

On Polar Coding for Binary Dirty Paper

Barak Beilin and David Burshtein
 School of Electrical Engineering
 Tel-Aviv University
 Tel-Aviv 6997801, Israel
 Email: barakbei@mail.tau.ac.il, burstyn@eng.tau.ac.il

Abstract—The problem of communication over binary dirty paper (DP) using nested polar codes is considered. An improved scheme, focusing on low delay, short to moderate blocklength communication is proposed. Successive cancellation list (SCL) decoding with properly defined CRC is used for channel coding, and SCL encoding without CRC is used for source coding. The performance is compared to the best achievable rate of any coding scheme for binary DP using nested codes. A well known problem with nested polar codes for binary DP is the existence of frozen channel code bits that are not frozen in the source code. These bits need to be retransmitted in a second phase of the scheme, thus reducing transmission rate. We observe that the number of these bits is typically either zero or a small number, and provide an improved analysis, compared to that presented in the literature, on the size of this set and on its scaling with respect to the blocklength when the power constraint parameter is sufficiently large or the channel crossover probability sufficiently small.

I. INTRODUCTION

Consider the problem of transmission over a side information channel with non-causal side information, also known as the Gelfand-Pinsker (GP) problem. Applications include watermarking codes, memories with defects, write once memories and transmission over broadcast channels. In the GP problem the encoder needs to send a message M reliably over some memoryless channel $W(y | x, s)$ where $x \in \mathcal{X}$ is the input to the channel, $y \in \mathcal{Y}$ is the output, and $s \in \mathcal{S}$ is the channel state. For each transmitted symbol, x , the state s is obtained by i.i.d. sampling of some given source random variable $S \in \mathcal{S}$. The encoder observes the channel state vector, $\mathbf{s} = (s_0, s_1, \dots, s_{N-1})$, non-causally, prior to transmission. It then constructs a codeword $\mathbf{x} = (x_0, x_1, \dots, x_{N-1})$ which is a function of the message \mathbf{m} and the state vector \mathbf{s} . The decoder observes only the vector of channel outputs, $\mathbf{y} = (y_0, y_1, \dots, y_{N-1})$ and constructs the decoded codeword $\hat{\mathbf{m}}$ from \mathbf{y} .

The binary dirty paper (DP) problem depicted in Figure 1 is a side information problem with $\mathcal{X} = \mathcal{S} = \mathcal{Y} = \{0, 1\}$, $S \sim \text{Ber}(1/2)$ ($\text{Ber}(q)$ denotes a Bernoulli $(0, 1)$ random variable with probabilities $(1 - q, q)$), and the channel $W(y | x, s)$ is defined by

$$Y = X \oplus S \oplus Z$$

where \oplus denotes XOR, and $Z \sim \text{Ber}(p)$ for $p \in (0, 1/2)$. In this problem there is a “power constraint” which comes in one of two possible forms. The first form is an average power constraint: Given some $D \in (0, 1/2)$, on average a

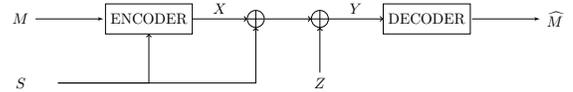


Fig. 1. The binary dirty paper problem.

fraction at most D of the channel input bits, X , are ones. That is, $Ew_H(\mathbf{X})/N \leq D$ where \mathbf{X} is the codeword and $w_H(\cdot)$ denotes Hamming weight. The second form is an individual codeword power constraint, which is stronger than the average power constraint. In this case, each codeword, \mathbf{x} , needs to satisfy $w_H(\mathbf{x})/N \leq D$. An error event in the communication happens when either the receiver does not decode the correct transmitted message, or (under the individual codeword power constraint) when the encoder cannot obtain a codeword that satisfies the required power constraint. For $D > p$ the capacity of the binary DP problem is given by

$$C_{\text{DP}} = h_2(D) - h_2(p)$$

where $h_2(\cdot)$ is the binary entropy function (base 2). Following [1], we will assume in this paper that $D > p$ since for the other case, $D < p$, the capacity as a function of D can be achieved by time sharing with the point with $D = 0$ and $R = 0$.

