IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received September 11, 2018, accepted October 13, 2018. Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2018.2877720

Serially Concatenated Polar Codes

ERDAL ARıKAN, (Fellow, IEEE)
Department of Electrical and Electronics Engineering, Bilkent University, Ankara, Turkey
e-mail: arikan@ee.bilkent.edu.tr

AQ:1 This work was supported by Huawei Technologies, Co., Ltd. under Grant HE2017110001.

ABSTRACT Simulation results show that the performance of polar codes is improved vastly by using polar codes as inner codes in serially concatenated coding schemes. Furthermore, this performance improvement is achieved using a relatively short cyclic redundancy check as the outer code and a practically implementable successive cancellation list decoder for decoding the overall code. This paper offers a theoretical analysis of such schemes by employing a random-coding method on the selection of the outer code and assuming that the channel is memoryless. It is shown that the probability of error for the concatenated coding scheme decreases exponentially in the code block length at any fixed rate below the symmetric capacity. Applications of this result include the design of polar codes for communication systems that require high reliability at small to moderate code lengths, such as control channels in wireless systems and machine-type communications for industrial automation.

INDEX TERMS Polar codes, serial concatenation, list decoding, error exponents.

I. INTRODUCTION

Polar codes are a class of linear block codes that achieve the capacity of certain classes of channels with explicit code constructions and practical encoding and decoding algorithms [1]. Early on, performance studies on polar codes with a low-complexity successive cancellation (SC) decoder revealed that the performance of polar codes was not on par with that of the state-of-the-art codes such as turbo and LDPC codes. The disappointing performance of polar codes could partly be attributed to the suboptimal nature of the SC decoder. Indeed, in [2], Tal and Vardy showed that the performance of polar codes could be improved significantly by using a successive cancellation list (SCL) decoder, which is a modified SC decoder, originally devised by Dumer and Shabunov [3] and Dumer [4] for Reed-Muller codes. An SCL decoder with list size L tracks a list of L candidate codewords and picks the most probable one as its decision in the final stage of decoding.

To be more specific, Fig. 1 presents a simulation study, originally from [2], in which the code is a polar code with rate $R = 1/2$ and block-length $N = 2048$, the modulation is quaternary Quadrature Amplitude Modulation (4-QAM), and the channel is an additive Gaussian noise channel. The frame (block) error rate (FER) is plotted as a function of the signal-to-noise ratio (SNR). We observe that SCL-32 (SCL with list-size 32) decoding provides significantly better performance compared to SC decoding especially at low to moderate SNR. Except for low SNR (close to channel capacity), increasing the list size from 32 to 1024 provides only marginal improvements. It is remarkable that the SCL decoder (even at list size 32) achieves near ML performance across a broad range of SNR. (The ML bound shown in the figure is obtained empirically by counting the number of times the SCL-1024 decoder produces a decision that is closer to the received word than the transmitted codeword is.)

On the bright side, the above simulation results promise that SCL decoders may achieve near ML performance with a practically feasible list size. On the other hand, the performance of polar codes at high SNR appears unsatisfactory even under ML decoding. The poor performance of polar codes at high SNR can be blamed on their poor minimum distance, which grows as $\mathcal{O}(\sqrt{N})$ as a function of the code block-length N at any fixed rate $0 < R < 1$, whereas optimal codes have a minimum distance that grows linearly with N. It appears that any method to improve the performance of polar codes beyond their native ML performance has to address the deficiency of the minimum distance of polar codes.

Tal and Vardy [2] provided a fix to this problem by introducing a cyclic redundancy check (CRC) into the data before it was encoded by the polar encoder and modified the SCL decoder so that at the end of decoding candidate

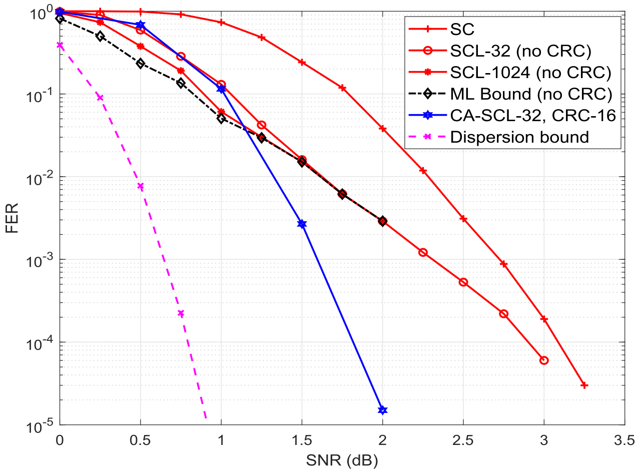

FIGURE 1. **Performance of polar codes.**

codewords that did not satisfy the CRC could be discarded. The CRC helped reduce the chances that the SCL decoder would make decision errors to near neighbors of the transmitted codeword. The remarkable performance improvement under this type of CRC-aided SCL (CA-SCL) decoding is shown in Fig. 1. Indeed, polar codes under CA-SCL decoding with list size $L = 8$ and CRC-length 24 were powerful and practical enough to be included as a coding scheme in a recent 3GPP NR standard [5].

Also shown in Fig. 1 is the theoretical limit (dispersion bound) [6] for any code of length 2048 and rate 1/2. Despite the performance improvement by CA-SCL decoding, there is still a substantial gap between the dispersion bound and the performance of polar coding under CA-SCL-32 decoding. Li *et al.* [7] investigated whether this gap could be closed by using CA-SCL decoders with larger list sizes. They carried out simulations using a CA-SCL decoder with a 24-bit CRC on a rate-1/2 polar code of length 2048 and observed that the performance at list size 262,144 came to within 0.2 dB of the dispersion bound at FER 10^{-3}.

