} \in \{0, 1\}$ or $\lim_{n \rightarrow \infty} J_{n-m} \in \{0, 1\}$, and

$$\lim_{n \rightarrow \infty} J_n \in \{0, 1\}, \quad S_n \in \{+, -\}. \tag{43}$$

We see from the operation of $\varphi_n : S_{n-1} \rightarrow S_n$ in Fig. 2 that

$$\Pr(S_n \in \{+, -\}) = \frac{2N(n-m)}{N(n)} \geq 0, \tag{44}$$

which implies that the event $\{S_n : S_{n-1} \in \{+, -\}\}$ occurs infinitely many times as $n \rightarrow \infty$ and $\sum_{n \rightarrow \infty} \Pr(S_{n-1} \in \{+, -\})$ diverges. Consequently, by using the first Borel–Cantelli lemma [7, p. 36] we conclude that

$$\lim_{n \rightarrow \infty} \Pr(J_n \in \{0, 1\}) = 1.$$

One-to-one correspondence between J_n and I_n implies

$$\lim_{n \rightarrow \infty} \Pr(I_n \in \{0, 1\}) = 1,$$

and having $E[I_n] = I(W)$ results in

$$\lim_{n \rightarrow \infty} \Pr(I_n = 1) = I(W),$$

and

$$\lim_{n \rightarrow \infty} \Pr(I_n = 0) = 1 - I(W)$$

which completes the proof. \triangle

4.3. A Typicality Result

In this section we use the method of types to investigate the state vectors \mathbf{s}_n obtained from the realizations of the process $\{S_n\}$. We let $s \in \{+, -, \star\}$ and write $P_{\mathbf{s}_n}^{(s)}, P_{\mathbf{s}_n}^{(s)} \in [0, 1]$, to denote the type (frequency) of s in \mathbf{s}_n as

$$P_{\mathbf{s}_n}^{(s)} = \#(\mathbf{s}_n | s) / n$$

where $\#(\mathbf{s}_n | s)$ denotes the number of times that the symbol s occurs in \mathbf{s}_n . From the state transition diagram of Fig. 3, we see that as n gets large, we have $P_{\mathbf{s}_n}^{(\star)} = (m - 1)P_{\mathbf{s}_n}^{(-)}$, because each $-$ state in \mathbf{s}_n is followed by $m - 1$ occurrences of state \star . Since the remaining states in \mathbf{s}_n will be $+$, we must have $P_{\mathbf{s}_n}^{(+)} = 1 - mP_{\mathbf{s}_n}^{(-)}$, indicating $P_{\mathbf{s}_n}^{(+)} \in [0, 1]$, $P_{\mathbf{s}_n}^{(-)} \in [0, 1/m]$, and $P_{\mathbf{s}_n}^{(\star)} \in [0, (m - 1)/m]$. As it turns out, depending on $P_{\mathbf{s}_n}^{(s)}$, not all realizations of $\{S_n\}$ are observed with the same probability. This is explained with the following theorem.

Theorem 2. *As n gets large, except for a vanishing fraction of $\mathbf{s}_n \in \mathcal{S}_n$, and for some $\varepsilon \in (0, 1)$, we have*

$$\begin{aligned} |P_{\mathbf{s}_n}^{(-)} - p^-| &\leq \varepsilon, \\ |P_{\mathbf{s}_n}^{(+)} - p^+| &\leq \varepsilon, \\ |P_{\mathbf{s}_n}^{(\star)} - p^\star| &\leq \varepsilon \end{aligned}$$

where $p^- = (\phi - 1) / [1 + m(\phi - 1)]$, $p^\star = (m - 1)p^-$, and $p^+ = 1 - mp^-$.

We can consider p^+ , p^- , and p^\star as the frequencies of states $+$, $-$, and \star in \mathbf{s}_n , respectively, that one typically observes as n gets large.

Proof of Theorem 2. The proof is based on the method of types [8]. We let $q \in [0, 1/m]$ and define

$$\mathcal{T}_n^{(q)} = \{\mathbf{s}_n : P_{\mathbf{s}_n}^{(-)} = q\}. \tag{45}$$

The $\mathcal{T}_n^{(q)}$ is a type class, and it consists of \mathbf{s}_n having $nq \in [0, n/m]$ occurrences of state $-$. For all $m \geq 1$, there are at most $n + 1$ different such type classes. However, the number $|\mathcal{S}_n|$ of all possible \mathbf{s}_n increases exponentially in n as $|\mathcal{S}_n| = N = O(\phi^n)$. The method of types ensures the existence of a type class with exponentially many elements. Our aim is to find this type class. Recalling that each \mathbf{s}_n is observed with probability $1/N$, the probability of observing a given \mathbf{s}_n in $\mathcal{T}_n^{(q)}$ is

$$\Pr(\mathbf{s}_n \in \mathcal{T}_n^{(q)}) = \frac{|\mathcal{T}_n^{(q)}|}{N}.$$

Lemma 3. *We have*

$$|\mathcal{T}_n^{(q)}| < 2^{n(G(m,q)+o(1))}, \tag{46}$$

where

$$G(m, q) = [1 - (m - 1)q]H\left(\frac{q}{1 - (m - 1)q}\right)$$

and H is the binary entropy function.

$G(m, q)$ is a concave function of $q \in [0, 1/m]$. We establish a similarity between $\partial G(m, q)/\partial q$ and $F(m, \rho)$ in (35). The following statement is a direct consequence of this result.

Lemma 4. *The function $G(m, q)$ attains its maximum when $q = p^-$, and its maximum value is*

$$G(m, p^-) = \log \phi.$$

Consequently, for every $\mathcal{T}_n^{(q)}$ with $|q - p^-| > 0$ there exists a $D(q, p^-) > 0$ such that

$$D(q, p^-) \triangleq G(m, p^-) - G(m, q) = \log \phi - G(m, q).$$

Using the above fact in (46) results in

$$|\mathcal{T}_n^{(q)}| \leq \phi^n 2^{n(-D(q, p^-) + o(1))}.$$

From the above result and the fact that $N = O(\phi^n)$ we obtain

$$\Pr(\mathbf{s}_n \in \mathcal{T}_n^{(q)}) \leq 2^{-n(D(q, p^-) + o(1))}. \tag{47}$$

Equation (47) shows that, depending on $D(q, p^-)$ and in turn q , the probabilities of some type classes decay exponentially in n . The following proposition is an outcome of this fact.

Proposition 8. *As n tends to infinity, $D(q, p^-)$ converges to 0 with probability 1.*

The above proposition implies the convergence of q to p^- as well, because $D(q, p^-) = 0$ only if $q = p^-$. Thus, among all $\mathcal{T}_n^{(q)}$, one observes the ones with $|q - p^-| \leq \varepsilon$ with probability 1.

4.4. Rate of Polarization

We define the Bhattacharyya process $\{Z_n\}$ where $Z_n = Z(K_n)$ is the Bhattacharyya parameter of the random channel K_n . By using the channel evolution model in (37), this process can be expressed as

$$Z_n \begin{cases} = Z_{n-1}Z_{n-m} & \text{if } S_n = +, \\ \leq Z_{n-1} + Z_{n-m} - Z_{n-1}Z_{n-m} & \text{if } S_n = -, \\ = Z_{n-1} & \text{otherwise,} \end{cases} \tag{48}$$

where $Z_n = Z(W)$ for $n < 1$.