In [1, Section VIII] Arikan’s polar codes [2] were extended to the problem of binary DP using a polar nested codes structure. Given some pair (p, D) such that $0 < p < D < 1/2$ and blocklength N , one constructs two polar codes with blocklength N . The first is a standard binary polar code, \mathcal{C}_c , with frozen set F_c , designed for reliable communication over the binary symmetric channel, $\text{BSC}(p)$. The second code, \mathcal{C}_s , with frozen set F_s , is a binary polar code for lossy source coding (quantization) [1] of the source S with distortion D . The design of \mathcal{C}_s is very similar to the design of a polar channel code for a $\text{BSC}(D)$ channel, except for the threshold defining its frozen set. The polar coding scheme for the binary DP problem was formulated in [1] under an average power constraint but it is also valid under an individual codeword power constraint [3]. To describe the method, first assume that $F_c \subseteq F_s$ such that \mathcal{C}_c and \mathcal{C}_s are nested polar codes. As usual, denote by \mathbf{u} the message and frozen bits of a polar code, and by $\mathbf{c} = \mathbf{u}G_N$ the corresponding polar codeword where G_N is the generating matrix defined in [2, Eq. (70)]. The encoder first sets $\mathbf{u}_{F_c} = \mathbf{0}$ and $\mathbf{u}_{F_s \setminus F_c} = \mathbf{m}$ where \mathbf{m} denotes the information bits that need to be transmitted. This defines a polar source code that we denote by $\mathcal{C}_s(F_s, \mathbf{u}_{F_s}(\mathbf{m}))$.

The encoder then observes \mathbf{s} and obtains a polar codeword $\mathbf{s}' \in \mathcal{C}_s(F_s, \mathbf{u}_{F_s}(\mathbf{m}))$ that satisfies the power constraint (e.g., under the individual codeword power constraint, $w_H(\mathbf{s} \oplus \mathbf{s}') \leq D$) using a successive cancellation (SC) encoding algorithm which is a randomized version of the standard SC decoding algorithm [2] (as explained in [1], in practice the standard SC decoding algorithm can be used by the encoder without modification, but the proof requires the randomized version of the algorithm). Now the encoder transmits $\mathbf{x} = \mathbf{s} \oplus \mathbf{s}'$. The decoder receives $\mathbf{y} = \mathbf{x} \oplus \mathbf{s} \oplus \mathbf{z} = \mathbf{s}' \oplus \mathbf{z}$. Noting that $\mathbf{s}' \in \mathcal{C}_c(F_c, \mathbf{u}_{F_c} = 0)$, and that F_c was designed to obtain a good channel code for the BSC(p), the decoder can decode $\hat{\mathbf{u}}_{F_c^c}$ (F_c^c denotes the complementary of F_c) from \mathbf{y} using the SC decoding algorithm. The decoded message is obtained as $\hat{\mathbf{m}} = \hat{\mathbf{u}}_{F_s \setminus F_c}$. The rate of this communication scheme is $R = (|F_s| - |F_c|)/N$. Since for N large enough $|F_s|/N \rightarrow h_2(D)$ and $|F_c|/N \rightarrow h_2(p)$, we have $R \rightarrow C_{DP}$. That is, the scheme approaches capacity.