This paper is motivated by the desire to provide a theoretical explanation for the above empirical findings. We study this problem in a more general setting by regarding the polar code and the CRC as the inner and outer codes, respectively, in a serially concatenated coding scheme. Such serially concatenated coding schemes are relevant in other contexts as well. For example, Narayanan and Stuber [8] used a serially concatenated coding scheme in which the outer code was a BCH or Reed-Solomon code and the inner code was a turbo code. They investigated list decoding in such a system to reduce the error floor of turbo codes. As another example, one may regard the Viterbi decoding algorithm for convolutional codes as a list-decoder in a concatenation scheme in which the termination bits of the convolutional code play the role of a CRC. Despite the importance of such concatenation and list decoding techniques in practical applications, the subject does not appear to have received sufficient theoretical attention; this is true at least in the context of polar codes. The goal of this paper is to fill this gap to some extent. The paper builds on some of our earlier results of [9] and extends them using some new techniques.

The rest of the paper is organized as follows. Section II contains a precise formulation of the problem considered in this paper and a statement of the main result. Section III gives a proof of the main result. The paper concludes with some remarks in Section IV.

II. PROBLEM FORMULATION AND THE MAIN RESULT

The scope of the paper is limited to linear codes over the binary field $\mathbb{F}_2 = \{0, 1\}$. We use bits as the unit of information and use base-two logarithms (denoted log) throughout. The rate of a code with M codewords and length N is defined as $R = (1/N)\log M$. The notation $[N]$ denotes the set of integers 1 through N. Vectors used in the paper are row vectors and are denoted by boldface letters such as $\mathbf{x}$. For $\mathbf{x} = (x_1, \ldots, x_N)$ a row vector of length N and $\mathcal{A} \subset [N]$, the notation $\mathbf{x}_{\mathcal{A}}$ denotes the subvector $(x_i : i \in \mathcal{A})$, with the elements of $\mathbf{x}_{\mathcal{A}}$ listed in increasing order of $i \in \mathcal{A}$.

A. SYSTEM MODEL

We will be studying a serially concatenated coding system as shown in Fig. 2. The channel in the system will be a binary-input discrete memoryless channel with input alphabet $\mathcal{X} = \{0, 1\}$, a finite but otherwise arbitrary output alphabet $\mathcal{Y}$, and transition probabilities $\{W(y|x) : x \in \mathcal{X}, y \in \mathcal{Y}\}$. We will be concerned with achieving the symmetric capacity of such channels, which is defined as

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

Details of the encoding and decoding operations in Fig. 2 are as follows. The input to the system is a transmitted word $\mathbf{d} = (d_1, \ldots, d_K) \in \mathbb{F}_2^K$. The outer code is an arbitrary linear code with dimension K, block-length K_{in}, and generator matrix $\mathbf{G}_{\text{out}} \in \mathbb{F}^{K\times K_{\text{in}}}$. The outer encoder maps $\mathbf{d}$ into an outer codeword $\mathbf{v} = (v_1, \ldots, v_{K_{\text{in}}}) \in \mathbb{F}_2^{K_{\text{in}}}$ by computing $\mathbf{v} = \mathbf{d}\mathbf{G}_{\text{out}}$.

The inner code is a polar code with dimension K_{in}, block-length $N = 2^n$ (for some $n \geq 1$), a transform matrix $\mathbf{G}_N = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}^{\otimes n}$ (the nth Kronecker power), and a frozen set $\mathcal{F} \subset [N]$ with $N - K_{\text{in}}$ elements. The inner encoder maps the outer codeword $\mathbf{v}$ to a polar codeword $\mathbf{x} = (x_1, \ldots, x_N) \in \mathbb{F}_2^N$ by computing $\mathbf{x} = \mathbf{u}\mathbf{G}_N$, where the transform input $\mathbf{u} \in \mathbb{F}^N$ is prepared by setting $\mathbf{u}_{\mathcal{F}} = \mathbf{0}$ (an all-zero word) and $\mathbf{u}_{\mathcal{F}^c} = \mathbf{v}$ ($\mathcal{F}^c$ is the complement of $\mathcal{F}$ in $[N]$).

The codeword $\mathbf{x}$ is transmitted over the channel W and a channel output $\mathbf{y} \in \mathcal{Y}^N$ is produced. The decoder receives $\mathbf{y}$ and produces an estimate $\hat{\mathbf{d}} \in \mathbb{F}_2^K$ of the transmitted message $\mathbf{d}$. The goal of the system is to have $\hat{\mathbf{d}} = \mathbf{d}$ with as high a probability as possible. Since we are interested in the best achievable performance with such systems, we will assume that the decoder is an ML decoder.

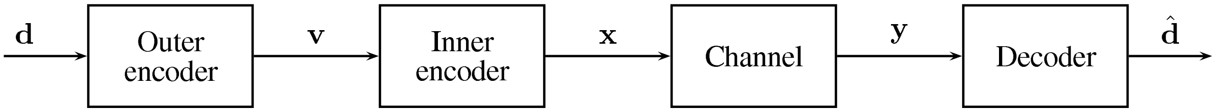

FIGURE 2. Serially concatenated polar code with a linear outer code.

B. PROBABILISTIC MODEL

Here, we specify a probabilistic model for the above system. This involves introducing randomness into the inner and outer code so as to make the analysis tractable.

For the outer code, we use the standard random code ensemble for linear codes, as the one in [10, p. 206]. This ensemble is characterized by a pair $(\mathbf{G}_{\text{out}}, \mathbf{C})$ where $\mathbf{G}_{\text{out}}$ is a random generator matrix of size $K \times K_{\text{in}}$ over $\mathbb{F}_2$ and $\mathbf{C}$ is a random offset word of length K_{in} over $\mathbb{F}_2$. Specific samples of $(\mathbf{G}_{\text{out}}, \mathbf{C})$ are denoted by $(\mathbf{g}_{\text{out}}, \mathbf{c})$. The two parameters are assumed independent, $\Pr(\mathbf{G}_{\text{out}} = \mathbf{g}_{\text{out}}, \mathbf{C} = \mathbf{c}) = \Pr(\mathbf{G}_{\text{out}} = \mathbf{g}_{\text{out}}) \Pr(\mathbf{C} = \mathbf{c})$, and uniformly distributed over their respective ranges, $\Pr(\mathbf{G}_{\text{out}} = \mathbf{g}_{\text{out}}) = 2^{-KK_{\text{in}}}$ and $\Pr(\mathbf{C} = \mathbf{c}) = 2^{-K_{\text{in}}}$ for any particular encoder setting $(\mathbf{g}_{\text{out}}, \mathbf{c})$.

