Rate-Dependent Analysis of the Asymptotic Behavior of Channel Polarization

S. Hamed Hassani, Ryuhei Mori, Toshiyuki Tanaka, and Rüdiger L. Urbanke

***Abstract*—We consider the asymptotic behavior of the polarization process in the large block-length regime when transmission takes place over a binary-input memoryless symmetric channel W. In particular, we study the asymptotics of the cumulative distribution $\mathbb{P}(Z_n \leq z)$, where $\{Z_n\}$ is the Bhattacharyya process associated with W, and its dependence on the rate of transmission. On the basis of this result, we characterize the asymptotic behavior, as well as its dependence on the rate, of the block error probability of polar codes using the successive cancellation decoder. This refines the original asymptotic bounds by Arıkan and Telatar. Our results apply to general polar codes based on $\ell \times \ell$ kernel matrices. We also provide asymptotic lower bounds on the block error probability of polar codes using the maximum *a posteriori* (MAP) decoder. The MAP lower bound and the successive cancellation upper bound coincide when $\ell = 2$, but there is a gap for $\ell > 2$.**

I. Introduction

A. Polar Codes

Polar codes, introduced by Arıkan [1], are a family of codes that provably achieve the capacity of binary-input memoryless symmetric (BMS) channels using low-complexity encoding and decoding algorithms. Since their invention, there has been a large body of work that has analyzed (see, e.g., [2]–[11]) and extended (see, e.g., [12]–[20]) these codes.

The construction of polar codes is based on an $\ell \times \ell$ matrix G, with entries in $\{0, 1\}$, called the kernel matrix. Besides being invertible, the matrix G should have the property that none of its column permutations is upper triangular [13]. We call a matrix G with such properties a *polarizing* matrix and in the following, whenever we speak of a kernel matrix G, we assume that G is polarizing.

Manuscript received October 02, 2011; revised June 08, 2012; accepted August 19, 2012. Date of publication November 20, 2012; date of current version March 13, 2013. S. H. Hassani was supported by the Swiss National Science Foundation under Grant 200021-121903. R. Mori was supported by the Grant-in-Aid for Scientific Research for the Japan Society for the Promotion of Science (JSPS) Fellows (22·5936), MEXT, Japan. T. Tanaka was supported by the Grant-in-Aid for Scientific Research (C) (22560375), JSPS, Japan. This paper was presented in part at the Graduate School of Informatics, Kyoto University, Kyoto, Japan [6], in part at the 2010 IEEE International Symposium on Information Theory [7], [8], and in part at the 2010 IEEE Information Theory Workshop, Dublin, Ireland [11].

S. H. Hassani and R. L. Urbanke are with the School of Computer and Communication Science, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland (e-mail: seyehamed.hassani@epfl.ch; rudiger.urbanke@epfl.ch).

R. Mori and T. Tanaka are with the Department of Systems Science, Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan (e-mail: rmori@sys.i.kyoto-u.ac.jp; tt@i.kyoto-u.ac.jp).

Communicated by E. Arıkan, Associate Editor for Coding Theory.

Digital Object Identifier 10.1109/TIT.2012.2228295

The rows of the generator matrix of a polar code with block length $N = \ell^n$ are chosen from the rows of the matrix

$$G^{\otimes n} \triangleq \overbrace{G \otimes G \otimes \cdots \otimes G}^{n}$$

where $\otimes$ denotes the Kronecker product. For the case $\ell = 2$ and the choice $G = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$, Reed–Muller (RM) codes also fall into this category. However, the crucial difference between polar codes and RM codes lies in the choice of the rows. For RM codes, the rows of the largest weights are chosen, whereas for polar codes, the choice depends on the channel and is made using a method called channel polarization. We briefly review this method and explain how polar codes are constructed from it. We refer the reader also to [1], [5], and [13] for a detailed discussion.

B. Channel Polarization

Let W be a BMS channel, and let $\mathcal{X} = \{0, 1\}$ denote its input alphabet, $\mathcal{Y}$ the output alphabet, and $W(y \,|\, x)$ the transition probabilities. Also, let $I(W) \in [0, 1]$ denote the mutual information between the input and output of W with uniform distribution on the input. The capacity of a BMS channel W is equal to $I(W)$. Also, the Bhattacharyya parameter of W, denoted by $Z(W)$, is defined as

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

It provides upper and lower bounds of the error probability $P_{\mathrm{e}}(W)$ in estimating the channel input x on the basis of the channel output y via the maximum-likelihood (ML) decoding of $W(y|x)$ as follows [5], [22, Ch. 4]:

$$\frac{1}{2}\left(1 - \sqrt{1 - Z(W)^2}\right) \leq P_{\mathrm{e}}(W) \leq \frac{1}{2} Z(W). \tag{1}$$

It is also related to the capacity $I(W)$ via

$$Z(W) + I(W) \geq 1,$$
$$[Z(W)]^2 + [I(W)]^2 \leq 1$$

both proved in [1]. In short, for $n \in \mathbb{N}$, the method of channel polarization takes $N = \ell^n$ copies of a BMS channel W and combines them by using the kernel matrix G to make a new set of ℓ^n channels $\{W_{\ell^n}^{(i)}\}_{1 \leq i \leq \ell^n}$. As $n \to \infty$, the set

0018-9448/$31.00 © 2012 IEEE

$\{W_{\ell^n}^{(i)}\}_{1\le i\le \ell^n}$ tends to have extremal properties. We explain in more detail the method of channel polarization through the following three steps.

Step 1 (Channel Splitting): Let $\mathcal{W}$ denote the class of BMS channels. Let us define a channel transform $W \to (W^0, W^1, \ldots, W^{\ell-1})$, called channel splitting, that maps $\mathcal{W}$ to $(\overbrace{\mathcal{W}, \mathcal{W}, \ldots, \mathcal{W}}^{\ell})$. In other words, channel splitting is a transform which takes a BMS channel W as input and outputs ℓ BMS channels W^j, $0 \le j \le \ell - 1$. The channels W^j are constructed using the channel W and matrix G, according to the following rule: Consider a random row vector $U_0^{\ell-1} = (U_0, \ldots, U_{\ell-1})$ that is uniformly distributed over $\{0,1\}^\ell$. Let $X_0^{\ell-1} = U_0^{\ell-1}G$, where the arithmetic is in GF(2). Also, let $Y_0^{\ell-1}$ be the result of passing each component of $X_0^{\ell-1}$ through an independent copy of W (i.e., Y_i is the outcome of passing X_i through an independent copy of W). We thus define the channel between $U_0^{\ell-1}$ and $Y_0^{\ell-1}$ by the transition probabilities

$$W_\ell(y_0^{\ell-1} \,|\, u_0^{\ell-1}) \stackrel{\triangle}{=} \prod_{i=0}^{\ell-1} W(y_i \,|\, x_i) = \prod_{i=0}^{\ell-1} W(y_i \,|\, (u_0^{\ell-1}G)_i). \tag{2}$$

The channel $W^j : \{0,1\} \to \mathcal{Y}^\ell \times \{0,1\}^j$ is defined as the BMS channel with input u_j, output $(y_0^{\ell-1}, u_0^{j-1})$ and transition probabilities

$$W^j(y_0^{\ell-1}, u_0^{j-1} \,|\, u_j) = \frac{1}{2^{\ell-1}} \sum_{u_{j+1}^{\ell-1}} W_\ell(y_0^{\ell-1} \,|\, u_0^{\ell-1}). \tag{3}$$

Here and hereafter, u_i^j denotes the subvector $(u_i, \ldots, u_j)$. An intuitive explanation behind the definition of W^j is as follows: Pick uniformly at random one of the 2^ℓ possible realizations of the vector $U_0^{\ell-1}$ and let it be denoted by $u_0^{\ell-1} = (u_0, \ldots, u_{\ell-1})$. Construct the vector $x_0^{\ell-1} = u_0^{\ell-1}G$ and send the ℓ components of $x_0^{\ell-1}$ through ℓ parallel (and independent) copies of the channel W, and finally, let $y_0^{\ell-1}$ denote the output (i.e., y_i is the result of passing x_i through an independent copy of W). It is easy to see that the channel between $u_0^{\ell-1}$ and $y_0^{\ell-1}$ is precisely the channel $W_\ell(y_0^{\ell-1} \,|\, u_0^{\ell-1})$ defined in (2). Fig. 1 gives a schematic representation of this channel. Thus, given the vector $y_0^{\ell-1}$, the optimal way to infer about the value of $u_0^{\ell-1}$ is via ML decoding of $W_\ell(y_0^{\ell-1} \,|\, u_0^{\ell-1})$. Now, besides having access to $y_0^{\ell-1}$, assume that a genie also gives us the values of the bits $u_0, \ldots, u_{j-1}$ and asks us to decide on the value of u_j based on the observed vector $(y_0^{\ell-1}, u_0^{j-1})$. A little thought shows that the optimal way to do this is ML decoding of the values of u_j by using the transition probabilities $W^j(y_0^{\ell-1}, u_0^{j-1} \,|\, u_j)$ defined in (3). In other words, W^j is precisely the channel between u_j and $(y_0^{\ell-1}, u_0^{j-1})$ when we do not have any information about the value of $u_{j+1}, \ldots, u_{\ell-1}$ (i.e., they are modeled as independent and identically distributed (i.i.d.) random variables with a uniform distribution). Fig. 2 gives a schematic representation of the channel W^j. Finally, a noteworthy point to repeat is that the actual implementation of the channel splitting transform $W \to (W^0, W^1, \ldots, W^{\ell-1})$ requires ℓ independent copies of W to generate $W^0, \ldots, W^{\ell-1}$. Furthermore, by applying the chain rule for mutual information, one can show that this transform preserves capacity [1], [13]