Theorem 3. *For any $\varepsilon \in (0, 1)$ there exists an n such that for $\beta < p^+$ we have*

$$\Pr(Z_n \leq 2^{-\phi^{n\beta}}) \geq I(W) - \varepsilon. \tag{49}$$

Proof. We consider another process $\{\widehat{Z}_n\}$, driven by $\{S_n\}$, so that for $i = 1, 2, \dots, n_0$, $n_0 < n$, we have $\widehat{Z}_i = Z_i$, and for $i > n_0$, \widehat{Z}_i obeys

$$\widehat{Z}_i = \begin{cases} \widehat{Z}_{i-1}\widehat{Z}_{i-m} & \text{if } S_n = +, \\ \widehat{Z}_{i-1} + \widehat{Z}_{i-m} - \widehat{Z}_{i-1}\widehat{Z}_{i-m} & \text{if } S_n = -, \\ \widehat{Z}_{i-1} & \text{otherwise.} \end{cases} \tag{50}$$

Comparing (48) and (50), we observe that Z_n is stochastically dominated by \widehat{Z}_n in the sense that for some $f_n \in (0, 1)$, $\Pr(Z_n \leq f_n) \geq \Pr(\widehat{Z}_n \leq f_n)$. For the proof it will suffice to show that $\Pr(\widehat{Z}_n \leq f_n) \geq I(W) - \varepsilon$ holds for $f_n = 2^{-\phi^{n\beta}}$ and $\beta < p_+$.

In [9, Lemma 1] the authors derive an upper bound on \widehat{Z}_n for the case $m = 1$ by using the frequency of state $+$ in the realizations of $\{S_{n_0+1}, S_{n_0+2}, \dots, S_n\}$ and the fact that Z_{n_0} gets arbitrarily close to 0 with probability $I(W)$ when n_0 is large enough. The following lemma is a generalization of this approach for arbitrary $m \geq 1$.

Lemma 5. *For some $\zeta \in (0, 1)$ and $\gamma \in (0, 1)$ define the events*

$$C_{n_0}(\zeta) = \{Z_{n_0} \leq \zeta\},$$

$$D_{n_0}^n(\gamma) = \{\#\left((S_{n_0+1}, \dots, S_n) \mid +\right) \geq \gamma(n - n_0)\}.$$

We have

$$\widehat{Z}_n \leq 2^{-\phi^{(\gamma-\varepsilon)(n-n_0)}}, \quad C_{n_0}(\zeta) \cap D_{n_0}^n(\gamma).$$

From the convergence of Z_n to Z_∞ with probability $\Pr(Z_\infty = 0) = I(W)$, we know that for any $\varepsilon \in (0, 1)$ there exists a fixed n_0 such that

$$\Pr(C_{n_0}(\zeta)) \geq I(W) - \varepsilon.$$

Next, from Theorem 2 we infer that when $m \ll n - n_0$, we have

$$\Pr(D_{n_0}^n(\gamma)) \geq 1 - \varepsilon, \quad \gamma \geq p^+ - \varepsilon. \tag{51}$$

This results from the fact that the probability of observing $+$ in $\{S_{n_0+1}, \dots, S_n\}$ approaches p^+ when $n - n_0$ is much larger than the memory m of the process $\{S_n\}$.

Choosing $n_0 = n\varepsilon$ and using Lemma 5 gives

$$\Pr(\widehat{Z}_n \leq 2^{-\phi^{n(p^+ - 2\varepsilon)(1-\varepsilon)}}) \geq (1 - \varepsilon)(I(W) - \varepsilon).$$

Since $\varepsilon \in (0, 1)$ can be chosen arbitrarily close to 0, the above result indicates that

$$\Pr(\widehat{Z}_n \leq 2^{-\phi^{n\beta}}) \geq I(W) - \varepsilon$$

for $\beta < p^+$. \triangle

Let us analyze the implications of Theorem 3 on the block-decoding error probability P_e of $\{\mathcal{C}_n^{(m)}\}$. It states that for $I(W) - \varepsilon$ fraction of $W_n^{(i)}$ the corresponding Bhattacharyya parameters will be bounded as $Z_n^{(i)} \leq 2^{-\phi^{n\beta}}$ for $\beta < p^+$. We have $P_e \leq \sum_{i=1}^N Z_n^{(i)} \leq N2^{-\phi^{n\beta}} = O(2^{-\phi^{n\beta}})$. Since the code length of $\{\mathcal{C}_n^{(m)}\}$ scales as $N = O(\phi^n)$, we also see that $P_e = O(2^{-N^\beta})$ holds for $\beta < p^+$.

The term p^+ is plotted in Fig. 6 as m increases from 1 to 50. We see that p^+ equals 0.5 when $m = 1$, which coincides with the bound for the exponent of polar codes presented by Arıkan and Telatar in [3]. As m increases from 1 to 50, p^+ (and thus the achievable exponent) decreases. This decrease is more steep for small values of m .

In order to fully characterize the asymptotic performance of $\{\mathcal{C}_n^{(m)}\}$, one needs to provide a converse bound on β , which may be a difficult task. We believe that for the case $m > 1$, the achievable β for $\{\mathcal{C}_n^{(m)}\}$ may show a dependency on the rate $R \in [0, 1]$ chosen for the code, a phenomenon that does not exist when $m = 1$ (see [10]). In order to explain our conjecture, consider the process $\{\widehat{Z}_n\}$ in (50) which we use to obtain an achievable bound on β as $\beta < p^+$. Our proof is based on the observation that once the realizations of \widehat{Z}_{n_0} are sufficiently close to 0,

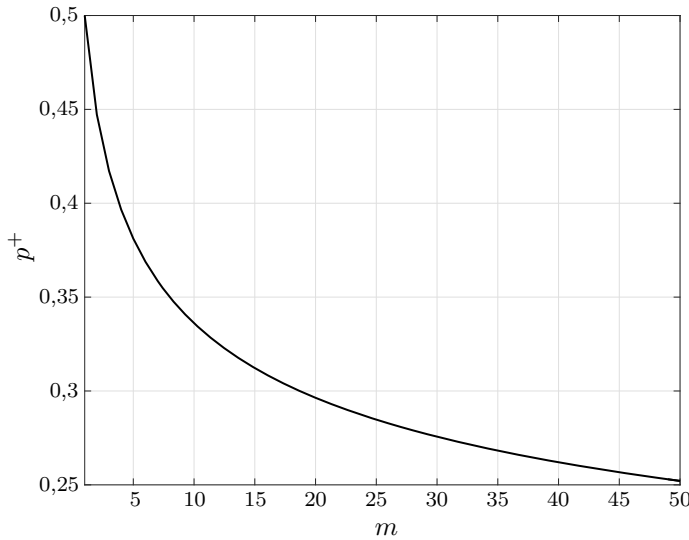


Fig. 6. Achievable exponent $\beta < p^+$ versus m .