When the assumption $F_c \subseteq F_s$ is violated, it was suggested [1] to use a two phase transmission scheme. The first phase is identical to the one described above. In the second phase, we transmit $\mathbf{u}_{F_c \cap F_s^c}$ (or accumulate the bits of $\mathbf{u}_{F_c \cap F_s^c}$ of many first phase transmissions and send them together). In the second phase the transmitter ignores the power constraint and transmits $\mathbf{x} = \mathbf{c} \oplus \mathbf{s}$ such that the state noise is canceled. However, if $|F_c \cap F_s^c|$ is small then the damage to the power constraint will be negligible (we can compensate by decreasing the design distortion D of \mathcal{C}_s). The decoder starts by decoding $\mathbf{u}_{F_c \cap F_s^c}$ from the second phase transmission, and then it can apply the decoding described above for the case $F_c \subseteq F_s$. Another possibility is to transmit in frames using the chaining construction [4] as used in [5] for achieving Marton's region for broadcast channels. When using a sufficiently large number of frames, the rate loss of the chaining construction becomes negligible. However, the use of L frames increases latency by a factor of L , and hence is not suitable for low delay communications.

In this paper we propose an improved, low delay communication scheme over binary DP using nested polar codes [1], with CRC aided SC list (SCL) decoding [6] for channel coding, and SCL encoding without CRC for source coding. The performance is compared to the best achievable rate of any coding scheme for binary DP using nested codes. We observed that typically the set $F_c \cap F_s^c$, that needs to be retransmitted in the second phase of the scheme [1] is zero for $D - p$ larger than some small threshold. Our main theoretical contribution is an improved analysis compared to that presented in [1] on $|F_c \cap F_s^c|$ and on its scaling with respect to the blocklength when D is sufficiently large or p sufficiently small.

II. POLAR SCL CODING FOR BINARY DIRTY PAPER

We now discuss the design of an SCL coding scheme based on the SCL decoder [6] and the nested polar coding scheme of [1]. We first observe that lists are useful for improved lossy source coding since the encoder can choose the least distorted codeword from several possibilities. For example, using an

SCL encoder with $L_s = 50$ lists to encode a Ber(1/2) source at rate 0.258, with a polar code with blocklength $N = 1024$, the distortion was 0.217, compared to 0.226 when lists are not used. The theoretical minimum distortion for this rate is 0.21. Repeating the experiment with $N = 4096$ yielded a distortion of 0.215 with $L_s = 50$ lists compared to a distortion of 0.223 without lists. Although the use of CRC verification provides a significant reduction in the error rate for channel coding [6], it is not helpful for lossy source coding. This is due to the fact that in this problem we are only interested in the distortion between the source vector and any codeword. Hence a range of codewords which are sufficiently close to the source vector can be used, rather than a single preferred codeword as in the channel coding problem, making the CRC rule irrelevant. Hence, we use SCL encoder with L_s lists and without CRC, and SCL decoder with L_c lists and with CRC.

The other important design consideration relates to the proper definition of the CRC code. In [6], r CRC bits are computed from $N - |F_c| - r$ information bits. The r CRC bits are then appended to the $N - |F_c| - r$ information bits. For polar coding over side information channels it would be problematic to compute the CRC this way since this would impose a difficult constraint on the encoder side (how to output a valid codeword for channel coding that satisfies the required distortion bound). To solve this problem, we compute the r CRC bits only from the message bits that need to be transmitted to the decoder. Without using CRC there are $|F_s \cap F_c^c|$ message bits. When using CRC we reduce this number to $|F_s \cap F_c^c| - r$ and compute the r CRC bits from these message bits. We then append the r CRC bits to the $|F_s \cap F_c^c| - r$ message bits, and place them in $F_s \cap F_c^c$ (as in [1] we set zeros in F_c). At the decoder side, in the last decoding stage only those lists for which the CRC is satisfied are considered as in [6].