$$\sum_{j=0}^{\ell-1} I(W^j) = \ell I(W). \tag{4}$$

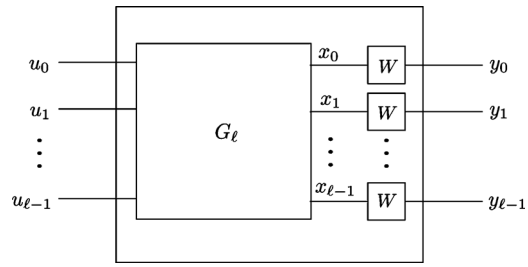

Fig. 1. Schematic representation of the channel between the random vectors $U_0^{\ell-1}$ and $Y_0^{\ell-1}$.

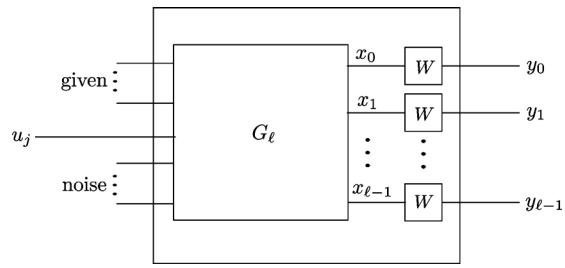

Fig. 2. Schematic representation of the channel W^j. This is the channel that the bit u_j "sees" when the value of $u_0, \ldots, u_{j-1}$ together with $y_0^{\ell-1}$ is given as output and the bits $u_{j+1}, \ldots, u_{\ell-1}$ are treated as noise (i.e., there is no information about their value and we assume their value is chosen uniformly at random).

Step 2 (Infinite ℓ-ary Tree): Consider an infinite ℓ-ary tree with the root node placed at the top. In this tree, each vertex has ℓ children and there are ℓ^n vertices at level n. Assume that we label these vertices from left to right from 1 to ℓ^n. Here, we intend to assign to each vertex of the tree a BMS channel. We do this by a recursive procedure. Assign to the root node the channel W itself. Now consider the channel splitting transform $W \to (W^0, W^1, \ldots, W^{\ell-1})$ and from left to right, assign W^0 to $W^{\ell-1}$ to the children of the root node. In general, if Q is the channel that is assigned to vertex v, we assign Q^0 to $Q^{\ell-1}$, from left to right, respectively, to the children of the node v. In this way, we recursively assign a channel to all the vertices of the tree. Fig. 3 shows the first two levels of the ℓ-ary tree when $\ell = 2$. Let $W_{\ell^n}^{(i)}$ denote the channel that is assigned to vertex with label i at level n of the tree, $1 \le i \le \ell^n$. As a result, one can equivalently relate the channel $W_{\ell^n}^{(i)}$ to W via the following procedure: let the ℓ-ary representation of $i-1$ be $b_1 b_2 \cdots b_n$, where b_1 is the most significant digit. Then, we have

$$W_{\ell^n}^{(i)} = (((W^{b_1})^{b_2})^{\cdots})^{b_n}.$$

As an example, assuming $i = 7$, $n = 3$, and $\ell = 2$, we have $W_8^{(7)} = ((W^1)^1)^0$.

Step 3 (Polarization Property): The channels $\{W_{\ell^n}^{(i)}\}_{1\le i\le \ell^n}$ have the property that (see[1] and [13]), as n grows large, a fraction close to $I(W)$ of the channels have capacity close to 1 (or

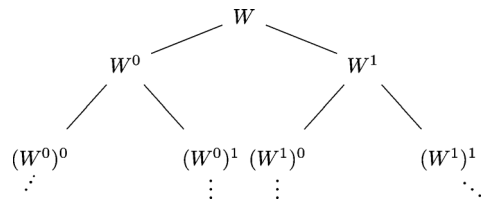

Fig. 3. Infinite ℓ-ary tree and the channels assigned to it for $\ell = 2$.

Bhattacharyya parameter close to 0); and a fraction close to $1 - I(W)$ of the channels have capacity close to 0 (or Bhattacharyya parameter close to 1). In other words, as n grows large, the channels $\{W_{\ell^n}^{(i)}\}_{1\le i\le \ell^n}$ tend to become polarized to one of the following extremal situations: an almost perfect channel (capacity is very close to 1) or a very noisy channel (capacity is very close to 0). The basic idea behind polar codes is to use those channels that have capacity close to 1 (or equivalently have Bhattacharyya parameter close to 0) for information transmission. Accordingly, given the rate $R < I(W)$ and block length $N = \ell^n$, the rows of the generator matrix of a polar code of block length N correspond to a subset of the rows of the matrix $G^{\otimes n}$ whose indices are chosen with the following rule: Choose a subset of size NR of the channels $\{W_{\ell^n}^{(i)}\}_{1\le i\le \ell^n}$ with the least values for the Bhattacharyya parameter and choose the rows $G^{\otimes n}$ with the indices corresponding to those of the channels. For example, if the channel $W_{\ell^n}^{(i)}$ is chosen, then the jth row of $G^{\otimes n}$ is selected, where the ℓ-ary representation of $j-1$ is the digit-reversed version of that of $i-1$. We decode using a successive cancellation (SC) decoder. This algorithm decodes the bits one-by-one in a prescribed order that is closely related to how the row indices of $G^{\otimes n}$ are chosen.

C. Problem Formulation and Relevant Work

Let $\mathcal{I}$ be the set of indices of the NR channels in the set $\{W_{\ell^n}^{(i)}\}_{1\le i\le \ell^n}$ with the least values for the Bhattacharyya parameter. Let $\mathbb{P}_{\mathrm{e}}^{\mathrm{SC}}(N,R)$ and $\mathbb{P}_{\mathrm{e}}^{\mathrm{MAP}}(N,R)$ denote the average block error probabilities of the SC and of the maximum *a posteriori* (MAP) decoders, respectively, with block length N and rate R. For the SC decoder, we have [1], [13],

$$\max_{i\in\mathcal{I}} \frac{1}{2}\left(1-\sqrt{1-Z(W_{\ell^n}^{(i)})^2}\right) \le \mathbb{P}_{\mathrm{e}}^{\mathrm{SC}}(N,R) \le \sum_{i\in\mathcal{I}} Z(W_{\ell^n}^{(i)}). \tag{5}$$

This relation shows that the distribution of the Bhattacharyya parameters of the channels $\{W_{\ell^n}^{(i)}\}_{1\le i\le \ell^n}$ plays a fundamental role in the analysis of polar codes. More precisely, for $n \in \mathbb{N} \stackrel{\triangle}{=} \{0, 1, 2, \ldots\}$ and $0 < z < 1$, we are interested in analyzing the asymptotic behavior of

$$F(n,z) = \frac{\#\{i : Z(W_{\ell^n}^{(i)}) \le z\}}{\ell^n} \tag{6}$$

where $\#A$ denotes the number of elements of the set A. There is an entirely equivalent probabilistic description of (6): Define the "polarization" process [2], [13] of the channel W as a channel-valued discrete-time stochastic process $\{W_n\}_{n\in\mathbb{N}}$ with $W_0 = W$ and

$$W_{n+1} = W_n^{B_n} \tag{7}$$

where $\{B_n\}_{n\in\mathbb{N}}$ is a sequence of i.i.d. random variables with distribution $\mathbb{P}(B_0 = j) = \frac{1}{\ell}$ for $j \in \{0, 1, \ldots, \ell-1\}$. Hence, assuming the value of the process $\{W_n\}_{n\in\mathbb{N}}$ at time n is Q and B_n takes the value $j \in \{0, 1, \ldots, \ell-1\}$, W_{n+1} will be equal to Q^j (recall from Section I-B the definition of channel splitting transform $Q \to (Q^0, Q^1, \ldots, Q^{\ell-1})$). In other words, the process begins at the root node of the infinite ℓ-ary tree introduced above, and in each step, it chooses one of the ℓ children of the current node with uniform probability. So at time n, the process $\{W_n\}_{n\in\mathbb{N}}$ outputs one of the ℓ^n channels at level n of the tree uniformly at random. The Bhattacharyya process $\{Z_n\}_{n\in\mathbb{N}}$ of the channel W is defined from the polarization process as $Z_n \stackrel{\triangle}{=} Z(W_n)$. In this setting, we have

$$\mathbb{P}(Z_n \le z) = F(n,z). \tag{8}$$

It was shown in [2] and [13] that the Bhattacharyya process $\{Z_n\}_{n\in\mathbb{N}}$ converges almost surely to a $\{0,1\}$-valued random variable Z_∞ with $\mathbb{P}(Z_\infty = 0) = I(W)$. Our objective is to investigate the asymptotic behavior of $\mathbb{P}(Z_n \le z)$. The analysis of the process $\{Z_n\}_{n\in\mathbb{N}}$ around the point $z = 0$ is of particular interest, as this indicates how the "good" channels (i.e., the channels that have mutual information close to 1) behave. The asymptotic behavior of the process is closely related to the "partial distances" of the kernel matrix G:

Definition 1 (Partial Distances): We define the partial distances $D_i(G)$, $i = 0, \ldots, \ell-1$, of an $\ell \times \ell$ matrix $G = \begin{bmatrix} g_0 \\ \vdots \\ g_{\ell-1} \end{bmatrix}$ (g_i's are row vectors) as

$$D_i(G) \stackrel{\triangle}{=} d_H(\{g_i\}, \langle g_{i+1}, \ldots, g_{\ell-1}\rangle), \quad i = 0, \ldots, \ell-2,$$
$$D_{\ell-1}(G) \stackrel{\triangle}{=} d_H(\{g_{\ell-1}\}, \{\mathbf{0}\}).$$

Here, $\langle g_{i+1}, \ldots, g_{\ell-1}\rangle$ denotes the linear space spanned by $g_{i+1}, \ldots, g_{\ell-1}$. Also, $d_H(a,b)$ denotes the Hamming distance between two binary vectors a, b of equal length and more generally if A, B are two sets of binary vectors all of the same length, then $d_H(A,B) = \min_{a\in A, b\in B} d_H(a,b)$. The *exponent* of G is then defined as

$$E(G) = \frac{1}{\ell}\sum_{i=0}^{\ell-1} \log_\ell D_i(G),$$

and the *second exponent* of G is defined as

$$V(G) = \frac{1}{\ell}\sum_{i=0}^{\ell-1} (\log_\ell D_i(G) - E(G))^2.$$

In other words, the exponent $E(G)$ and the second exponent $V(G)$ are the mean and the variance of the random variable $\log_\ell D_B(G)$, where B is a random variable taking a value in $\{0, 1, \ldots, \ell-1\}$ with uniform probability. It should be noted that the invertibility of G implies the partial distances $\{D_i(G)\}$ to be strictly positive, making the exponent $E(G)$ finite [13]. Note also that the condition for a matrix G to be polarizing, that none of column permutations of G is upper triangular, implies

that at least one of the $D_i(G)$'s is strictly greater than 1, yielding $E(G)$ to be strictly positive.

The following theorem partially characterizes the behavior of the process $\{Z_n\}_{n\in\mathbb{N}}$ around $z = 0$.

Theorem 2 (See [2] and [13]): Let W be a BMS channel and assume that we are using as the kernel matrix an $\ell \times \ell$ matrix G with exponent $E(G)$. For any fixed β with $0 < \beta < E(G)$

$$\lim_{n\to\infty} \mathbb{P}(Z_n \leq 2^{-\ell^{n\beta}}) = I(W). \tag{9}$$

Conversely, if $I(W) < 1$, then for any fixed $\beta > E(G)$

$$\lim_{n\to\infty} \mathbb{P}(Z_n \geq 2^{-\ell^{n\beta}}) = 1. \tag{10}$$

■

An important consequence of Theorem 2 is as follows. Let $\mathbb{P}_{\mathrm{e}}^{\mathrm{SC}}(N, R)$ be the block error probability when using polar codes with the kernel matrix G, of block length $N = \ell^n$ and rate $R < I(W)$ under SC decoding. By using the inequality on the right-hand side in (5) and the limit in (9), we can easily conclude that for any $0 < \beta < E(G)$, the value of $\mathbb{P}_{\mathrm{e}}^{\mathrm{SC}}(N, R)$ is less than $2^{-\ell^{n\beta}}$ for sufficiently large n. Also, by using the inequality on the left-hand side in (5) and (10), we can easily conclude that for $\beta > E(G)$, the value of $\mathbb{P}_{\mathrm{e}}^{\mathrm{SC}}(N, R)$ is greater than $2^{-\ell^{n\beta}}$ for sufficiently large n. Hence, $\mathbb{P}_{\mathrm{e}}^{\mathrm{SC}}(N, R)$ behaves as $2^{-\ell^{nE(G)+o(n)}}$ as n tends to infinity. A noteworthy point about this result is that it is rate independent, provided that the rate R is less than the capacity $I(W)$. In this paper, we provide a refined estimate for $\mathbb{P}(Z_n \leq z)$. Specifically, we derive the asymptotic relation between $\mathbb{P}(Z_n \leq z)$ and the rate of transmission R. From this, we derive the asymptotic behavior of $\mathbb{P}_{\mathrm{e}}^{\mathrm{SC}}(N, R)$ and its dependence on the rate of transmission. We further derive lower bounds on the error probability when we perform MAP decoding instead of SC decoding.

An important point to mention here is that the results of this paper are obtained in the asymptotic limit of the block length for any *fixed* rate value R. Considering the regime where R also varies with the block length is a problem of different interest, for which we refer the reader to [9], [10], and [21].

The outline of the paper is as follows. In Section II, we state the main results of the paper. In Section III, we first define several auxiliary processes and provide bounds on their asymptotic behavior. Using these bounds, we then prove the main results. We discuss the implications of the proofs in selecting the set of channel indices in Section IV. It should be noted that in the following the logarithms are in base 2 unless explicitly stated otherwise.

II. Main Results

Theorem 3: Consider an $\ell \times \ell$ polarizing kernel matrix $G = \begin{bmatrix} g_0 \\ \vdots \\ g_{\ell-1} \end{bmatrix}$. For a BMS channel W, let $\{Z_n = Z(W_n)\}_{n\in\mathbb{N}}$ be the Bhattacharyya process of W. Let $Q(t) \stackrel{\triangle}{=} \int_t^\infty e^{-z^2/2}\, dz/\sqrt{2\pi}$ be the error function and $Q^{-1}(\cdot)$ be its inverse function.

1) For $R < I(W)$

$$\lim_{n\to\infty} \mathbb{P}\left(Z_n \leq 2^{-\ell^{nE(G)+\sqrt{nV(G)}Q^{-1}\left(\frac{R}{I(W)}\right)+f(n)}}\right) = R.$$

2) Let $H = [g_{\ell-1}^{\mathrm{T}}, \ldots, g_0^{\mathrm{T}}]^{-1}$ ($\cdot^{\mathrm{T}}$ denotes the transpose) and assume that $D_i(H) \leq D_{i-1}(H)$ for $1 \leq i \leq \ell - 1$. Then, for $R' < 1 - I(W)$ we have

$$\lim_{n\to\infty} \mathbb{P}\left(Z_n \geq 1 - 2^{-\ell^{nE(H)+\sqrt{nV(H)}Q^{-1}\left(\frac{R'}{1-I(W)}\right)+f(n)}}\right) = R'.$$

Here, $f(n)$ is any function satisfying $f(n) = o(\sqrt{n})$. ■

Theorem 3 characterizes the asymptotic behavior of $\mathbb{P}(Z_n \leq z)$ and refines Theorem 2 in the following way. According to Theorem 2, if we transmit at rate R below the channel capacity, then the quantity $\log_\ell(-\log(\mathbb{P}_{\mathrm{e}}^{\mathrm{SC}}(N = \ell^n, R)))$ scales like $nE(G) + o(n)$. The first part of Theorem 3 gives one further term by stating that $o(n)$ is in fact $\sqrt{nV(G)}Q^{-1}\left(\frac{R}{I(W)}\right) + o(\sqrt{n})$. The second part of Theorem 3, on the other hand, characterizes the asymptotic behavior of $\mathbb{P}(Z_n \leq z)$ near $z = 1$, which is important in applications of polar codes for source coding [12], [23]. Put together, Theorem 3 characterizes the scaling of the error probability of polar codes with the SC decoder. Note that the condition $D_i(H) \leq D_{i-1}(H)$ can be removed by using the $(1/2)\sum_{y\in\mathcal{Y}} |W(y\,|\,0) - W(y\,|\,1)|$ instead of $Z(W)$ [26]. Similar results hold for the case of the MAP decoder.