which happens with probability $I(W)$, the scaling of Z_n is mostly determined by the number of occurrences of state $+$ in $\{S_{n_0+1}, S_{n_0+2}, \dots, S_n\}$. From Theorem 2 we know that one typically observes $(n - n_0)p^+$ occurrences of $+$ in $\{S_{n_0+1}, S_{n_0+2}, \dots, S_n\}$; therefore, the value of $\log Z_n$ decreases $(n - n_0)p^+$ times with the same speed as the code length: $\log \hat{Z}_n = \log \hat{Z}_{n-1} + \log \hat{Z}_{n-m}$, scaling as $\log Z_n = -\phi^{(n-n_0)p^+} = -\phi^{n(1-\varepsilon)p^+}$. This results in the achievable exponent $\beta < p^+$. However, when $m > 1$, the value of $\log \hat{Z}_n$ may also decrease with a faster rate as compared with that of the code length. To see this, consider the case $(S_{n-1}, S_{n-2}, \dots, S_{n-(m-1)}) = (\star, \star, \dots, \star)$ and $S_n = +$, where we have $\hat{Z}_{n-1} = \hat{Z}_{n-2} = \dots = \hat{Z}_{n-(m-1)}$ and $\log \hat{Z}_n = \log \hat{Z}_{n-1} + \log \hat{Z}_{n-m} = \log \hat{Z}_{n-1}^2$. There may be times when $\log Z_n$ decreases with a faster rate as $\log \hat{Z}_n = \log Z_{n-1}^2$ instead of $\log \hat{Z}_n = \log \hat{Z}_{n-1} + \log \hat{Z}_{n-m}$, and this may yield a higher achievable β . In order to quantify this, we need to know not only the number of times that state $+$ occurs in $\{S_n\}$ but also the number of times that a state $+$ in $\{S_n\}$ is preceded by \star states. Therefore, we need to refine Theorem 2 in terms of the number of transitions between states $+$, $-$, and \star as well. This might be a difficult but important problem, whose solution will provide a full characterization of the asymptotic polarization performance of $\{\mathcal{C}_n^{(m)}\}$, and we leave it as a future work.

5. COMPLEXITY

We consider a single-core processor with random access memory and investigate the time complexity of the encoding and decoding of $\{\mathcal{C}_n^{(m)}\}$. Let χ_n^E denote the complexity of encoding the information vector \mathbf{u}_N to encoded bits \mathbf{x}_N . We take complexity of each XOR operation as 1 unit. By inspection of Fig. 1, we have

$$\chi_n^E = \chi_{n-1}^E + \chi_{n-m}^E + N_{n-m}, \quad n, m \geq 1, \tag{52}$$

where $\chi_1^E = 1$ and $\chi_0^E = \chi_{-1}^E = \dots = \chi_{1-m}^E = 0$.

Similarly, let χ_n^D be the complexity of decoding the inputs of $W_n^{(i)}$ channels when SCD is employed. We take the complexity of computing the LR relationships in (32) as 1 unit. Note that calculating the LR in (33) does not require any operation. From Fig. 1, we have

$$\chi_n^D = \chi_{n-1}^D + \chi_{n-m}^D + 2N_{n-m}, \quad n, m \geq 1, \tag{53}$$

where $\chi_0^D = \chi_{-1}^D = \dots = \chi_{1-m}^D = 0$.

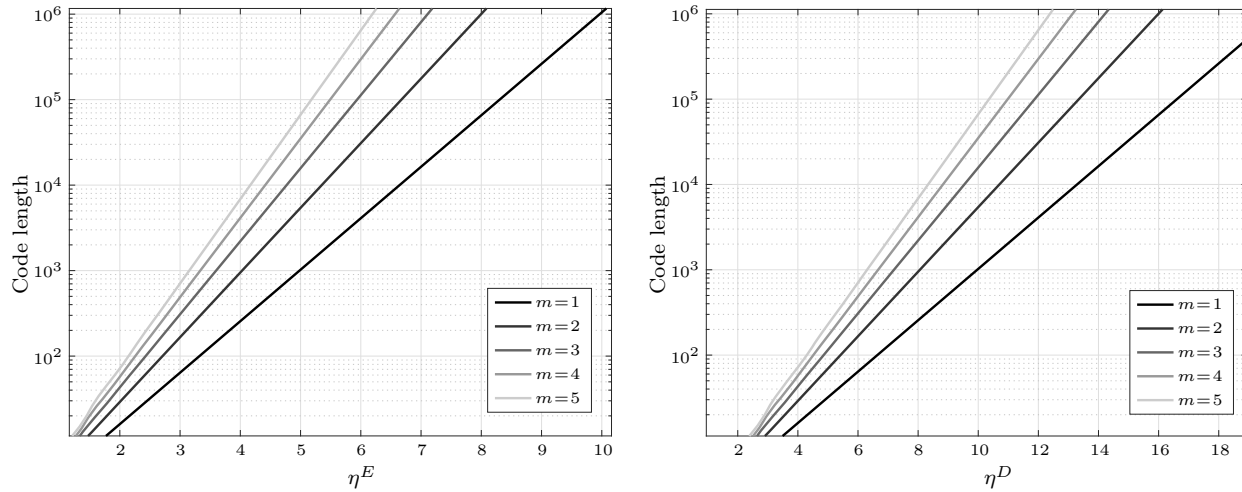


Fig. 7. Scaling of η^E and η^D with code length for $m = 1, 2, 3, 4$, and 5 .

The recursions in (52) and (53) are cumbersome to deal with. To observe the scaling behavior of χ_n^E and χ_n^D in m , we define

$$\eta^E \triangleq \frac{\chi_n^E}{N}, \quad \eta^D \triangleq \frac{\chi_n^D}{N} \quad (54)$$

where η^E and η^D are the encoding and decoding complexities per bit, respectively. Figure 7 demonstrates the scaling of η^E and η^D with code length for $m = 1, 2, 3, 4$ and 5 , where it can be seen that both η^E and η^D decrease as m grows from 1 to 5 . These trends observed in Fig. 7 are valid for $m > 5$ as well. Therefore, when the code length is fixed, it is possible to decrease the η^E and η^D values by increasing m . The same result also holds for encoding and decoding complexities. This property implies that when complexity is fixed, one can increase the code length by opting for a larger m . The memory order m brings a flexibility to the design of $\{\mathcal{C}_n^{(m)}\}$ by allowing for the adjustment of the complexity per code length.

6. PERFORMANCE AND SIMULATIONS

The asymptotic exponent analysis of Section 4 does not reflect the finite-length performance of $\{\mathcal{C}_n^{(m)}\}$. In this section we discuss the performance of $\{\mathcal{C}_n^{(m)}\}$ for practical code lengths and present our simulation results.

We have shown that the encoding and decoding complexities of $\{\mathcal{C}_n^{(m)}\}$ decrease with increasing m . This reduction is somehow expected, because as m grows, the fraction of synthesized channels with \star states increases, and these channels do not contribute to the encoding and decoding complexities. However, not being polarized with respect to the previous level, those same channels undermine the polarization performance. Therefore, complexity reduction inevitably brings performance degradation.

When complexity is fixed, the code length of $\{\mathcal{C}_n^{(m)}\}$ increases in m . The unevenness in the code length and the number of bit channels makes it hard to obtain a fair performance comparison for $\{\mathcal{C}_n^{(m)}\}$. This problem can be overcome by fixing the complexity per bit instead. By doing so, one can compare the polarization performance of $\{\mathcal{C}_n^{(m)}\}$ for different m by plotting the Bhattacharyya parameters of bit channels against the code rate, and by observing how fast the Bhattacharyya curves fall as the code rate is lowered.

In the simulations we assign different values to η^D on a BEC with erasure probability ε . We also vary ε to bring out the rate-dependent nature of $\{\mathcal{C}_n^{(m)}\}$, as is conjectured at the end of Section 4. Two sets of codes are considered. The first one contains the original polar codes ($m = 1$) with $n = 10, 11$, and 12 . These codes have block lengths $N = 1024, 2048, 4096$ and decoding complexity per bit $\eta^D = 10, 11, 12$, respectively. The second set consists of codes with $m = 2$ and $n = 18, 19, 20, 21$. The latter codes have $N = 6765, 10946, 17711, 28657$ and $\eta^D = 10.2, 10.8, 11.3, 11.9$, respectively. Note that it is impossible to have exactly the same decoding complexity per bit for codes obtained with $m = 1$ and $m = 2$, because this parameter takes discrete values. Therefore, we consider the performance of codes that have similar decoding complexities per bit.