The offset vector $\mathbf{C}$ ensures that the outer code has the pairwise-independence property, namely, the property that, for any two distinct data words, $\mathbf{d}, \mathbf{d}' \in \mathbb{F}_2^K$, $\mathbf{d} \neq \mathbf{d}'$,

$$\Pr\big(\mathbf{V}(\mathbf{d}) = \mathbf{v}, \mathbf{V}(\mathbf{d}') = \mathbf{v}'\big) = 2^{-2K_{\text{in}}}, \tag{1}$$

where $\mathbf{V}(\mathbf{d}) = \mathbf{d}\mathbf{G}_{\text{out}} + \mathbf{C}$ and $\mathbf{V}(\mathbf{d}') = \mathbf{d}'\mathbf{G}_{\text{out}} + \mathbf{C}$ are the codewords corresponding to $\mathbf{d}$ and $\mathbf{d}'$.

The analysis that follows will be valid for any ensemble of outer codes satisfying the pairwise-independence property. We will hide all other aspects of the outer code in the following analysis and view the outer code as a list of codewords $\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_{2^K}$.

We also simplify the representation of input data and instead of data vectors $\mathbf{d} \in \mathbb{F}^K$ use integers $m \in [2^K]$ to represent messages carried by the system. The integer m in turn is regarded as a realization of a message random variable M, distributed uniformly over the message set $[2^K]$. The message estimates at the output of the decoder will be denoted by a random variable $\hat{M}$ and realizations of $\hat{M}$ by $\hat{m}$.

The inner code has been specified above as a polar code with a frozen part $\mathbf{u}_{\mathcal{F}}$ equal to $\mathbf{0}$. For the analysis, we use randomization over the frozen part and set $\mathbf{u}_{\mathcal{F}} = \mathbf{a}$ where $\mathbf{a}$ is a sample of a random word $\mathbf{A}$ taking values uniformly at random over $\mathbb{F}_2^{N-K_{\text{in}}}$. The inner encoding operation for a particular outer codeword $\mathbf{v}$ and a frozen word $\mathbf{a}$ takes the form $\mathbf{x} = \mathbf{u}\mathbf{G}_N$ with $\mathbf{u}_{\mathcal{F}} = \mathbf{a}$ and $\mathbf{u}_{\mathcal{F}^c} = \mathbf{v}$. Each realization $\mathbf{a}$ of $\mathbf{A}$ defines a specific code in an ensemble of $2^{N-K_{\text{in}}}$ polar codes.

With these randomizations, we now have a joint ensemble

$$(M, \mathbf{V}_1, \ldots, \mathbf{V}_{2^K}, \mathbf{A}, \mathbf{X}, \mathbf{Y}, \hat{M})$$

representing all parts of the system. The joint probability mass function (PMF) is of the form

$$\begin{aligned} p(m, \mathbf{v}_1, \ldots, \mathbf{v}_{2^K}, \mathbf{a}, \mathbf{x}, \mathbf{y}, \hat{m}) &= p(m) p(\mathbf{v}_1, \ldots, \mathbf{v}_{2^K}) p(\mathbf{a}) \\ &\quad \times p(\mathbf{x}|\mathbf{a}, \mathbf{v}_m) p(\mathbf{y}|\mathbf{x}) p(\hat{m}|\mathbf{v}_1, \ldots, \mathbf{v}_{2^K}, \mathbf{a}, \mathbf{y}). \end{aligned}$$

For notational simplicity, we omitted the names of the random variables from the PMFs, writing $p(m)$ instead of $p_M(m)$ and $p(\mathbf{y}|\mathbf{x})$ instead of $p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})$, etc.

The structure of the joint PMF shows that the message M, the outer code $\{\mathbf{V}_1, \ldots, \mathbf{V}_{2^K}\}$, and the frozen word $\mathbf{A}$ are jointly independent. The joint model inherits the pairwise independence property of the outer code: $p(\mathbf{v}_i, \mathbf{v}_j) = p(\mathbf{v}_i)p(\mathbf{v}_j) = 2^{-2K_{\text{in}}}$ for any $i, j \in [2^K]$, $i \neq j$. The inner encoder is characterized by the conditional PMF $p(\mathbf{x}|\mathbf{a}, \mathbf{v})$, which equals 1 if $\mathbf{x} = \mathbf{u}\mathbf{G}_N$ for $\mathbf{u}_{\mathcal{F}} = \mathbf{a}$ and $\mathbf{u}_{\mathcal{F}^c} = \mathbf{v}$, and equals 0 otherwise. The channel PMF $p(\mathbf{y}|\mathbf{x})$ equals $\prod_{i=1}^N W(y_i|x_i)$. The decoder PMF $P(\hat{m}|\mathbf{v}_1, \ldots, \mathbf{v}_{2^K}, \mathbf{a}, \mathbf{y})$ equals 1 if the specific ML decoder in the system produces $\hat{m}$ in response to channel output $\mathbf{y}$, and equals 0 otherwise. The presence of $\{\mathbf{v}_1, \ldots, \mathbf{v}_{2^K}, \mathbf{a}\}$ as part of the conditioning variables in the decoder PMF signifies that the ML decoder operates with knowledge of the encoder settings for the outer and inner codes. In the following analysis, we will use the notation $\Pr(E)$ to denote the probability of an arbitrary event E according to the joint probability model above.