Theorem 4: Let W be a BMS channel and let $R < I(W)$ be the rate of transmission. Consider an $\ell \times \ell$ kernel matrix G with $\{w_0(G), \ldots, w_{\ell-1}(G)\}$ the Hamming weights of its rows and define

$$E_w(G) = \frac{1}{\ell}\sum_{i=0}^{\ell-1} \log_\ell w_i(G)$$
$$V_w(G) = \frac{1}{\ell}\sum_{i=0}^{\ell-1} (\log_\ell w_i(G) - E_w(G))^2.$$

If we use polar codes of length $N = \ell^n$ and rate R for transmission, then the probability of error under MAP decoding, $\mathbb{P}_{\mathrm{e}}^{\mathrm{MAP}}(N, R)$, satisfies

$$\log_\ell(-\log(\mathbb{P}_{\mathrm{e}}^{\mathrm{MAP}}(N, R))) \leq nE_w(G) + \sqrt{nV_w(G)}Q^{-1}\left(\frac{R}{I(W)}\right) + o(\sqrt{n}).$$

■

Corollary 5: Let G be according to Arıkan's original construction [1], i.e., $G = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$, which is the only polarizing matrix for the case $\ell = 2$. For this G, we have $w_i(G) = D_i(G)$ for $i = 0$ and 1. Hence, the block error probability for the SC decoder and the MAP block error probability share the same asymptotic behavior according to Theorems 3 and 4.

For a general $\ell \times \ell$ matrix G, however, one may have strict inequality $E_w(G) > E(G)$, in which case one still has an asymptotic gap between the error probability with SC decoding and the lower bound of MAP error probability. Whether or not this gap can be filled or made narrower is an open problem. We conclude this section by stating a corollary that is deduced from the proof of Theorem 4.

Corollary 6: Assuming $G = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$, the fraction of the common chosen row indices of $G^{\otimes n}$ between polar codes of rate R and RM codes of rate R' tends to $I(W)\min\{\frac{R}{I(W)}, R'\}$ as $n \to \infty$.

III. Proof of the Main Result

A. Preliminaries

Let $\{B_n\}_{n\in\mathbb{N}}$ be a sequence of i.i.d. random variables that take their values in $\{0, 1, \ldots, \ell - 1\}$ with uniform probability, i.e., $\mathbb{P}(B_0 = j) = \frac{1}{\ell}$ for $j \in \{0, 1, \ldots, \ell - 1\}$. Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote the probability space generated by the sequence $\{B_n\}_{n\in\mathbb{N}}$ and let $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ be the probability space generated by $(B_0, \ldots, B_n)$. By using the bounds given in [5, Lemma 5.7, Lemma 5.10], we have the following relationship between the Bhattacharyya parameters of W^i and that of W: Recall that $\{D_i(G)\}_{0\le i\le \ell-1}$ are the partial distances of the matrix G. We have [5, Lemma 5.10]

$$Z(W)^{D_i(G)} \le Z(W^i) \le 2^{\ell-i} Z(W)^{D_i(G)}. \tag{11}$$

Also, let $H = [g_{\ell-1}^{\mathrm{T}}, \ldots, g_0^{\mathrm{T}}]^{-1}$. Assuming $D_i(H) \le D_{i-1}(H)$, we have [5, Lemma 5.7]

$$(1 - Z(W))^{D_i(H)} \le 1 - Z(W^i) \le 2^{2i+1}(1 - Z(W))^{D_i(H)}. \tag{12}$$

B. Proof of Theorem 3

We first provide an intuitive picture behind the result of Theorem 3. For simplicity, assume $\ell = 2$ and $G = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$. Also, assume that W is a binary erasure channel with erasure probability ϵ. The capacity of this channel is $1 - \epsilon$. For such a channel, the Bhattacharyya process has a simple closed form [1] as $Z_0 = \epsilon$ and

$$Z_{n+1} = \begin{cases} Z_n^2, & B_n = 0, \\ 2Z_n - Z_n^2, & B_n = 1. \end{cases} \tag{13}$$

We know from Section I-C that as n grows large, Z_n tends almost surely to a $\{0, 1\}$-valued random variable Z_∞ with $\mathbb{P}(Z_\infty = 0) = 1 - \epsilon$. The asymptotic behavior of $\{Z_n\}_{n\in\mathbb{N}}$ can be explained roughly by considering the behavior of $\{-\log Z_n\}_{n\in\mathbb{N}}$. In particular, it is clear from (13) that at time $n + 1$, $-\log Z_n$ is either doubled (when $B_n = 0$), or decreased by at most 1 (when $B_n = 1$). Also, observe that once $-\log Z_n$ becomes sufficiently large, subtracting 1 from it has negligible effect compared with the doubling operation. Now assume that m is a sufficiently large number. Conditioned on the event that $-\log Z_m$ is a very large value (or equivalently, the value of Z_m is very close to 0: this happens with probability very close to $1 - \epsilon$), for $n > m$ the process $\{-\log Z_n\}_{n\in\mathbb{N}}$ evolves each time by being doubled if $B_n = 0$ or remaining roughly the same if $B_n = 1$. We can then use the central limit theorem to characterize the asymptotic behavior of $\{-\log Z_n\}_{n\in\mathbb{N}}$ for $n \gg m$.

The proof of Theorem 3 is done by making the above intuitive steps rigorous for a BMS channel W and a polarizing $\ell \times \ell$ kernel matrix G. In a slightly more general setting, we study the asymptotic properties of $\mathbb{P}(X_n \le x)$ for any generic process $\{X_n\}_{n\in\mathbb{N}}$ satisfying the conditions (c1)–(c4) defined as follows.

Definition 7: Let S be a random variable taking values in $[1, \infty)$. Assume that the expectation and the variance of $\log S$ exist and are denoted by $\mathbb{E}[\log S]$ and $\mathbb{V}[\log S]$, respectively. Also, assume that $\{S_n\}_{n\in\mathbb{N}}$ are i.i.d. samples of S. Let $\{(X_n, S_n) \in (0, 1) \times [1, \infty)\}_{n\in\mathbb{N}}$ be a random process satisfying the following conditions.

(c1) There exists a random variable X_∞ such that $X_n \to X_\infty$ holds almost surely.
(c2) With probability 1 we have $X_n^{S_n} \le X_{n+1}$.
(c3) There exists a constant $c \ge 1$ such that $X_{n+1} \le cX_n^{S_n}$ holds with probability 1.
(c4) S_n is independent of X_m for $m \le n$.

The random processes $\{(Z_n, D_{B_n}(G))\}_{n\in\mathbb{N}}$ and $\{(1 - Z_n, D_{B_n}(H))\}_{n\in\mathbb{N}}$ satisfy the above four conditions. The fact that these processes satisfy the condition (c1) has been proved in [5, Lemma 5.4], and the result reads that if G is polarizing, then Z_∞ takes only 0 and 1, with probabilities $I(W)$ and $1 - I(W)$, respectively. Conditions (c2) and (c3) also hold because of (11) and (12).

Our objective now is to prove that for such a process $\{(X_n, S_n)\}_{n\in\mathbb{N}}$, we have

$$\lim_{n\to\infty} \mathbb{P}\left(X_n \le 2^{-2^{n\mathbb{E}[\log S] + t\sqrt{n\mathbb{V}[\log S]} + f(n)}}\right) = \mathbb{P}(X_\infty = 0)Q(t) \tag{14}$$

where $f(n)$ is any function such that $f(n) = o(\sqrt{n})$ holds. The results of Theorem 3 then follow by noting that $\mathbb{P}(Z_\infty = 0) = I(W)$ and $\mathbb{P}(1 - Z_\infty = 0) = \mathbb{P}(Z_\infty = 1) = 1 - I(W)$ hold, and by substituting $t = Q^{-1}(R/I(W))$ and $t = Q^{-1}(R'/(1 - I(W)))$, respectively, into (14).

We prove (14) by showing the two inequalities obtained by replacing the equality in(14) by inequality in both directions. As the first step, we have the following lemma.