In Fig. 8 we plot the sorted Bhattacharyya parameters of bit channels against the code rate. Specifically, we compare the performances of the codes as ε is increased from 0.6 to 0.95 and inspect the steepness of the Bhattacharyya curves as the code rate is lowered. Examining Fig. 8a, we see that when $\varepsilon = 0.6$, the original polar codes $\{\mathcal{C}_n^{(1)}\}$ perform better than those obtained with $m = 2$, i.e., $\{\mathcal{C}_n^{(2)}\}$. Figure 8b demonstrates that when $\varepsilon = 0.7$, the code with $m = 2$ and $\eta = 10.8$ lies between the original polar codes with $\eta = 10$ and $\eta = 11$. Similarly, the code with $m = 2$ and $\eta = 11.9$ interpolates between the original polar codes with $\eta = 11$ and $\eta = 12$. The latter interpolation is not uniform, since the code with $m = 2$ and $\eta = 11.9$ provides a slightly better polarization performance compared to the code with $m = 1$ and $\eta = 11$. Nevertheless, the two codes with $m = 2$ and $\eta = 10.8, \eta = 11.9$ operate in a region which is not attainable by the original polar codes, and they thereby become competitive. Figures 8c and 8d demonstrate the performances for $\varepsilon = 0.8$ and $\varepsilon = 0.9$, respectively. All of the considered codes with $m = 2$ interpolate between the original polar codes. As ε tends to 1 the codes with $m = 2$ perform better than the original polar codes. This is observed in Fig. 8d as the performance of the code with $m = 2$ and $\eta = 10.8$ matches the performance of polar code with $\eta = 11$. In Fig. 8e we see that for $\varepsilon = 0.95$ the performance of the codes with $m = 2$ are better than the original polar codes, because the former are better polarized per decoding complexity per bit. Below is a summary of our findings.

1. For $\eta^D > 12$, $\{\mathcal{C}_n^{(1)}\}$ perform better than $\{\mathcal{C}_n^{(2)}\}$. When $m = 1$, $\eta^D = 12$ is obtained with a code length $N = 4096$. Thus, when the decoding complexity per bit is fixed, the original polar codes with lengths larger than 4096 provide a better polarization performance compared to those obtained with $m = 2$.
2. For $\varepsilon \leq 0.6$, $\{\mathcal{C}_n^{(1)}\}$ perform better than $\{\mathcal{C}_n^{(2)}\}$.
3. For $\eta \leq 12$ and as ε is increased from 0.6 to 1, $\{\mathcal{C}_n^{(2)}\}$ start to be competitive with respect to $\{\mathcal{C}_n^{(1)}\}$.
4. In the regime $\eta \leq 12$ and $\varepsilon \approx 1$, $\{\mathcal{C}_n^{(2)}\}$ provide better performance per decoding complexity per bit.

The main reason for the above results is the nonuniform polarization which happens when $m \geq 2$. This nonuniformity takes place because the bit-channels with $-$ states are penalized and put into $m - 1$ levels of waiting states. On the other hand, the code favors the channels with $+$ states by spending the complexity budget to populate and polarize more channels from them. This favoring mechanism makes the codes with $m \geq 2$ perform worse above a certain code rate, which manifests itself when $\varepsilon > 0.6$ for $m = 2$, because the penalization of channels with $-$ states causes a sacrifice on a certain fraction of bit-channels when the polarization levels are completed (and not extended to infinity; i.e., the asymptotic case). It seems that for $\varepsilon > 0.6$, uniform polarization (i.e., not favoring any channel) works best. However, as ε moves above 0.6, the performance of $\{\mathcal{C}_n^{(2)}\}$ starts to match that of the original polar codes, since the favoring mechanism now pays off. Eventually, as ε tends to 1, $\{\mathcal{C}_n^{(2)}\}$ provides better performance. In the regime where $\eta^D > 12$ or $\varepsilon \leq 0.6$, there exists a code-rate region where the polarization performance of $\{\mathcal{C}_n^{(2)}\}$ matches $\{\mathcal{C}_n^{(1)}\}$, but this

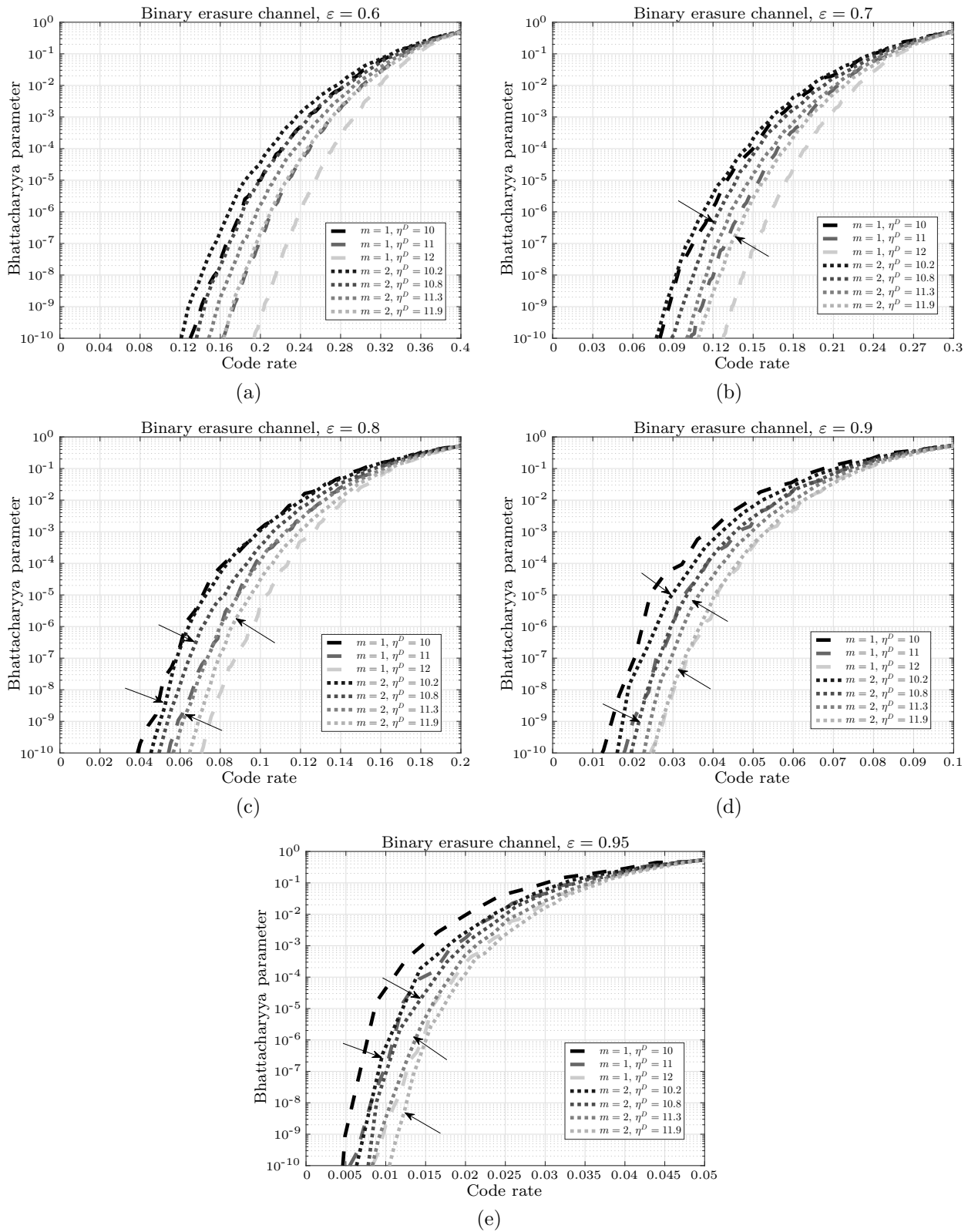


Fig. 8. Polarization performance of $\{\mathcal{C}_n^{(m)}\}$ for $m = 1, 2$, while keeping decoding complexity per bit η^D fixed. The performance of codes with $m = 2$ that offer a competitive polarization performance compared to the original polar codes are indicated with arrows.

is again the impractical low code-rate region where the Bhattacharyya parameters are very small. When $\varepsilon > 0.6$, this advantageous code-rate region becomes more relevant, since the corresponding Bhattacharyya parameters have practical values.