The probability of ML decoding error for a given setting of the encoder parameters given by

$$\begin{aligned} P_e(\mathbf{v}_1, \ldots, \mathbf{v}_{2^K}, \mathbf{a}) &\overset{\Delta}{=} \Pr\big(\hat{M} \neq M \big| \mathbf{v}_1, \ldots, \mathbf{v}_{2^K}, \mathbf{a}\big) \\ &= \sum_m \sum_{\hat{m} \neq m} p\big(m, \hat{m} \big| \mathbf{v}_1, \ldots, \mathbf{v}_{2^K}, \mathbf{a}\big). \end{aligned}$$

The average probability of ML decoding error over the ensemble of all encoder settings is given by

$$\overline{P}_e \overset{\Delta}{=} \sum_{\mathbf{v}_1, \ldots, \mathbf{v}_{2^K}, \mathbf{a}} p\big(\mathbf{v}_1, \ldots, \mathbf{v}_{2^K}\big) p(\mathbf{a}) P_e(\mathbf{v}_1, \ldots, \mathbf{v}_{2^K}, \mathbf{a}).$$

This completes the problem formulation. In the rest of the paper, our goal will be to derive upper-bounds on $\overline{P}_e$ by using random-coding methods.

C. THE MAIN RESULT

The main result of the paper is an upper bound on $\overline{P}_e$ under certain constraints on the target data rates for the inner and outer codes in the concatenated coding scheme. We will denote the target rates for the inner, outer, and the overall code by R_{in}, R_{out}, and R, respectively. For consistency we will require that $R = R_{\text{in}}R_{\text{out}}$. Clearly, in a serially concatenated coding scheme, we also have to have $R_{\text{in}} \ge R$ and $R_{\text{out}} \ge R$.

For a given target rate $R < I(W)$, there is a wide range of possible choices for R_{out} and R_{in} satisfying $R = R_{\text{out}}R_{\text{in}}$. Our primary interest is in cases where $R_{\text{in}} < I(W)$ and $R_{\text{out}} \approx 1$. By having $R_{\text{in}} < I(W)$, we wish to ensure that the inner code can be decoded using a low-complexity decoder designed for polar codes. By having $R_{\text{out}} \approx 1$, we desire to have a lightweight outer code. The main result below will cover these cases of interest.

Theorem 1: Consider serially concatenated coding with an inner polar code on a binary-input memoryless channel W with a strictly positive symmetric capacity, $I(W) > 0$. Let $(R_{\text{in}}, R_{\text{out}}, R)$ be the desired rates such that $R = I(W) - \gamma$, $R_{\text{in}} = I(W) - \gamma_{\text{in}}$, and $0 < \gamma_{\text{in}} < \gamma$. Consider the class of concatenated coding ensembles with parameter set (K, K_{in}, N) such that $K = \lfloor NR \rfloor$, $K_{\text{in}} = \lfloor NR_{\text{in}} \rfloor$, and $N = 2^n$ for some n. The average probability of error for any such ensemble satisfies $\overline{P}_e \le 2^{-Nf(R+o(N))}$ where f is a function independent of N, $f(\tilde{R}) > 0$ for all $0 \le \tilde{R} < I(W)$, and $o(N)$ is a quantity that depends on W but goes to zero as N increases.

III. PROOF OF THEOREM 1

We split the proof into two parts. In Section III-A, a method from [11] is used to upper bound the average probability of error $\overline{P}_e$. In Section III-B, the bound of Section III-A is reduced to a single-letter form and the proof is completed.

A. THRESHOLD DECODER BOUND

Consider a serially concatenated code ensemble whose parameters (K, K_{in}, N) satisfy the hypothesis of Theorem 1. Let $\overline{P}_e$ denote the ML probability of error for this ensemble. We will upperbound $\overline{P}_e$ by considering the performance of a suboptimal threshold decoder that is easier to analyze. Given the channel output $\mathbf{y}$, the threshold decoder that we consider here computes the metric

$$i(\mathbf{y}; \mathbf{v}_j|\mathbf{a}) = \log \frac{p(\mathbf{y}|\mathbf{v}_j, \mathbf{a})}{p(\mathbf{y}|\mathbf{a})}$$

for each message $j \in [2^K]$. The computed metrics are then compared against a threshold T. If there is only one message j such that $i(\mathbf{y}; \mathbf{v}_j|\mathbf{a}) > T$ is true, the decoder declares its decision as $\hat{m} = j$; in all other cases, a decoder error is declared. In the rest of the discussion, the threshold will be fixed as $T = N(R + \theta)$ where R is the target rate for the overall coding scheme and $\theta > 0$ is an arbitrary constant.

Proceeding to the random-coding analysis of the threshold decoder, we define $\mathcal{E}_j \overset{\Delta}{=} \{i(\mathbf{Y}; \mathbf{V}_j|\mathbf{A}) \le N(R + \theta)\}$ for each $j \in [2^K]$. Conditional on the transmitted message being m (the event $\{M = m\}$), the threshold decoder makes an error if $\mathcal{E}_m$ or $\cup_{m' \ne m}\mathcal{E}^c_{m'}$ occurs. Let $\overline{P'}_e$ denote the probability of error by the threshold decoder and let $\overline{P'}_{e,m}$ denote the conditional probability of error by the threshold decoder given that message m is transmitted.

$$\begin{aligned}
\overline{P}_e \le \overline{P'}_e &= \sum_m p(m)\overline{P'}_{e,m} \\
&\le \sum_m p(m)\Big[\Pr(\mathcal{E}_m|M = m) + \Pr(\bigcup_{m' \ne m} \mathcal{E}^c_{m'}|M = m)\Big] \\
&\le \sum_m p(m)\Big[\Pr(\mathcal{E}_m|M = m) + \sum_{m' \ne m} \Pr(\mathcal{E}^c_{m'}|M = m)\Big] \\
&= \Pr(\mathcal{E}_1|M = 1) + (2^K - 1)\Pr(\mathcal{E}^c_2|M = 1) \qquad (2)
\end{aligned}$$

where (2) follows by observing that $\Pr(\mathcal{E}_m|\{M = m\})$ and $\Pr(\mathcal{E}^c_{m'}|\{M = m\})$ do not depend on the particular choice of m and $m' \ne m$. We now bound each of these error terms.