Lemma 8: Let $\{(X_n, S_n)\}_{n\in\mathbb{N}}$ be a random process satisfying (c1), (c3) and (c4). For any $f(n) = o(\sqrt{n})$

$$\liminf_{n\to\infty} \mathbb{P}\left(X_n \le 2^{-2^{n\mathbb{E}[\log S] + t\sqrt{n\mathbb{V}[\log S]} + f(n)}}\right) \ge \mathbb{P}(X_\infty = 0)Q(t).$$

Proof: Without loss of generality, we can assume that c in condition (c3) satisfies $c \ge 2$. Define the process $\{L_n\}_{n\in\mathbb{N}}$ as $L_n \triangleq \log X_n$. From (c3), we have

$$L_n \le \log c + S_{n-1}L_{n-1}$$

and by applying the above relation recursively, for $m \leq n-1$, we obtain

$$L_n \leq \left(\sum_{j=m}^{n-1}\prod_{i=j+1}^{n-1} S_i\right)\log c + \left(\prod_{i=m}^{n-1} S_i\right) L_m$$
$$\leq \left(\prod_{i=m}^{n-1} S_i\right)((n-m)\log c + L_m). \quad (15)$$

Fix $\beta \in (0, \mathbb{E}[\log S])$ and let

$$m \stackrel{\triangle}{=} (\log n + \log\log c)/\beta. \quad (16)$$

Conditioned on the event $\mathcal{D}_m(\beta) \stackrel{\triangle}{=} \{X_m < 2^{-2^{\beta m}}\}$, by using (15), we obtain

$$L_n \leq -\left(\prod_{i=m}^{n-1} S_i\right) m \log c.$$

Let the event $\mathcal{H}_m^{n-1}(t)$ be defined as

$$\mathcal{H}_m^{n-1}(t) \stackrel{\triangle}{=} \Bigg\{\sum_{i=m}^{n-1}\log S_i \geq (n-m)\mathbb{E}[\log S] + t\sqrt{(n-m)\mathbb{V}[\log S]} + f(n-m)\Bigg\}$$

where f is any function such that $f(k) = o(\sqrt{k})$ holds. Conditioned on $\mathcal{D}_m(\beta)$ and $\mathcal{H}_m^{n-1}(t)$, we have

$$\log(-L_n) \geq \log m + \log\log c + (n-m)\mathbb{E}[\log S] + t\sqrt{(n-m)\mathbb{V}[\log S]} + f(n-m).$$

Hence

$$\mathbb{P}\Big(\log(-L_n) \geq \log m + \log\log c + (n-m)\mathbb{E}[\log S] + t\sqrt{(n-m)\mathbb{V}[\log S]} + f(n-m)\Big)$$
$$\geq \mathbb{P}(\mathcal{D}_m(\beta) \cap \mathcal{H}_m^{n-1}(t)) = \mathbb{P}(\mathcal{D}_m(\beta))\mathbb{P}(\mathcal{H}_m^{n-1}(t)).$$

The last equality follows from the independence condition (c4).

Note that taking the limit $n \to \infty$ also implies $m \to \infty$ and $n - m \to \infty$ via (16). From Theorem 12 (in the Appendix), we have $\lim_{n\to\infty}\mathbb{P}(\mathcal{D}_m(\beta)) = \mathbb{P}(X_\infty = 0)$. We also have $\lim_{n\to\infty}\mathbb{P}(\mathcal{H}_m^{n-1}(t)) = Q(t)$ due to the central limit theorem for $\{\log S_i\}$. We consequently have

$$\liminf_{n\to\infty}\mathbb{P}\Big(\log(-\log X_n) \geq n\mathbb{E}[\log S] + t\sqrt{n\mathbb{V}[\log S]} + f(n)\Big) \geq \mathbb{P}(X_\infty = 0)Q(t)$$

for any $f(n) = o(\sqrt{n})$. ■

The second step of the proof of (14) is to prove the other direction of the inequality. We have the following.

Lemma 9: Let $\{(X_n, S_n)\}_{n\in\mathbb{N}}$ be a random process satisfying (c1), (c2), and (c4). For any $f(n) = o(\sqrt{n})$

$$\limsup_{n\to\infty}\mathbb{P}\left(X_n \leq 2^{-2^{n\mathbb{E}[\log S] + t\sqrt{n\mathbb{V}[\log S]} + f(n)}}\right) \leq \mathbb{P}(X_\infty = 0)Q(t).$$

Proof: Let $L_n \stackrel{\triangle}{=} \log X_n$. From (c2), for $m \leq n-1$, we have

$$L_n \geq S_{n-1}L_{n-1} \geq \left(\prod_{i=m}^{n-1} S_i\right) L_m$$

and thus

$$\log(-L_n) \leq \sum_{i=m}^{n-1}\log S_i + \log(-L_m). \quad (17)$$

Hence, for any fixed m and any $\delta \in (0,1)$

$$\limsup_{n\to\infty}\mathbb{P}\Big(\log(-L_n) > n\mathbb{E}[\log S] + t\sqrt{n\mathbb{V}[\log S]} + f(n)\Big) \leq A + B \quad (18)$$

where

$$A = \limsup_{n\to\infty}\mathbb{P}\Big(\log(-L_n) > n\mathbb{E}[\log S] + t\sqrt{n\mathbb{V}[\log S]} + f(n), X_m \leq \delta\Big) \quad (19)$$

and

$$B = \limsup_{n\to\infty}\mathbb{P}\Big(\log(-L_n) > n\mathbb{E}[\log S] + t\sqrt{n\mathbb{V}[\log S]} + f(n), X_m > \delta\Big). \quad (20)$$

We proceed now by providing upper bounds on the terms A and B. By using (17), the term A in (19) is upper bounded as

$$A \leq \limsup_{n\to\infty}\mathbb{P}\Big(\sum_{i=m}^{n-1}\log S_i + \log(-L_m) > n\mathbb{E}[\log S] + t\sqrt{n\mathbb{V}[\log S]} + f(n),\ X_m \leq \delta\Big). \quad (21)$$

Also, from the condition (c4) and the central limit theorem, we obtain that the right-hand side of (21) is equal to $Q(t)\mathbb{P}(X_m \leq \delta)$. Hence, we can write

$$A \leq Q(t)\mathbb{P}(X_m \leq \delta). \quad (22)$$

The term B in (20) is upper bounded as

$$B \leq \limsup_{n\to\infty}\mathbb{P}\left(X_n \leq \frac{\delta}{2},\ X_m > \delta\right) \stackrel{(a)}{\leq} \mathbb{P}\left(X_\infty \leq \frac{\delta}{2},\ X_m > \delta\right) \quad (23)$$

where (a) follows from (c1). Applying (22) and (23) to (18), for any $\delta \in (0,1)$, we have

$$\begin{aligned}
&\limsup_{n\to\infty} \mathbb{P}\left(\log(-L_n) > n\mathbb{E}[\log S] + t\sqrt{n\mathbb{V}[\log S]} + f(n)\right) \\
&\le \limsup_{m\to\infty} \left\{ Q(t)\mathbb{P}(X_m \le \delta) + \mathbb{P}\left(X_\infty \le \frac{\delta}{2},\ X_m > \delta\right)\right\} \\
&\le Q(t)\mathbb{P}(X_\infty \le \delta) + \mathbb{P}\left(X_\infty \le \frac{\delta}{2},\ X_\infty \ge \delta\right) \\
&= Q(t)\mathbb{P}(X_\infty \le \delta).
\end{aligned}$$

By letting $\delta \to 0$, we obtain the result. ■

C. Proof of Theorem 4

We start the proof of Theorem 4 by stating a general fact regarding the MAP error probability of linear codes.

Lemma 10: The MAP error probability of a linear code $\mathcal{C}$ over a BMS channel W is lower bounded by $Z(W)^{2d_{\min}}/4$ where $d_{\min}$ is the minimum distance of $\mathcal{C}$.

Proof: Within this proof, the notation $\mathbb{P}(\cdots)$ should be understood as generically denoting the probability of an event $(\cdots)$. Since the MAP error probability of a linear code over a BMS channel does not depend on the transmitted codeword, we can assume without loss of generality that the transmitted codeword is the all-zero codeword, which is denoted by $\mathbf{0}$. Let $\boldsymbol{Y}$ be the random variable corresponding to a received sequence when $\mathbf{0}$ is transmitted and let $P(y \,|\, c)$ be the likelihood of received sequence y given a codeword c. Consider an arbitrary codeword $c \in \mathcal{C}\backslash\{\mathbf{0}\}$. Since MAP and ML are equivalent for equiprobable codewords, the MAP error probability is clearly lower bounded as

$$\mathbb{P}(\cup_{c'\in\mathcal{C}\backslash\{\mathbf{0}\}}\{P(\boldsymbol{Y}|c') \ge P(\boldsymbol{Y}|\mathbf{0})\}) \ge \mathbb{P}(P(\boldsymbol{Y}|c) \ge P(\boldsymbol{Y}|\mathbf{0})).$$

That is, assuming that $\mathbf{0}$ has been sent, the MAP error probability is lower bounded by the probability that the codeword c is more likely than $\mathbf{0}$. We now provide a lower bound for the probability of the latter event. Let us consider c as a binary vector of length N, i.e., $c = (c_0, c_1, \ldots, c_{N-1})$. We let A be the set of indices $i \in \{0, 1, \ldots, N-1\}$ such that $c_i = 1$. Thus, the set A has cardinality equal to the Hamming weight of c which we write as $w(c)$. We thus obtain

$$\mathbb{P}(P(\boldsymbol{Y} \,|\, c) \ge P(\boldsymbol{Y} \,|\, \mathbf{0})) = \mathbb{P}\big(\prod_{i\in A} P(y_i \,|\, 1) \ge \prod_{i\in A} P(y_i \,|\, 0)\big). \tag{24}$$

For a positive integer m, let us define the BMS channel $W^{\otimes m}$: $\{0,1\} \to \mathcal{Y}^m$ as

$$W^{\otimes m}(y_1^m \,|\, x) \stackrel{\triangle}{=} \prod_{i=1}^{m} W(y_i \,|\, x). \tag{25}$$