7. DISCUSSION AND CONCLUSION

We have introduced a method to design a class of code sequences $\{\mathcal{C}_n^{(m)}; n \geq 1, m \geq 1\}$ with code length $N = O(\phi^n)$, $\phi \in (1, 2]$, and memory order m . The design of $\{\mathcal{C}_n^{(m)}\}$ is based on the channel polarization idea of Arikan [1], and $\{\mathcal{C}_n^{(m)}\}$ coincides with the polar codes presented by Arikan when $m = 1$. We have shown that $\{\mathcal{C}_n^{(m)}\}$ achieves the symmetric capacity of BDMCs for arbitrary but fixed m . We have obtained an achievable bound on the asymptotic polarization performance of $\{\mathcal{C}_n^{(m)}\}$ as scaled with m and showed that the encoding and decoding complexities of $\{\mathcal{C}_n^{(m)}\}$ decrease with increasing m .

While the original polar codes provide better polarization performance than the codes obtained with $m \geq 2$, there are special cases where $\{\mathcal{C}_n^{(m)}; m \geq 2\}$ may work competitively, or even better. For example, in the regime where $0.6 < \varepsilon < 1$ and $10 \leq \eta^D \leq 12$, there exist codes with $m = 2$ that provide a competitive polarization performance per decoding complexity per bit. Moreover, as ε approaches 1, i.e., when the code rate is close to 0, $\{\mathcal{C}_n^{(2)}\}$ provides a performance advantage over the original polar codes. This fact implies that the codes with higher-order memory may find application in systems that operate in very low signal-to-noise ratios, e.g., transmission over very noisy channels or deep space communications where the received signal is extremely weak. Further investigation is needed on these channels to understand the performance of $\{\mathcal{C}_n^{(m)}; m \geq 2\}$, which we leave as a future work.

APPENDIX

Proof of Proposition 1. We have

$$J(W^-) = \log \frac{2}{1 + Z(W^-)}, \quad J(W^+) = \log \frac{2}{1 + Z(W^+)}.$$

By using (16) and (17), we obtain

$$\begin{aligned} J(W^+) + J(W^-) &\geq \log \frac{2}{1 + Z(W')Z(W'')} + \log \frac{2}{1 + Z(W') + Z(W'') - Z(W')Z(W'')} \\ &= \log \frac{4}{1 + Z(W') + Z(W'') + w(W', W'')Z(W')Z(W'')}, \end{aligned} \quad (55)$$

where $w(W', W'') = Z(W') + Z(W'') - Z(W')Z(W'') \leq 1$, so that

$$J(W^+) + J(W^-) \geq \log \frac{2}{1 + Z(W')} + \log \frac{2}{1 + Z(W'')} = J(W') + J(W''). \quad (56)$$

In order to have $J(W^+) + J(W^-) = J(W') + J(W'')$, the equalities in (55) and (56) must be achieved. From (17) we know that the equality in (55) is achieved only if $Z(W') \in \{0, 1\}$ or $Z(W'') \in \{0, 1\}$ or if W' and W'' are BECs. When $(Z(W'), Z(W'')) \in (0, 1)^2$, we have $w(W', W'') < 1$, and the inequality in (56) is always strict whether or not W' and W'' are BECs. Consider the cases $Z(W') = 1$ or $Z(W'') = 1$. Then we have $w(W', W'') = 1$, and the equalities in (55) and (56) are achieved. When $Z(W') = 0$, we have $J(W') = 1$ and $J(W^+) + J(W^-) = J(W') + J(W'')$, and the case $Z(W'') = 0$ follows from the symmetry in (55) and (56). Hence the equalities in (55) and (56) are both achieved only if $Z(W') \in \{0, 1\}$ or $Z(W'') \in \{0, 1\}$, or alternatively only if $J(W') \in \{0, 1\}$ or $J(W'') \in \{0, 1\}$. \triangle

Proof of Proposition 3. From the operation of φ_n in Definition 1 we obtain $\mathcal{S}_1 = \{+, -\}$ such that $\mathbf{s}_1^{(1)} = (+)$ and $\mathbf{s}_1^{(2)} = (-)$, indicating that $\mathbf{s}_1^{(1)}$ and $\mathbf{s}_1^{(2)}$ are unique. Proof is by induction. Assume that $\mathbf{s}_{n-1}^{(j)} \in \mathcal{S}_{n-1}$ are unique. Let $j \in \mathbb{N}_{n-m}$ and consider $\mathbf{s}_{n-1}^{(j)}$ from which one obtains the distinct vectors $\mathbf{s}_n^{(j)}$ and $\mathbf{s}_n^{(j+N(n-1))}$ by, respectively, appending $+$ and $-$. Next, let $j \in \mathbb{N}_{n-1} \setminus \mathbb{N}_{n-m}$. Then $\mathbf{s}_n^{(j)}$ are obtained by appending \star to $\mathbf{s}_{n-1}^{(j)}$, which, by assumption, are unique. Combining these results, we see that for all $j \in \mathbb{N}_n$ the vectors $\mathbf{s}_n^{(j)} \in \mathcal{S}_n$ are different from each other. \triangle

Proof of Proposition 4. Consider the operation of φ_{n-1} presented in Fig. 2, where $\mathbf{s}_{n-2}^{(k)} = (s_1, s_2, \dots, s_{n-2})$, $k \in \mathbb{N}_{n-2}$, holds at level $n - 1$. Next, consider the operation of φ_{n-2} , where one has $\mathbf{s}_{n-3}^{(k)} = (s_1, s_2, \dots, s_{n-3})$ for $k \in \mathbb{N}_{n-3}$. By induction through $\varphi_{n-2}, \varphi_{n-3}, \dots, \varphi_{n-(m-1)}$ we conclude that $\mathbf{s}_{n-m}^{(j)} = (s_1, s_2, \dots, s_{n-m})$, $j \in \mathbb{N}_{n-m}$. \triangle

Proof of Proposition 6. 1. For $m > 1$ we have $F(m, 1) = -1 < 0$ and $F(m, 2) = 2^{m-1} - 1 \geq 0$, so that there exists at least one real root in $(1, 2]$. Proof is by contradiction. Let $\rho_1, \rho_2 \in (1, 2]$ be two real roots of $F(m, \rho)$. From (35) we have

$$\rho_1^{m-1}(\rho_1 - 1) = 1, \tag{57}$$

$$\rho_2^{m-1}(\rho_2 - 1) = 1. \tag{58}$$

Let $\rho_1 < \rho_2$, so that $\rho_2^{m-1} > \rho_1^{m-1}$ and $\rho_2 - 1 > \rho_1 - 1 > 0$, implying $\rho_2^{m-1}(\rho_2 - 1) > 1$ if $\rho_1^{m-1}(\rho_1 - 1) = 1$, which contradicts (58). Carrying out a similar analysis for $\rho_1 < \rho_2$ also contradicts (58), which means $\rho_1 = \rho_2 = \phi$.