For the first type of error, we have

$$\begin{aligned}
\Pr(\mathcal{E}_1|M = 1) &= \Pr\big[i(\mathbf{Y}; \mathbf{V}_1|\mathbf{A}) \le N(R + \theta)\big|M = 1\big] \\
&= \Pr\big[i(\mathbf{Y}; \mathbf{X}) - i(\mathbf{Y}; \mathbf{A}) \le N(R + \theta)\big]. \qquad (3)
\end{aligned}$$

In writing (3), we have used the identities

$$\begin{aligned}
i(\mathbf{Y}; \mathbf{V}_1|\mathbf{A}) &= i(\mathbf{Y}; \mathbf{V}_1, \mathbf{A}) - i(\mathbf{Y}; \mathbf{A}) \\
&= i(\mathbf{Y}; \mathbf{U}) - i(\mathbf{Y}; \mathbf{A}) \\
&= i(\mathbf{Y}; \mathbf{X}) - i(\mathbf{Y}; \mathbf{A})
\end{aligned}$$

where $\mathbf{U}$ and $\mathbf{X}$ are related by the polar transform $\mathbf{X} = \mathbf{U}\mathbf{G}_N$ and $\mathbf{U}$ is composed of $\mathbf{U}_{\mathcal{F}} = \mathbf{A}$ and $\mathbf{U}_{\mathcal{F}^c} = \mathbf{V}_1$. Since $\mathbf{X}$ and M are independent, the conditioning on $\{M = 1\}$ was dropped in (3).

For the second type of error, we have

$$\begin{aligned}
\Pr(\mathcal{E}^c_2|M = 1) &= \Pr\big[i(\mathbf{Y}; \mathbf{V}_2|\mathbf{A}) > N(R + \theta)\big|M = 1\big] \\
&= \sum_{\mathbf{v}_2, \mathbf{a}, \mathbf{y}} p(\mathbf{v}_2)p(\mathbf{y}|\mathbf{a})\mathbb{1}\big(\{i(\mathbf{y}; \mathbf{v}_2|\mathbf{a}) > N(R+\theta)\}\big) \\
&\le \sum_{\mathbf{v}_2, \mathbf{a}, \mathbf{y}} p(\mathbf{v}_2)p(\mathbf{y}|\mathbf{a})\frac{p(\mathbf{y}|\mathbf{v}_2, \mathbf{a})}{p(\mathbf{y}|\mathbf{a})}2^{-N(R+\theta)} \\
&= 2^{-N(R+\theta)}. \qquad (4)
\end{aligned}$$

where $\mathbb{1}(\cdot)$ is the indicator function of the enclosed event. Combining (3) and (4) and noting that $2^K - 1 < 2^K \le 2^{NR}$, the bound (2) yields

$$\overline{P}_e \le \Pr\big[i(\mathbf{Y}; \mathbf{X}) - i(\mathbf{Y}; \mathbf{A}) \le N(R + \theta)\big] + 2^{-N\theta}. \qquad (5)$$

This bound is a generalization of a similar bound in [11, Th. 1]; the two bounds become identical when $\mathbf{A}$ is a null vector (the frozen set $\mathcal{F}$ is empty). Note that the bound (5) does not have a single-letter form and it is not clear yet if the bound decreases exponentially as N is increased. To resolve this question, we proceed to derive a single-letter form of the bound.

B. SINGLE-LETTER FORM OF THE BOUND

In this part, we use the assumption that the channel is memoryless and simplify the bound (5) to a single-letter expression. The task is to prove that the event

$$\mathcal{A} \overset{\Delta}{=} \{i(\mathbf{Y}; \mathbf{X}) - i(\mathbf{Y}; \mathbf{A}) \le N(R + \theta)\}$$

has a probability that decreases to zero exponentially in N. To that end, let

$$\mathcal{B} \triangleq \{i(\mathbf{Y};\mathbf{A}) \geq N\delta\},$$

with

$$\delta \triangleq \frac{1}{N} I(\mathbf{Y};\mathbf{A}) + \lambda, \quad \lambda > 0. \tag{6}$$

(Note that $I(\mathbf{Y};\mathbf{A}) = \mathbb{E}[i(\mathbf{Y};\mathbf{A})]$.) We now write

$$\begin{aligned}\Pr(\mathcal{A}) &= \Pr\left[\mathcal{A}\cap\mathcal{B}\right) + \Pr(\mathcal{A}\cap\mathcal{B}^c\right]\\ &\leq \Pr(\mathcal{B}) + \Pr\left[i(\mathbf{Y};\mathbf{X}) \leq N(R+\theta+\delta)\right]\end{aligned} \tag{7}$$

We will show that each term on the right hand side of (7) decreases to zero exponentially in N.

For the first term, we use McDiarmid's inequality to show that

$$\Pr(\mathcal{B}) \leq e^{-2N\lambda^2/\alpha} \tag{8}$$

where α is a constant that depends on the channel W but is independent of N. Details are given in Appendix.