It is now easy to see that the right-hand side of (24) is equal to the probability that having sent the symbol 0 on the channel $W^{\otimes w(c)}$, we receive an output for which the symbol 1 is more likely than 0. Hence

$$\begin{aligned}
&\mathbb{P}\big(\prod_{i\in A} P(y_i \,|\, 1) \ge \prod_{i\in A} P(y_i \,|\, 0)\big) \\
&= P_{\mathrm{e}}(W^{\otimes w(c)}) \\
&\stackrel{(a)}{\ge} \frac{1}{2}\left(1 - \sqrt{1 - Z(W^{\otimes w(c)})^2}\right) \\
&\stackrel{(b)}{=} \frac{1}{2}\left(1 - \sqrt{1 - Z(W)^{2w(c)}}\right) \\
&\ge \frac{1}{4} Z(W)^{2w(c)}
\end{aligned}$$

where step (a) follows from (1) and where (b) follows from the fact that for $m \ge 1$ we have $Z(W^{\otimes m}) = Z(W)^m$ [1]. ■

It should be noted that the lower bound $P_{\mathrm{e}}(W^{\otimes w(c)}) \ge (1/4)Z(W)^{2w(c)}$ in the proof of Lemma 10 is not asymptotically tight in terms of the conventional exponents. It is possible to obtain tighter lower bounds via more elaborate arguments as in [22, Ch. 4]. However, since we are only interested in the behavior of double exponents, the above bound turns out to be sufficient for the purpose of proving Theorem 4.

In order to prove Theorem 4, from Lemma 10, it is sufficient to prove that given any $\epsilon > 0$, there exists an integer $M \in \mathbb{N}$ such that for $n \ge M$

$$\log_\ell(d(n,R)) \le nE_w(G) + \sqrt{nV_w(G)}\left(Q^{-1}\left(\frac{R}{I(W)}\right) + \epsilon\right)$$

where $d(n,R)$ is the minimum distance of a polar code using the kernel matrix G, with block length $N = \ell^n$ and rate R. We note that a row weight of the generator matrix is an upper bound of the minimum distance for a linear code, and also that the weight of the ith row of $G^{\otimes n}$ is equal to $\prod_{j=1}^{n} w_{i_j}(G)$, where i_j is the jth digit of the ℓ-ary representation of $i-1$. As a result, it is sufficient to prove that given any $\epsilon > 0$, there exists an integer $M \in \mathbb{N}$ such that for a polar code of block length $N = \ell^n \ge \ell^M$, rate R and set of chosen indices $\mathcal{I}$, there exists $i \in \mathcal{I}$ for which the inequality

$$\sum_{j=1}^{n} \log_\ell w_{i_j}(G) \le nE_w(G) + \sqrt{nV_w(G)}\left(Q^{-1}\left(\frac{R}{I(W)}\right) + \epsilon\right) \tag{26}$$

holds. In the proof of Theorem 3, one can observe that the key idea is to apply the central limit theorem for the i.i.d. sequence $\{\log S_n = \log D_{B_n}(G)\}_{n\in\mathbb{N}}$. In order to prove Theorem 4, we also consider the i.i.d. sequence $\{\log w_{B_n}(G)\}_{n\in\mathbb{N}}$ in addition to $\{\log D_{B_n}(G)\}_{n\in\mathbb{N}}$. Note that the two sequences $\{\log D_{B_n}(G)\}_{n\in\mathbb{N}}$ and $\{\log w_{B_n}(G)\}_{n\in\mathbb{N}}$ are in general correlated since they are both coupled to the same process $\{B_n\}_{n\in\mathbb{N}}$ and they are equal with probability one if and only if $D_i(G) = w_i(G)$ holds for all $i \in \{0,\ 1,\ \ldots,\ \ell-1\}$. In the same manner as the proof of Theorem 3, we move on to a more abstract setting. We first introduce a random variable U taking values in $[1,\infty)$, for which we assume that the expectation and the variance of

$\log U$ exist and are denoted by $\mathbb{E}[\log U]$ and $\mathbb{V}[\log U]$, respectively. We also let $\{(S_n, U_n)\}_{n\in\mathbb{N}}$ be i.i.d. drawings of (S, U), where S is defined as in Definition 7. Let $\{(X_n, S_n, U_n)\}_{n\in\mathbb{N}}$ be a random process such that $\{(X_n, S_n)\}_{n\in\mathbb{N}}$ satisfies the conditions (c1) to (c4) together with the additional condition (c5) for $\{(X_n, U_n)\}_{n\in\mathbb{N}}$: (c5) U_n is independent of X_m for $m \leq n$.

It is easy to see that the stochastic process of the triplets $\{(Z_n, D_{B_n}(G), w_{B_n}(G))\}_{n\in\mathbb{N}}$ satisfies (c1) to (c5). We first note from the proof of Theorem 4 that for any generic process $\{(X_n, S_n, U_n)\}_{n\in\mathbb{N}}$ satisfying (c1) to (c5), relation (14) holds for any function $f(n) = o(\sqrt{n})$. We also claim that for real numbers v, t such that $v > t$ and for any function $g(n) = o(\sqrt{n})$, we have

$$\limsup_{n\to\infty} \mathbb{P}\Bigg(X_n \leq 2^{-2^{n\mathbb{E}[\log S]+t\sqrt{n\mathbb{V}[\log S]}+f(n)}}, \sum_{i=0}^{n-1} \log U_i > n\mathbb{E}[\log U] + v\sqrt{n\mathbb{V}[\log U]} + g(n)\Bigg) < \mathbb{P}(X_\infty = 0)Q(v). \quad (27)$$

Before proving (27), let us see how it leads to the proof of Theorem 4. Since the stochastic process of the triplets $\{(Z_n, D_{B_n}(G), w_{B_n}(G))\}_{n\in\mathbb{N}}$ satisfies (c1) to (c5), we can use relations (14) and (27) by letting $(X_n, S_n, U_n) = (Z_n, D_{B_n}(G), w_{B_n}(G))$. Now, by (14) and (27), it is easy to see that for generator matrices of polar codes with rate R, the number of rows satisfying (26) is asymptotically proportional to the block length, and hence, there exists at least one row satisfying(26) which concludes the proof of Theorem 4. Thus, it remains to prove the claim (27).

Lemma 11: Let $\{(X_n, S_n, U_n)\}_{n\in\mathbb{N}}$ be a random process satisfying (c1) to (c5). For any $f(n) = o(\sqrt{n})$ and $g(n) = o(\sqrt{n})$

$$\lim_{n\to\infty} \mathbb{P}\Bigg(X_n \leq 2^{-2^{n\mathbb{E}[\log S]+t\sqrt{n\mathbb{V}[\log S]}+f(n)}}, \sum_{i=0}^{n-1} \log U_i > n\mathbb{E}[\log U] + v\sqrt{n\mathbb{V}[\log U]} + g(n)\Bigg) = \mathbb{P}(X_\infty = 0)\mathbb{P}(A_S \geq t, A_U \geq v)$$

where (A_S, A_U) are Gaussian random variables of mean zero whose covariance matrix is equal to that of

$$\left(\frac{\log S - \mathbb{E}[\log S]}{\sqrt{\mathbb{V}[\log S]}}, \frac{\log U - \mathbb{E}[\log U]}{\sqrt{\mathbb{V}[\log U]}} \right).$$

The proof of this Lemma is the same as the proofs of Lemmas 8 and 9. The difference is that the central limit theorem is replaced by the 2-D central limit theorem. From the fact that $\mathbb{P}(A_S \geq t, A_U \geq v) \leq Q(\max\{t, v\})$, relation (27) is obtained for $v > t$. This completes the proof of Theorem 4.

D. Proof of Corollary 6

Let $G = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$. For this choice of G, we have $w_i(G) = D_i(G)$ for $i = 0$ and 1. Hence, the random variables $S_n = D_{B_n}(G)$ and $U_n = w_{B_n}(G)$ are equal for all $n \in \mathbb{N}$. Also note that S_n takes its value in the set $\{1, 2\}$ uniformly at random. From the proof of Theorem 4, the set of indices of the rows of polar codes with the kernel matrix G and rate R corresponds to the event

$$\left\{ X_n \leq 2^{-2^{n\mathbb{E}[\log S]+Q^{-1}\left(\frac{R}{I(W)}\right)\sqrt{n\mathbb{V}[\log S]}+f(n)}} \right\}.$$

Also, with the same G, the set of indices of a RM code with rate R' corresponds to the event

$$\left\{ \sum_{i=0}^{n-1} \log U_i > n\mathbb{E}[\log U] + Q^{-1}(R')\sqrt{n\mathbb{V}[\log U]} + g(n) \right\}.$$

From Lemma 11, it is now easy to conclude that the fraction of the common chosen row indices of $G^{\otimes n}$ between polar codes of rate R and RM codes of rate R' tends to $I(W)\min\{\frac{R}{I(W)}, R'\}$ as $n \to \infty$.