2. Assume that ρ is a complex root of $F(m, \rho)$ with $\sqrt{\rho\rho^*} = \sigma > 1$, where $*$ denotes the conjugation. Since the coefficients of $F(m, \rho)$ are real, its complex roots must be in conjugate pairs. From (35),

$$\begin{aligned} \rho^{m-1}(\rho - 1) &= 1, \\ \rho^{*m-1}(\rho^* - 1) &= 1. \end{aligned}$$

Multiplying the above equations, we obtain

$$\sigma^{2(m-1)}(\sigma^2 - 2\text{Re}(\rho) + 1) = 1,$$

or

$$\sigma^{2(m-1)}(\sigma^2 - 2\sigma\alpha + 1) = 1, \tag{59}$$

where $0 \leq \alpha < 1$. In turn, for any ρ , σ must be a root of

$$g(\sigma, \alpha) = \sigma^{2(m-1)}(\sigma^2 - 2\sigma\alpha + 1) - 1. \tag{60}$$

Observe that when σ is fixed, $g(\sigma, \alpha)$ is decreasing in α . We also have

$$\frac{\partial g(\sigma, \alpha)}{\partial \sigma} = 2(m-1)\sigma^{2(m-1)-1}(\sigma^2 - 2\sigma\alpha + 1) + \sigma^{2(m-1)}(2\sigma - 2\alpha).$$

From (59) observe that $(\sigma^2 - 2\sigma\alpha + 1) > 0$, and since $(2\sigma - 2\alpha) > 0$ for $\sigma > 1$, we have $\partial g(\sigma, \alpha)/\partial \sigma > 0$, and $g(\sigma, \alpha)$ is increasing in σ . But ϕ is a root of $g(\sigma, \alpha)$ with $\alpha = 1$, and thus $g(\phi, 1) = 0$. Since $g(\sigma, \alpha)$ is decreasing in α , we have $g(\phi, \alpha) \geq 0$, and $g(\sigma, \alpha) = 0$ is achieved only if $\sigma < \phi$, because $g(\sigma, \alpha)$ is increasing in σ .

3. Observe that for some $\rho \in (1, 2]$ we have $\partial F(m, \rho)/\partial \rho > 0$, so that $F(m, \rho)$ is increasing in ρ . When ρ is fixed, $F(m, \rho)$ is also increasing in m . Assume that $\rho_1, \rho_2 \in (1, 2]$ are real roots of $F(m_1, \rho)$ and $F(m_2, \rho)$, respectively, where $m_1, m_2 \geq 1$. Then $f(m_1, \rho_1) < f(m_2, \rho_1)$ if $m_2 > m_1$, and $f(m_1, \rho_1) = f(m_2, \rho_2) = 0$ is satisfied only if $\rho_1 < \rho_2$. \triangle

Proof of Lemma 1. Let $J_n^{(i)} = J(W_n^{(i)})$ denote the symmetric cut-off rate of $W_n^{(i)}$. From Proposition 5 we know that for $j \in \mathbb{N}_{n-m}$ we have $W_n^{(j)} = W_{n-1}^{(j)} \boxplus W_{n-m}^{(j)}$ and $W_n^{(j+N(n-1))} = W_{n-1}^{(j)} \boxplus W_{n-m}^{(j)}$. Proposition 1 indicates that these transforms increase the sum cut-off rate: $J_n^{(j)} + J_n^{(j+N(n-1))} \geq J_{n-1}^{(j)} + J_{n-1}^{(j)}$, where the equality is achieved only if $J_{n-1}^{(j)} \in \{0, 1\}$ or $J_{n-m}^{(j)} \in \{0, 1\}$ holds. For $j \in \mathbb{N}_{n-1} \setminus \mathbb{N}_{n-m}$, from Proposition 5 we have $J_n^{(j)} = \gamma(n)J_{n-1}^{(j)}$, which implies $J_n^{(j)} = J_{n-1}^{(j)}$. Combining the above results gives

$$\sum_{i \in \mathbb{N}_n} J_n^{(i)} \geq \sum_{j \in \mathbb{N}_{n-1}} J_n^{(j)} + \sum_{k \in \mathbb{N}_{n-m}} J_n^{(k)}, \tag{61}$$

where the equality is achieved only if $J_{n-1}^{(j)} \in \{0, 1\}$ or $J_{n-m}^{(j)} \in \{0, 1\}$ for all $j \in \mathbb{N}_{n-m}$. In the probabilistic domain of Section 4, (61) is equivalent to

$$\sum_{s_n \in \mathcal{S}_n} J_n \geq \sum_{s_{n-1} \in \mathcal{S}_{n-1}} J_{n-1} + \sum_{s_{n-m} \in \mathcal{S}_{n-m}} J_{n-m},$$

with equality only if $J_{n-1} \in \{0, 1\}$ or $J_{n-m} \in \{0, 1\}$ holds for all $S_n \in \{+, -\}$. Dividing both sides of the above inequality by $N(n)$ and using $E[J_n] = \frac{1}{N(n)} \sum_{s_n \in \mathcal{S}_n} J_n$, we obtain

$$E[J_n] \geq \frac{N(n-1)}{N(n)} E[J_{n-1}] + \frac{N(n-m)}{N(n)} E[J_{n-m}].$$

Noticing that $\frac{N(n-1)}{N(n)} = \mu(n)$ and $\frac{N(n-m)}{N(n)} = 1 - \mu(n)$ completes the proof. \triangle

Proof of Lemma 2. From (38) we have

$$\begin{aligned} E[J_n] &\geq \mu E[J_{n-1}] + (1 - \mu) E[J_{n-m}] \\ &\geq \min\{E[J_{n-1}], E[J_{n-m}]\}. \end{aligned} \tag{62}$$

Let us define the set

$$\mathcal{E}_k^{(m)} \triangleq \{E_{km}, E_{km-1}, \dots, E_{km-(m-1)}\}.$$

By the definition in (39) we have $E[\widehat{J}_k] = \min \mathcal{E}_k^{(m)}$. Proof is by induction. We use (62) to upper bound the elements of $\mathcal{E}_k^{(m)}$ with respect to $\min \mathcal{E}_{k-1}^{(m)} = E[\widehat{J}_{k-1}]$. Let $n = km - (m - 1)$ and use (62) to obtain

$$E_{km-(m-1)} \geq \min\{E_{(k-1)m}, E_{(k-1)m-(m-1)}\} \geq \min \mathcal{E}_{k-1}^{(m)}$$

For $i = 2, 3, \dots, m - 1$, assume that

$$E_{km-(m-i)} \geq \min \mathcal{E}_{k-1}^{(m)}$$

is true. Next, let $n = km - (m - (i + 1))$ in (62) to write

$$E_{km-(m-(i+1))} \geq \min\{E_{km-(m-i)}, E_{(k-1)m-(m-(i+1))}\}.$$

By the assumption, $E_{km-(m-i)} \geq \min \mathcal{E}_{k-1}^{(m)}$, and by the definition, $E_{(k-1)m-(m-(i+1))} \geq \min \mathcal{E}_{k-1}^{(m)}$ holds, indicating

$$E_{km-(m-(i+1))} \geq \min \mathcal{E}_{k-1}^{(m)}.$$

Combining the above results tells us that for $i = 1, 2, \dots, m$ we have $E_{km-(m-i)} \geq \min \mathcal{E}_{k-1}^{(m)} = E[\widehat{J}_{k-1}]$, which implies that $E[\widehat{J}_k] \geq E[\widehat{J}_{k-1}]$. \triangle