The second term $\Pr\left[i(\mathbf{Y};\mathbf{X}) \leq N(R+\theta+\delta)\right]$ is readily upper-bounded by noting that $i(\mathbf{Y};\mathbf{X})$ for a memoryless channel is the sum of i.i.d. random variables: $i(\mathbf{Y};\mathbf{X}) = \sum_{j=1}^{N} i(Y_j;X_j)$. Using the Chernoff bound (see [11] or [10, Eq. 5.4.12]), we obtain

$$\Pr\left[i(\mathbf{Y};\mathbf{X}) \leq N(R+\theta+\delta)\right] \leq 2^{N[\tilde{\mu}(s)-s(R+\theta+\delta)]}, \tag{9}$$

which is valid for $s < 0$. Here, $\tilde{\mu}(s)$ is the semi-logarithmic moment generating function for the random variable $i(Y_j;X_j)$,

$$\begin{aligned}\tilde{\mu}(s) &\triangleq \log \sum_{x_j,y_j} p(x_j,y_j) 2^{si(Y_j;X_j)}\\ &= \log \sum_{x_j,y_j} p(x_j)p(y_j|x_j)^{1+s}p(y_j)^{-s}.\end{aligned}$$

(Clearly, the value of $\tilde{\mu}(s)$ does not depend on the index $j \in [N]$ since the channel is memoryless. The function $\tilde{\mu}(s)$ defined here is related to the function $\mu(s)$ in [11] by $\tilde{\mu}(s) = \mu(s)\log 2$. We use $\tilde{\mu}(s)$ here since we have chosen bits instead of nats as the unit of information.)

Optimizing the bound (9) over s, we obtain

$$\Pr\left[i(\mathbf{Y};\mathbf{X}) \leq N(R+\theta+\delta)\right] \leq 2^{-NE(R+\theta+\delta)}, \tag{10}$$

where

$$E(R') \triangleq -\inf_{s<0}\,[\tilde{\mu}(s) - sR'].$$

Shannon [11] shows that $E(R') > 0$ provided $I(W) > 0$ and $R' < I(W)$. Combining and (7), (8), and (10), we have

$$\overline{P}_e \leq 2^{-N\theta} + 2^{-NE(R+\theta+\delta)} + e^{-2N\lambda^2/\alpha}. \tag{11}$$

Until now, the analysis did not make use of the assumption that the inner code is a polar code. Now, we use this assumption. Since $\mathbf{A} = \mathbf{U}_{\mathcal{F}}$, we may write $I(\mathbf{Y};\mathbf{A}) = I(\mathbf{Y};\mathbf{U}_{\mathcal{F}})$. By the channel polarization theorem of [1] or [12], we can choose the frozen set $\mathcal{F}$ so that

$$\frac{1}{N} I(\mathbf{Y};\mathbf{U}_{\mathcal{F}}) = I(W) - R_{\text{in}} + o(N) = \gamma_{\text{in}} + o(N).$$

Thus,

$$\delta = \frac{1}{N} I(\mathbf{Y};\mathbf{A}) + \lambda = \frac{1}{N} I(\mathbf{Y};\mathbf{U}_{\mathcal{F}}) + \lambda = \gamma_{\text{in}} + o(N).$$

Substituting this in (11), we obtain

$$\overline{P}_e \leq 2^{-N\theta} + 2^{-NE(R+\theta+\lambda+\gamma_{\text{in}}+o(N))} + e^{-2N\lambda^2/\alpha}. \tag{12}$$

Optimizing (11) over θ and λ appears infeasible and unnecessary in view of the ad hoc nature of the bound. Instead, we set $\theta = \lambda = (\gamma - \gamma_{\text{in}})/4$. Then, the bound (12) becomes

$$\overline{P}_e \leq 2^{-\frac{N(\gamma-\gamma_{\text{in}})}{4}} + 2^{-NE[R+\frac{\gamma+\gamma_{\text{in}}}{2}+o(N)]} + e^{-\frac{N(\gamma-\gamma_{\text{in}})^2}{8\alpha}}. \tag{13}$$

Since $\gamma - \gamma_{\text{in}} > 0$, the first and third terms on the right side of (13) go to zero exponentially in N. Since $R + \frac{\gamma+\gamma_{\text{in}}}{2} = I(W) - (\gamma - \gamma_{\text{in}})/2 < I(W)$, the second term on the right side of (13) also goes to zero exponentially in N. The function f in the statement of Theorem 1 may be taken as

$$f(\tilde{R}) \triangleq -\frac{1}{N}\log\left[2^{-\frac{N(\gamma-\gamma_{\text{in}})}{4}} + 2^{-NE(\tilde{R}+\frac{\gamma+\gamma_{\text{in}}}{2})} + e^{-\frac{N(\gamma-\gamma_{\text{in}})^2}{8\alpha}}\right].$$

This completes the proof.

IV. CONCLUDING REMARKS

We conclude the paper with some complementary remarks. Theorem 1 showed that the probability of error for serially concatenated coding with an inner polar code goes to zero exponentially in N provided that the target rate R is less than the symmetric capacity $I(W)$. This result was proved under the additional constraint $R_{\text{in}} < I(W)$ on the rate of the inner polar code. The constraint $R_{\text{in}} < I(W)$ was placed to leave open the possibility that a low-complexity decoder can be used to decode the inner polar code, in anticipation that the ML performance guaranteed by Theorem 1 can perhaps be approximated in practice. Overall, Theorem 1 and its proof provide some insight into why CA-SCL achieves vastly superior performance compared to a stand-alone polar code.

The proof of Theorem 1 relied heavily on techniques from [11], which gives a bound to ML performance for standalone codes without any concatenation (equivalent to having $R_{\text{in}} = 1$ or equivalently $\mathcal{F} = \emptyset$ in our framework). It is of interest to compare the bounds here with those in [11]. For this, let $\overline{P}_{e,\emptyset}$ denote $\overline{P}_e$ for the special case $\mathcal{F} = \emptyset$. Shannon [11] shows that

$$\overline{P}_{e,\emptyset} \leq 2^{-N\theta} + 2^{N[\tilde{\mu}(s)-s(R+\theta))]}, \quad s < 0. \tag{14}$$

The comparable bound for the concatenated coding scheme is

$$\overline{P}_e \leq 2^{-N\theta} + 2^{N[\tilde{\mu}(s)-s(R+\theta+\delta)]} + e^{-2N\lambda^2/\alpha}, \quad s < 0, \tag{15}$$

which is obtained by combining (5), (8), and (9). Comparing these bounds, the cost of concatenation becomes visible. The bound is worsened by the inclusion of an extra term $e^{-2N\lambda^2/\alpha}$ and the inflation of the effective code rate from R to $R+\delta$.