IV. Selection Rule of Rows

The proof of Lemma 8 suggests a way to help us select the good indices in a computationally efficient way. Let us recall that the construction of polar codes relies on finding the "quality" of the channels $W_{\ell^n}^{(i)}$, $1 \leq i \leq \ell^n$. "Quality" here can either refer to large capacity (capacity close to 1) or small Bhattacharyya value. We also recall from Section I-B that for each i, $1 \leq i \leq N$, the channel $W_{\ell^n}^{(i)}$ is constructed from W as follows. We first compute the ℓ-ary representation of $i-1$, which is denoted by $b_1 b_2 \cdots b_n$, with b_1 being the most significant digit and

$$W_{\ell^n}^{(i)} = (((W^{b_1})^{b_2})^{\cdots})^{b_n}.$$

The goal of this section is to show that the quality of a channel $W_{\ell^n}^{(i)}$ depends on the channel W only through the first $o(n)$ digits of the sequence $b_1 b_2 \cdots b_n$. In other words, to choose the indices of the channels $W_{\ell^n}^{(i)}$, $1 \leq i \leq N$, that have the best quality, the first $o(n)$ significant digits of the ℓ-ary expansion of $i-1$ should be determined depending on W and the rest are determined in a RM-like fashion (i.e., are chosen according to their Hamming weight).

In the proof, the ℓ-ary expansion of row indices of $G^{\otimes n}$ corresponds to realizations of $B_1, \ldots, B_n$. The proof of Lemma 8 implies that it is sufficient to select the rows in $\mathcal{D}_m(\beta) \cap \mathcal{H}_m^{n-1}(t)$ in order to achieve the asymptotically optimum performance. It should be noted that the event $\mathcal{D}_m(\beta)$ applied to the Bhattacharyya process $\{Z_n = Z(W_n)\}_{n\in\mathbb{N}}$ of W depends on the channel W, whereas the event $\mathcal{H}_m^{n-1}(t)$ is channel independent. This observation leads to the following selection rule: The first $m = s(n) \triangleq (\log n + \log\log c)/\beta$ digits of the row indices are determined in the channel-dependent way. Then, the following $(n-m)$ digits are determined in the RM way, i.e., those combinations of digits $(B_m, \ldots, B_{n-1})$ giving large values of $\sum_{i=m}^{n-1} \log D_{B_i}(G)$ are selected. In this rule, only the first $\Theta(\log n)$ digits should be determined depending on the channel.

The above argument can further be extended in a recursive manner. Let $\mathcal{C}_m^{n-1}(\epsilon) \triangleq \{(n-m)^{-1}\sum_{i=m}^{n-1} \log S_i \geq \mathbb{E}[\log S] - \epsilon\}$. Then, it is sufficient to select rows in $\mathcal{D}_{m_0}(\beta) \cap \mathcal{C}_{m_0}^{m_1-1}(\epsilon) \cap \mathcal{H}_{m_1}^{n-1}(t)$, where $m_1 = s(n)$ and $m_0 = s(m_1)$ since $\mathcal{D}_{m_1}(\beta)$ and $\mathcal{D}_{m_0}(\beta) \cap \mathcal{C}_{m_0}^{m_1-1}(\mathbb{E}[\log S] - \beta)$ are asymptotically equal. (Use $\mathcal{C}_m^{n-1}(\epsilon)$ instead of $\mathcal{H}_m^{n-1}(t)$ in the proof of Lemma 8. A

similar argument can be found in [1, Sec. IV-B].) From this observation, only $\Theta(\log\log n)$ digits have to be determined depending on the channel. By iterating this argument, we obtain the selection rule in which only

$$\Theta(\overbrace{\log\cdots\log}^{k} n) \tag{28}$$

digits depend on the channel for any $k \in \mathbb{N}$. From the argument so far, we deduce that even though the behavior of $Z_n = Z(W_n)$ depends on the channel W as well as the whole sequence $\{B_0, B_1, \ldots, B_{n-1}\}$, whether it approaches 0 or 1 when n is large, is mostly determined by the channel W and a prefix of $\{B_0, B_1, \ldots, B_{n-1}\}$ with a relatively small length. Thus, to choose the indices of the channels $W_{\ell^n}^{(i)}$ that have the best quality, the first sublinear number of significant digits of the ℓ-ary expansion of $i-1$ are determined depending on the channel and the rest are determined in an RM-like fashion. It should be noted that the above argument is valid in the large-n asymptotics. It does not mean that one can make the number of digits to be determined in the channel-dependent manner arbitrarily small.

Although the good indices of the rows of $G^{\otimes n}$ can be selected using density evolution [3], in practice, storage and convolution of probability density functions is exponentially (in block length N) costly in terms of memory and computation. Recently, several authors have considered efficient algorithms that closely approximate the density evolution procedure [24], [25]. The aforementioned construction rule can be useful in reducing the number of convolutions and the number of levels in the quantization of channels.

APPENDIX

Theorem 12: Let $\{(X_n, S_n) \in (0,1)\times[1,\infty)\}_{n\in\mathbb{N}}$ be a random process satisfying (c1) and (c3). For any fixed $\beta \in (0, \mathbb{E}[\log S])$

$$\lim_{n\to\infty} \mathbb{P}\left(X_n \le 2^{-2^{\beta n}}\right) = \mathbb{P}(X_\infty = 0).$$

Remark: Although Theorem 12 has already been stated for Bhattacharyya processes $\{Z_n\}_{n\in\mathbb{N}}$ in [2] and [5], we would nevertheless like to confirm that the result is obtained by using only the two conditions (c1) and (c3).

Proof of Theorem 12: As the inequality

$$\limsup_{n\to\infty} \mathbb{P}\left(X_n \le 2^{-2^{\beta n}}\right) \le \mathbb{P}(X_\infty = 0)$$

obviously holds, a proof of the lower bound

$$\liminf_{n\to\infty} \mathbb{P}\left(X_n \le 2^{-2^{\beta n}}\right) \ge \mathbb{P}(X_\infty = 0)$$

is given in the following. Fix $\epsilon \in (0,1)$. Let $\{J_n\}_{n\in\mathbb{N}}$ be the random process defined as

$$J_n \stackrel{\triangle}{=} \begin{cases} \log(-\log X_n), & \text{for } n = 0,\ldots,m \\ \log(S_{n-1}-\epsilon) + J_{n-1}, & \text{for } n > m \end{cases}$$

which is to be used for deriving a probabilistic bound for $\{X_n\}_{n\in\mathbb{N}}$. Let $\mathcal{T}_m^n(\gamma) \stackrel{\triangle}{=} \{X_i < \gamma, \text{ for } i = m, m+1, \ldots, n\}$. Fix $k \in \{1, 2, \ldots\}$. From (c3), conditioned on $\mathcal{T}_m^{m+k-1}(c^{-1/\epsilon})$, the inequality $\log(-\log X_n) \ge J_n$ holds for $n = m, m+1, \ldots, m+k$. For the process $\{J_n\}_{n\in\mathbb{N}}$, the inequality

$$\begin{aligned} J_{m+k} &= J_m + \sum_{i=m}^{m+k-1} \log(S_i - \epsilon) \\ &\ge J_m + \sum_{i=m}^{m+k-1} (\log S_i + \log(1-\epsilon)) \end{aligned}$$

holds since $S_i \ge 1$. This inequality immediately implies the following conditional bound: Conditioned on $\mathcal{C}_m^{m+k-1}(\epsilon) \stackrel{\triangle}{=} \{(1/k)\sum_{i=m}^{m+k-1} \log S_i \ge \mathbb{E}[\log S] - \epsilon\}$, one has

$$J_{m+k} \ge J_m + k(\mathbb{E}[\log S] - \epsilon + \log(1-\epsilon)).$$

We have, therefore, obtained a probabilistic bound of $\log(-\log X_{m+k})$ of the form

$$\begin{aligned} \mathbb{P}(\log(-\log X_{m+k}) &\ge J_m + k(\mathbb{E}[\log S] - \epsilon + \log(1-\epsilon))) \\ &\ge \mathbb{P}\left(\mathcal{T}_m^{m+k-1}(c^{-1/\epsilon}) \cap \mathcal{C}_m^{m+k-1}(\epsilon)\right) \\ &\ge \mathbb{P}\left(\mathcal{T}_m^{m+k-1}(c^{-1/\epsilon})\right) + \mathbb{P}\left(\mathcal{C}_m^{m+k-1}(\epsilon)\right) - 1 \end{aligned}$$

for any $m \in \mathbb{N}$, $k \in \mathbb{N}$ and $\epsilon > 0$. From the law of large numbers, $\lim_{k\to\infty} \mathbb{P}\left(\mathcal{C}_m^{m+k-1}(\epsilon)\right) = 1$. From (c1), $\lim_{m\to\infty}\lim_{k\to\infty} \mathbb{P}\left(\mathcal{T}_m^{m+k-1}(c^{-1/\epsilon})\right) \ge \mathbb{P}(X_\infty < c^{-1/\epsilon})$. Hence