Proof of Lemma 3. In order to bound $|\mathcal{T}_n^{(q)}|$, we decompose $\mathcal{T}_n^{(q)}$ into two different sets

$$\begin{aligned} \mathcal{T}_n^{(a,q)} &\triangleq \{\mathbf{s}_n : P_{\mathbf{s}_n}^{(-)} = q, s_n = +\}, \\ \mathcal{T}_n^{(b,q)} &\triangleq \{\mathbf{s}_n : P_{\mathbf{s}_n}^{(-)} = q, s_n \neq +\}, \end{aligned}$$

where we have $T_n^{(q)} = \mathcal{T}_n^{(a,q)} \cup \mathcal{T}_n^{(b,q)}$. Recall that each state $-$ in \mathbf{s}_n is followed by $m - 1$ occurrences of state \star . In turn, $\mathcal{T}_n^{(a,q)}$ consists of \mathbf{s}_n having $k = nq$, $0 \leq k \leq n/m$, occurrences of the vector $\mathbf{a} = (-, \underbrace{\star, \star, \dots, \star}_{m-1 \text{ times}})$ and $n - km$ occurrences of state $+$. By combinatorial analysis we have

$$|\mathcal{T}_n^{(a,q)}| = \binom{n - (m - 1)k}{k}.$$

An element of $\mathcal{T}_n^{(b,q)}$ consists of $k - 1$ occurrences of \mathbf{a} , an occurrence of $\mathbf{b} = (-, \underbrace{\star, \star, \dots, \star}_p)$,

$1 \leq p < m - 1$, and $n - mk - (p + 1)$ occurrences of state $+$. The vector \mathbf{b} can only occur in the last $p + 1$ entries in \mathbf{s}_n , and it will be completed to a vector \mathbf{a} if we prolong the channel combining operation by $m - 1 - p \leq m$ more levels. Therefore,

$$|\mathcal{T}_n^{(b,q)}| \leq \binom{n + m - (m - 1)k}{k}.$$

For some $c, d \in \mathbb{Z}$ with $c < d$, we have

$$\binom{d}{c} = \frac{d}{d - c} \binom{d - 1}{c} \leq d \binom{d - 1}{c}. \tag{63}$$

Using (63), we obtain

$$\begin{aligned} \binom{n + m - (m - 1)k}{k} &\leq (n + m) \binom{n + (m - 1) - (m - 1)k}{k} \\ &< (n + m)^2 \binom{n + (m - 2) - (m - 1)k}{k} \\ &\vdots \\ &< (n + m)^m \binom{n - (m - 1)k}{k}. \end{aligned}$$

Then

$$\begin{aligned} |\mathcal{T}_n^{(q)}| &= |\mathcal{T}_n^{(a,q)}| + |\mathcal{T}_n^{(b,q)}| \\ &< [1 + (n + m)^m] \binom{n - (m - 1)k}{k} \\ &< [1 + (n + m)]^m \binom{n - (m - 1)k}{k} \\ &= 2^{nB(m,n)} \binom{n - (m - 1)k}{k}, \end{aligned} \tag{64}$$

where $B(m, n) = [m \log(1 + n + m)] / n = o(1)$. Next, we use the upper bound $\binom{n}{k} \leq 2^{nH(k/n)}$ in [8] to have

$$\binom{n - (m - 1)k}{k} \leq 2^{n(1 - (m - 1)(k/n))H(\frac{k/n}{1 - (m - 1)(k/n)}} = 2^{nG(m,q)}. \tag{65}$$

Combining (64) and (65), we obtain the desired bound: $|\mathcal{T}_n^{(q)}| < 2^{n(G(m,q) + B(m,n))} = 2^{n(G(m,q) + o(1))}$.

Proof of Lemma 4. We have

$$G(m, q) = [1 - (m - 1)q]H\left(\frac{q}{1 - (m - 1)q}\right).$$

We know that, for $q \in [0, 1/m]$, $H(q/[1 - (m - 1)q])$ is concave in q and $1 - (m - 1)q$ is linear in q , which ensure that $G(m, q)$ is concave in q . Let q^* denote the maximizer of $G(m, q)$. The maximum of $H(q/[1 - (m - 1)q])$ is attained when $q/[1 - (m - 1)q] = 1/2$, or equivalently when $q = 1/(m + 1)$, and because $1 - (m - 1)q$ is decreasing in q , we have $q^* \in [0, 1/(m + 1)]$. Moreover,

$$\frac{\partial G(m, q)}{\partial q} = (m - 1) \log(1 - (m - 1)q) + \log q - m \log(1 - mq).$$

Setting $\partial G(m, q)/\partial q|_{q=q^*} = 0$ gives

$$(m - 1) \log(1 - (m - 1)q^*) + \log q^* = m \log(1 - mq^*). \quad (66)$$

Rearranging the above equation,

$$m \log \frac{(1 - (m - 1)q^*)}{1 - mq^*} + \log \frac{q^*}{1 - mq^*} = \log \frac{(1 - (m - 1)q^*)}{1 - mq^*}. \quad (67)$$

Define

$$\eta = \frac{1 - (m - 1)q^*}{1 - mq^*},$$

so that for $q^* \in [0, 1/(m + 1)]$ we have $\eta \in [1, 2]$. Substituting η into (67) yields

$$m \log \eta + \log(\eta - 1) = \log \eta,$$

or alternatively,

$$\eta^m(\eta - 1) = \eta.$$

Dividing both sides of the above relationship by η , we obtain

$$\eta^m - \eta^{m-1} - 1 = 0. \quad (68)$$

But the above polynomial is the same as (35). Consequently, from part 1 of Proposition 6, we conclude that $\eta = \phi$, i.e., $[1 - (m - 1)q^*]/(1 - mq^*) = \phi$, and hence, $q^* = 1/[1 + m(\phi - 1)] = p^-$. Next, we evaluate the maximum of $G(m, q)$ attained at $q = q^*$:

$$G(m, q^*) = -q^* \log \frac{q^*}{1 - (m - 1)q^*} + (mq^* - 1) \log \frac{1 - mq^*}{1 - (m - 1)q^*}. \quad (69)$$

From (66) we observe that

$$\log \frac{q^*}{1 - (m - 1)q^*} = m \log \frac{1 - mq^*}{1 - (m - 1)q^*}. \quad (70)$$

Putting (70) in (69) gives

$$G(m, q^*) = \log \frac{1 - (m - 1)q^*}{1 - mq^*} = \log \phi. \quad \triangle$$

Proof of Proposition 8. We define a typical set $\mathcal{T}_n^{(q,\varepsilon)}$ as

$$\mathcal{T}_n^{(q,\varepsilon)} = \{s_n : P_{s_n}^{(-)} = q, D(q, p^-) \leq \varepsilon\}.$$

The probability that $\mathcal{T}_n^{(q)}$ is not typical is

$$\begin{aligned} 1 - \Pr(\mathcal{T}_n^{(q,\varepsilon)}) &= \sum_{\Pr(D(q,p^-) > \varepsilon)} \Pr(\mathcal{T}_n^{(q)}) \\ &\stackrel{(a)}{\leq} \sum_{\Pr(D(q,p^-) > \varepsilon)} 2^{-n(D(q,s_-)+o(1))} \\ &\leq \sum_{\Pr(D(q,p^-) > \varepsilon)} 2^{-n(\varepsilon+o(1))} \\ &\stackrel{(b)}{\leq} (n+1)2^{-n(\varepsilon+o(1))} \\ &= 2^{-n(\varepsilon+o(1))}. \end{aligned} \tag{71}$$