Finally, a case of interest is when $R+\theta+\delta$ is near $I(W)$. In that case (see [11]),

$$\inf_{s<0}[\mu(s) - s(R+\theta+\delta)] \doteq -\frac{\big(I(W)-(R+\theta+\delta)\big)^2}{2\mu''(0)} \tag{16}$$

where $\mu''(0)$ is the variance of channel mutual information random variable $i(X_j; Y_j)$. Thus, the second term on the right side of (15) has an exponent that is quadratic in $(I(W) - R - \theta - \delta)$. The quadratic form of the exponent at rates near capacity replicates the behavior of optimal ensembles (see [10, Prob. 5.23, p. 539]); however, the term δ appears again as a penalty term that reduces the effective gap-to-capacity and worsens the exponent.

In summary, the price of having a structured inner code that can be decoded at low complexity is captured by the parameter $\delta = \gamma_{\text{in}} + \lambda \approx \gamma_{\text{in}}$. The larger γ_{in} is, the more structure there is in the concatenated coding scheme, and the worse the error exponent becomes. Still, the remarkable fact that should stand out at the end of this study is that polar codes, when concatenated with an outer code of rate $R_{\text{out}} \approx 1$, can achieve rates $R < I(W)$ with a probability of error that goes to zero exponentially in the block-length N.

APPENDIX

PROOF OF (8)

We will use McDiarmid's inequality, also known as the method of bounded differences (see [13, p. 20]).

First, we note that the frozen word $\mathbf{A}$ can be obtained from the transmitted codeword $\mathbf{X}$ simply by computing the inverse transform $\mathbf{U} = \mathbf{X}\mathbf{G}_N^{-1}$ and looking at the frozen part $\mathbf{U}_{\mathcal{F}}$. The computation of $\mathbf{A}$ from $\mathbf{X}$ is in fact a linear operation of the form $\mathbf{A} = \mathbf{X}\mathbf{H}$ where $\mathbf{H} = \left(\mathbf{G}_N^{-1}\right)_{\mathcal{F}}$ is the submatrix of $\mathbf{G}_N^{-1}$ consisting of columns with indices in $\mathcal{F}$. Thus, $i(\mathbf{Y}; \mathbf{A})$ is a function of the input-output pair $(\mathbf{X}, \mathbf{Y})$ of the channel W; specifically, $i(\mathbf{Y}; \mathbf{A}) = g(\mathbf{X}; \mathbf{Y})$ with $g(\mathbf{X}; \mathbf{Y}) = i(\mathbf{Y}; \mathbf{X}\mathbf{H})$. Furthermore, the argument $(\mathbf{X}, \mathbf{Y})$ of the function g consists of i.i.d. pairs of random variables $(X_j, Y_j), j \in [N]$.

Next, we show that g is Lipschitz in the following sense. Let $(\mathbf{x}, \mathbf{y})$ and $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ be two pairs from $\mathcal{X}^N \times \mathcal{Y}^N$ such that (i) $(x_i, y_i) \neq (\tilde{x}_i, \tilde{y}_i)$ for some i but $(x_j, y_j) = (\tilde{x}_j, \tilde{y}_j)$ for all $j \neq i$, with $i, j \in [N]$, and (ii) $p(\mathbf{x}, \mathbf{y}) > 0$ and $p(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) > 0$. The function g is Lipschitz in the sense that

$$\left| g(\mathbf{x}, \mathbf{y}) - g(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) \right| \leq \Delta_i, \tag{17}$$

for some constant Δ_i that depends only on the distribution $p(x_i, y_i)$.

We now show that (17) holds. Instead of (17), it is more convenient to consider the equivalent expression

$$2^{-\Delta_i} \leq 2^{g(\mathbf{x},\mathbf{y}) - g(\tilde{\mathbf{x}},\tilde{\mathbf{y}})} \leq 2^{\Delta_i}$$

and note that

$$2^{g(\mathbf{x},\mathbf{y}) - g(\tilde{\mathbf{x}},\tilde{\mathbf{x}})} = \frac{p(\mathbf{y}|\mathbf{a})}{p(\tilde{\mathbf{y}}|\tilde{\mathbf{a}})} \, \frac{p(\tilde{\mathbf{y}})}{p(\mathbf{y})}, \tag{18}$$

where $\mathbf{a} = \mathbf{x}\mathbf{H}$ and $\tilde{\mathbf{a}} = \tilde{\mathbf{x}}\mathbf{H}$ are the frozen words corresponding to $\mathbf{x}$ and $\tilde{\mathbf{x}}$, respectively. Let $\mathcal{C} = \{\bar{\mathbf{x}} \in \mathbb{F}_2^N : \bar{\mathbf{x}}\mathbf{H} = \mathbf{0}\}$ where $\mathbf{0}$ is the all-zero vector. We can now write the first factor on the right hand side of (18) as

$$\frac{p(\mathbf{y}|\mathbf{a})}{p(\tilde{\mathbf{y}}|\tilde{\mathbf{a}})} = \frac{\sum_{\bar{\mathbf{x}} \in \mathcal{C}} p(\mathbf{x} + \bar{\mathbf{x}}|\mathbf{a})\, p(\mathbf{y}|\mathbf{x} + \bar{\mathbf{x}})}{\sum_{\bar{\mathbf{x}} \in \mathcal{C}} p(\tilde{\mathbf{x}} + \bar{\mathbf{x}}|\tilde{\mathbf{a}})\, p(\tilde{\mathbf{y}}|\tilde{\mathbf{x}} + \bar{\mathbf{x}})}.$$