$$\begin{aligned} &\liminf_{m\to\infty}\liminf_{k\to\infty} \\ &\mathbb{P}(\log(-\log X_{m+k}) \ge J_m + k(\mathbb{E}[\log S] - \epsilon + \log(1-\epsilon))) \\ &\qquad \ge \mathbb{P}(X_\infty < c^{-1/\epsilon}) \ge \mathbb{P}(X_\infty = 0) \end{aligned}$$

holds for any $\epsilon > 0$. On the other hand, we observe that

$$\begin{aligned} &\liminf_{n\to\infty} \mathbb{P}\left(\frac{1}{n}\log(-\log X_n) \ge \mathbb{E}[\log S] - \gamma\right) \\ &\ge \liminf_{k\to\infty} \\ &\mathbb{P}(\log(-\log X_{m+k}) \ge J_m + k(\mathbb{E}[\log S] - \epsilon + \log(1-\epsilon))) \end{aligned}$$

holds for any fixed $m \in \mathbb{N}$ and $\gamma > \phi(\epsilon) \stackrel{\triangle}{=} \epsilon - \log(1-\epsilon)$. Hence

$$\begin{aligned} &\liminf_{n\to\infty} \mathbb{P}\left(\frac{1}{n}\log(-\log X_n) \ge \mathbb{E}[\log S] - \gamma\right) \\ &\qquad \ge \mathbb{P}(X_\infty = 0) \end{aligned}$$

for any $\gamma > 0$ since $\phi(\epsilon) > 0$ for $\epsilon > 0$ and $\lim_{\epsilon\to 0}\phi(\epsilon) = 0$. ■

REFERENCES

[1] E. Arıkan, "Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," *IEEE Trans. Inf. Theory*, vol. 55, no. 7, pp. 3051–3073, Jul. 2009.

[2] E. Arıkan and E. Telatar, "On the rate of channel polarization," in *Proc. IEEE Int. Symp. Inf. Theory*, Seoul, South Korea, 2009, pp. 1493–1495.

[3] R. Mori and T. Tanaka, "Performance and construction of polar codes on symmetric binary-input memoryless channels," in *Proc. IEEE Int. Symp. Inf. Theory*, Seoul, South Korea, 2009, pp. 1496–1500.
[4] S. H. Hassani, S. B. Korada, and R. Urbanke, "The compound capacity of polar codes," in *Proc. 47th Annu. Allerton Conf. Commun., Control, Comput.*, 2009, pp. 16–21.
[5] S. B. Korada, "Polar codes for channel and source coding," Ph.D. dissertation, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 2009.
[6] R. Mori, "Properties and construction of polar codes," Master's thesis, Graduate School Informat., Kyoto Univ., Kyoto, Japan, 2010.
[7] T. Tanaka and R. Mori, "Refined rate of channel polarization," in *Proc. IEEE Int. Symp. Inf. Theory*, Austin, TX, 2010, pp. 889–893.
[8] S. H. Hassani and R. Urbanke, "On the scaling of polar codes: I. The behavior of polarized channels," in *Proc. IEEE Int. Symp. Inf. Theory*, Austin, TX, 2010, pp. 874–878.
[9] S. H. Hassani, K. Alishahi, and R. Urbanke, "On the scaling of polar codes: II. The behavior of un-polarized channels," in *Proc. IEEE Int. Symp. Inf. Theory*, Austin, TX, 2010, pp. 879–883.
[10] S. B. Korada, A. Montanari, E. Telatar, and R. Urbanke, "An empirical scaling law for polar codes," in *Proc. IEEE Int. Symp. Inf. Theory*, Austin, TX, 2010, pp. 884–888.
[11] T. Tanaka, "On the speed of channel polarization," presented at the IEEE Inf. Theory Workshop, Dublin, Ireland, 2010.
[12] S. B. Korada and R. Urbanke, "Polar codes are optimal for lossy source coding," *IEEE Trans. Inf. Theory*, vol. 56, no. 12, pp. 1751–1768, Dec. 2010.
[13] S. B. Korada, E. Şaşoğlu, and R. Urbanke, "Polar codes: Characterization of exponent, bounds, and constructions," *IEEE Trans. Inf. Theory*, vol. 56, no. 12, pp. 6253–6264, Dec. 2010.
[14] E. Şaşoğlu, E. Telatar, and E. Arıkan, Polarization for arbitrary discrete memoryless channels [Online]. Available: arXiv:0908.0302 [cs.IT]
[15] H. Mahdavifar and A. Vardy, Achieving the secrecy capacity of wiretap channels using polar codes [Online]. Available: arXiv:1001.0210v2 [cs.IT]
[16] R. Mori and T. Tanaka, "Channel polarization on q-ary discrete memoryless channels by arbitrary kernels," in *Proc. IEEE Int. Symp. Inf. Theory*, Austin, TX, 2010, pp. 894–898.
[17] M. Bakshi, S. Jaggi, and M. Effros, "Concatenated polar codes," in *Proc. IEEE Int. Symp. Inf. Theory*, Austin, TX, 2010, pp. 918–922.
[18] E. Hof and S. Shamai, Secrecy-achieving polar-coding for binary-input memoryless symmetric wire-tap channels [Online]. Available: arXiv:1005.2759v2 [cs.IT]
[19] E. Hof, I. Sason, and S. Shamai, Polar coding for reliable communications over parallel channels [Online]. Available: arXiv:1005.2770v1 [cs.IT]
[20] E. Abbe and E. Telatar, Polar codes for the m-user MAC [Online]. Available: arXiv:1002.0777v2 [cs.IT]
[21] K. Alishahi, S. H. Hassani, S. B. Korada, and R. Urbanke, *On the finite-length scaling of polar codes*, in preparation.
[22] T. Richardson and R. Urbanke, *Modern Coding Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2008.
[23] E. Arikan, "Source polarization," in *Proc. IEEE Int. Symp. Inf. Theory*, Austin, TX, 2010, pp. 899–903.
[24] I. Tal and A. Vardy, "How to construct polar codes," presented at the IEEE Inf. Theory Workshop, Dublin, Ireland, 2010, arXiv:1105.6164v1 [cs.IT].
[25] R. Pedarsani, H. Hassani, I. Tal, and E. Telatar, "On the construction of polar codes," in *Proc. IEEE Int. Symp. Inf. Theory*, St. Petersberg, Russia, 2011, pp. 11–15.
[26] R. Mori and T. Tanaka, *Source and channel polarization over finite fields and Reed-Solomon matrix*, in preparation.

S. Hamed Hassani received B.Sc. degrees in electrical engineering and in pure mathematics in 2007 from Sharif University of Technology, Tehran, Iran. Since 2008, he has been a Ph.D. student in School of Computer and Communication Sciences of EPFL, Lausanne, Switzerland. His fields of interests include coding theory, graphical models, statistical physics, and theoretical computer science.

Ryuhei Mori received the B.E. degree from Tokyo Institute of Technology, Tokyo, Japan in 2008, and the M.Inf. degree from Kyoto University, Kyoto, Japan in 2010. Since 2010, he has been a Ph.D. student in Kyoto University. His research interests include information theory, coding theory and statistical physics.

Toshiyuki Tanaka received the B.E., M.E., and D.E. degrees in electronics engineering from the University of Tokyo, Tokyo, Japan, in 1988, 1990, and 1993, respectively. From 1993 to 2005, he was with the Department of Electronics and Information Engineering at Tokyo Metropolitan University, Tokyo, Japan. He is currently a professor at the Graduate School of Informatics, Kyoto University, Kyoto, Japan. He received DoCoMo Mobile Science Prize in 2003, and Young Scientist Award from the Minister of Education, Culture, Sports, Science and Technology, Japan, in 2005. His research interests are in the areas of information and communication theory, statistical mechanics of information processing, machine learning, and neural networks. He is a member of the IEEE, the Japanese Neural Network Society, the Acoustical Society of Japan, the Physical Society of Japan, and the Architectural Institute of Japan.

Rüdiger L. Urbanke received the Diplomingenieur degree from the Vienna Institute of Technology, Vienna, Austria, in 1990 and the M.S. and Ph.D. degrees in electrical engineering from Washington University, St. Louis, MO, in 1992 and 1995 respectively. From 1995 to 1999, he held a position at the Mathematics of Communications Department at Bell Labs. Since November 1999, he has been a faculty member at the School of Computer and Communication Sciences of EPFL, Lausanne, Switzerland. Dr. Urbanke is a recipient of a Fulbright Scholarship. From 2000–2004 he was an Associate Editor of the IEEE Transactions on Information Theory and he is currently on the board of the series 'Foundations and Trends in Communications and Information Theory'. He is a co-recipient of the IEEE Information Theory Society 2002 Best Paper Award and of the 2011 IEEE Koji Kobayashi award. He co-authored the book *Modern Coding Theory* published by Cambridge University Press.