In the above derivation, (a) follows from (47) and (b) from the fact that there exist at most $n+1$ different type classes having $\Pr(D(q, s_-) > \varepsilon)$. Equation (71) indicates that $\sum_{n \rightarrow \infty} \Pr(D(q, s_-) \geq \varepsilon)$ converges. Thus, the expected number of occurrences of the event $D(q, s_-) > \varepsilon$ for all n is finite. By using the first Borel–Cantelli Lemma [7, p. 59], we conclude that $D(q, s_-)$ converges to 0 with probability 1. \triangle

Proof of Lemma 5. Conditioned on the event $D_{n_0}^n(\gamma) = \#\{(s_{n_0+1}, \dots, s_n) | +\} \geq \gamma(n - n_0)$, there exist at least $\gamma(n - n_0)$ occurrences of state $+$ in $\{S_{n_0+1}, S_{n_0+2}, \dots, S_n\}$. From (50), we have $\hat{Z}_n \leq \hat{Z}_{n-1}$ when $S_n = +$, and $Z_n \geq Z_{n-1}$ when $S_n \neq +$. Moreover, Z_n is increasing in Z_{n-1} when S_n is fixed. Consequently, if we fix \hat{Z}_m , the largest value of \hat{Z}_n will occur if $\{S_{n_0+1}, S_{n_0+2}, \dots, S_n\}$ has the following realization:

$$\overbrace{\{\mathbf{a}, \mathbf{a}, \dots, \mathbf{a}\}}^{(1-\gamma)(n-n_0)/m \text{ times}}, \underbrace{\{+, +, \dots, +\}}_{\gamma(n-n_0) \text{ times}}$$

where $\mathbf{a} = (-, \star, \star, \dots, \star)$. In order to upper bound \hat{Z}_n , we assume that the above realization

has occurred for $\{S_{n_0+1}, S_{n_0+2}, \dots, S_n\}$. During consecutive runs of $+$, the value of $\log \hat{Z}_n$ increases with the same recursion as the code length in (1), i.e., $\log \hat{Z}_n = \log \hat{Z}_{n-1} + \log \hat{Z}_{n-m}$. This recursion happens $\gamma(n - m)$ times, and since the code length obeying the same recursion scales as $\phi^{\gamma(n-m)}$, $\phi \in (1, 2]$, we have

$$\log \hat{Z}_n = \phi^{\gamma(n-n_0)} \log \hat{Z}_k, \tag{72}$$

where $k = n_0 + (1 - \gamma)(n - m)$. During the consecutive runs of \mathbf{a} , the value of \hat{Z}_i does not change with respect to \hat{Z}_{i-1} when $S_i = \star$, and increases as $\hat{Z}_i = \hat{Z}_{i-1} + \hat{Z}_{i-m} - \hat{Z}_{n-1}\hat{Z}_{i-m}$ when $S_i = -$. By the construction of $\{S_{n_0+1}, S_{n_0+2}, \dots, S_n\}$, each state $-$ is preceded by $m - 1$ occurrences of \star . Therefore, if $S_i = -$, then we have $(S_{i-1}, S_{i-2}, \dots, S_{i-(m-1)}) = (\star, \star, \dots, \star)$, indicating $\hat{Z}_{i-1} = \hat{Z}_{i-2} = \dots = \hat{Z}_{i-(m-1)}$. During each occurrence of state $-$ in \mathbf{a} , we see the recursion $\hat{Z}_{i-1} + \hat{Z}_{i-m} - \hat{Z}_{i-1}\hat{Z}_{i-m} = 2\hat{Z}_{i-1} - \hat{Z}_i^{(i)}$, or equivalently $1 - \hat{Z}_i = (1 - \hat{Z}_i^{(i)})^2$. The latter takes place $(1 - \gamma)(n - n_0)$ times, resulting in $1 - \hat{Z}_k = (1 - \hat{Z}_{n_0})^{2(1-\gamma)(n-n_0)}$ and $\hat{Z}_k = 1 - (1 - \hat{Z}_{n_0})^{2(1-\gamma)(n-n_0)}$. Employing the inequality $\log x \leq x - 1, x \in [0, 1]$, we obtain

$$\log \hat{Z}_k \leq -(1 - \hat{Z}_{n_0})^{2(1-\gamma)(n-n_0)}. \tag{73}$$

Using (73) in (72) gives

$$\begin{aligned} \log \widehat{Z}_n &= -\phi^{\gamma(n-n_0)}(1-Z_{n_0})^{2(1-\gamma)(n-n_0)} \\ &\leq -\phi^{\gamma(n-n_0)}(1-Z_{n_0})^{2(n-n_0)} \\ &= -\phi^{(\gamma-\varepsilon)(n-n_0)} \left((1-Z_{n_0})^2 \phi^\varepsilon \right)^{(n-n_0)}. \end{aligned}$$

Now choose $\zeta \in (0, 1)$ so that $\zeta \leq 1 - \phi^{-\frac{\varepsilon}{2}}$. Conditioned on $C_{n_0}(\zeta) = \{Z_{n_0} \leq \zeta\}$, we have $(1 - Z_{n_0})^2 \phi^\varepsilon \geq 1$, resulting in

$$\log_2 \widehat{Z}_n \leq -\phi^{(\gamma-\varepsilon)(n-m)}, \quad C_{n_0}(\zeta) \cap D_{n_0}^n(\gamma),$$

which proves the lemma. \triangle

REFERENCES

1. Arıkan, E., Channel Polarization: A Method for Constructing Capacity-Achieving Codes for Symmetric Binary-Input Memoryless Channels, *IEEE Trans. Inform. Theory*, 2009, vol. 55, no. 7, pp. 3051–3073.
2. Arıkan, E., Channel Combining and Splitting for Cutoff Rate Improvement, *IEEE Trans. Inform. Theory*, 2006, vol. 52, no. 2, pp. 628–639.
3. Arıkan, E. and Telatar, I.E., On the Rate of Channel Polarization, in *Proc. 2009 IEEE Int. Sympos. on Information Theory (ISIT'2009)*, Seoul, Korea, June 28–July 3, 2009, pp. 1493–1495.
4. Korada, S.B., Şaşıoğlu, E., and Urbanke, R., Polar Codes: Characterization of Exponent, Bounds, and Constructions, *IEEE Trans. Inform. Theory*, 2010, vol. 56, no. 12, pp. 6253–6264.
5. Tal, I. and Vardy, A., How to Construct Polar Codes, *IEEE Trans. Inform. Theory*, 2013, vol. 59, no. 10, pp. 6562–6582.
6. Bartle, R.G., *The Elements of Real Analysis*, New York: Wiley, 1995, 2nd ed.
7. Billingsley, P., *Probability and Measure*, New York: Wiley, 1995, 3rd ed.
8. Cover, T.M. and Thomas, J.A., *Elements of Information Theory*, Hoboken, NJ: Wiley, 2006, 2nd ed.
9. Afşer, H. and Deliç, H., On the Channel-Specific Construction of Polar Codes, *IEEE Commun. Lett.*, 2015, vol. 19, no. 9, pp. 1480–1483.
10. Hassani, S.H. and Urbanke, R., On the Scaling of Polar Codes: I. The Behavior of Polarized Channels, in *Proc. 2010 IEEE Int. Sympos. on Information Theory (ISIT'2010)*, Austin, TX, USA, June 13–18, 2010, pp. 874–878.