Term by term, we have the bound

$$\begin{aligned}\frac{p(\mathbf{x} + \bar{\mathbf{x}}|\mathbf{a})\, p(\mathbf{y}|\mathbf{x} + \bar{\mathbf{x}})}{p(\tilde{\mathbf{x}} + \bar{\mathbf{x}}|\tilde{\mathbf{a}})\, p(\tilde{\mathbf{y}}|\tilde{\mathbf{x}} + \bar{\mathbf{x}})} &= \frac{p(x_i + \bar{x}_i)\, p(y_i|x_i + \bar{x}_i)}{p(\tilde{x}_i + \bar{x}_i)\, p(\tilde{y}_i|\tilde{x}_i + \bar{x}_i)} \\ &= \frac{W(y_i|x_i + \bar{x}_i)}{W(\tilde{y}_i|\tilde{x}_i + \bar{x}_i)}\end{aligned}$$

Thus,

$$\frac{W_{\min}}{W_{\max}} \leq \left| \frac{p(\mathbf{y}|\mathbf{a})}{p(\tilde{\mathbf{y}})|\tilde{\mathbf{a}})} \right| \leq \frac{W_{\max}}{W_{\min}} \tag{19}$$

where $W_{\max}$ ($W_{\min}$) is the largest (smallest) non-zero channel transition probability. (Here, we have used the fact that any two sequences $\{a_1, \ldots, a_m\}$ and $\{b_1, \ldots, b_m\}$ of positive numbers, $\min\{a_1/b_1, \ldots, a_m/b_m\} \leq (\sum_{i=1}^m a_i)/(\sum_{i=1}^m b_i) \leq \max\{a_1/b_1, \ldots, a_m/b_m\}$.)

Likewise,

$$\frac{p(\mathbf{y})}{p(\tilde{\mathbf{y}})} = \frac{p(y_i)}{p(\tilde{y}_i)} = \frac{p(y_i|0) + p(y_i|1)}{p(\tilde{y}_i|0) + p(\tilde{y}_i|1)},$$

and

$$\frac{W_{\min}}{W_{\max}} \leq \left| \frac{p(\mathbf{y})}{p(\tilde{\mathbf{y}})} \right| \leq \frac{W_{\max}}{W_{\min}}. \tag{20}$$

Combining (19) and (20), Lipschitz condition (17) follows with $\Delta_i = 2\log(W_{\max}/W_{\min})$.

Now, McDiarmid's inequality [13, p. 20] states that

$$\begin{aligned}\Pr(B) &= \Pr(i(\mathbf{Y}; \mathbf{A}) \geq N(I(\mathbf{Y}; \mathbf{A}) + \lambda)) \\ &\leq \exp\left(-\frac{2(N\lambda)^2}{\sum_{i=1}^N \Delta_i^2}\right) = \exp\left(-\frac{2N\lambda^2}{\alpha}\right)\end{aligned}$$

where $\alpha = (2\log(W_{\max}/W_{\min}))^2$. This completes the proof of (8).

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] I. Tal and A. Vardy, "List decoding of polar codes," *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2213–2226, May 2015.

[3] I. Dumer and K. Shabunov. (Mar. 2017). "Recursive list decoding for Reed-Müller codes." [Online]. Available: https://arxiv.org/abs/1703.05304

[4] I. Dumer. (Mar. 2017). "On decoding algorithms for polar codes." [Online]. Available: https://arxiv.org/abs/1703.05307

[5] *Multiplexing and Channel Coding (Release 15)*, document 3GPP TS 38.212, 3GPP, Jun. 2018.

[6] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[7] B. Li, H. Shen, and D. Tse. (Aug. 2012). "An adaptive successive cancellation list decoder for polar codes with cyclic redundancy check." [Online]. Available: https://arxiv.org/abs/1208.3091

[8] K. R. Narayanan and G. L. Stuber, "List decoding of turbo codes," *IEEE Trans. Commun.*, vol. 46, no. 6, pp. 754–762, Jun. 1998.

[9] E. Arıkan, "A packing lemma for polar codes," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Hong Kong, China, Jun. 2015, pp. 2441–2445.

[10] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY, USA: Wiley, 1968.

[11] C. E. Shannon, "Certain results in coding theory for noisy channels," *Inf. Control*, vol. 1, no. 1, pp. 6–25, Sep. 1957.

[12] E. Arikan, "Source polarization," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2010, pp. 899–903.

[13] M. Raginsky and I. Sason, "Concentration of measure inequalities in information theory, communications, and coding," *Found. Trends Commun. Inf. Theory*, vol. 10, nos. 1–2, pp. 1–246, 2013.

ERDAL ARıKAN (S'84–M'79–SM'94–F'11) was born in Ankara, Turkey, in 1958. He received the B.S. degree from the California Institute of Technology, Pasadena, CA, USA, in 1981, and the M.S. and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 1982 and 1985, respectively, all in electrical engineering. Since 1987, he has been with the Electrical and Electronics Engineering Department, Bilkent University, Ankara, Turkey, where he is currently a Professor. He was a recipient of the 2010 IEEE Information Theory Society Paper Award, the 2013 IEEE W.R.G. Baker Award, the 2018 IEEE Hamming Medal, and the 2019 Claude E. Shannon Award.

AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES

PLEASE NOTE: We cannot accept new source files as corrections for your paper. If possible, please annotate the PDF proof we have sent you with your corrections and upload it via the Author Gateway. Alternatively, you may send us your corrections in list format. You may also upload revised graphics via the Author Gateway.

AQ:1 = Author: Please confirm or add details for any funding or financial support for the research of this article.