As in [1, Eqs. (18)-(19)], the frozen sets of the codes \mathcal{C}_s and \mathcal{C}_c are defined by

$$F_s = \left\{ i : Z_N^{(i)}(D) \geq \delta_N(D) \right\} \quad (1)$$

$$F_c = \left\{ i : Z_N^{(i)}(p) \geq \delta_N(p) \right\} \quad (2)$$

where $Z_N^{(i)}(D)$ ($Z_N^{(i)}(p)$, respectively) is the Bhattacharyya parameter of the i -th sub-channel after¹ $n = \log N$ polarization steps of a BSC(D) (BSC(p)) channel. In [1], $\delta_N(D) = 1 - \delta_N^2$ and $\delta_N(p) = \delta_N$. Setting $\delta_N = \delta/N$, yields error probability at most δ and, by [1, Lemmas 5, 6 and 7], average distortion at most $D + \sqrt{2}\delta$. Thus, to meet a required distortion constraint, D , we design F_s using a BSC(D') channel with $D' = D - \sqrt{2}\delta$. A similar statement can also be made regarding the individual codeword power constraint [3, Theorem 2]. In practice we set the thresholds $\delta_N(p)$ ($\delta_N(D)$, respectively) such that the performance of a polar code under SCL decoding (encoding) with the set F_c (F_s) yields the required error rate (distortion) performance.

¹The basis of all the logarithms in this paper is 2.

III. ANALYSIS OF $|F_c \cap F_s^c|$

As was explained above, for low delay communications with polar codes using our SCL coding variant of the method in [1], $|F_c \cap F_s^c|$ needs to be small (ideally zero).

Consider the definition of F_s and F_c in (1)-(2) and suppose that $\delta_N(D) = 1 - \delta_N^2$ and $\delta_N(p) = \delta_N$ as in [1]. Then

$$F_c \cap F_s^c = \left\{ i : Z_N^{(i)}(p) \geq \delta_N \text{ and } Z_N^{(i)}(D) < 1 - \delta_N^2 \right\} \quad (3)$$

Define

$$\tilde{F}_c = \left\{ i : Z_N^{(i)}(p) \geq 1 - \delta_N^2 \right\}$$

The analysis in [1] asserts the following

$$|F_c \cap F_s^c| \leq |F_c \setminus \tilde{F}_c| = o(N)$$

where the inequality is due to the degradedness of the sub-channel $W_N^{(i)}(D)$, corresponding to a BSC(D), with respect to $W_N^{(i)}(p)$, corresponding to a BSC(p) [1]. The equality is due to the polarization of the BSC(p) channel. A more refined argument, using scaling results of polar codes [7], [8] shows

$$|F_c \cap F_s^c| \leq |F_c \setminus \tilde{F}_c| = O(N^{1-\alpha}) \quad (4)$$

for $\alpha = (1 + 1/0.2127) = 0.175$ (this can be verified using the proofs of Theorem 1 and Theorem 2 in [8]).

However, we observed that for small to moderate values of N , in the above bound of [1] (now formulated in terms of general threshold values, $\delta_N(D)$ and $\delta_N(p)$),

$$|F_c \cap F_s^c| \leq |\hat{F}_c| \triangleq \left\{ i : \delta_N(p) \leq Z_N^{(i)}(p) < \delta_N(D) \right\}$$

$|\hat{F}_c|$ is quite large for actual practical thresholds, $\delta_N(D)$ and $\delta_N(p)$, in the definitions of F_s and F_c . Fortunately, we have observed empirically that even though $|\hat{F}_c|$ tends to be relatively large, $|F_c \cap F_s^c|$ tends to be much smaller, and it vanishes for sufficiently large D or sufficiently small p . For example, consider the case where $L_c = L_s = 8$ and frozen set thresholds designed for block error rate below 0.001 for channel crossover p , and average distortion below D . Then for $N = 1024$ and $p \in \{0.11, 0.21, 0.31\}$ we have $F_c \cap F_s^c = \emptyset$ for $D - p \geq 0.1$. For larger values of N , $|F_c \cap F_s^c|$ vanishes even starting from smaller values of $D - p$. On the other hand, $|\hat{F}_c|$ is much larger, e.g. for $p = 0.11$ and $D = 0.25, 0.45$ we have $|\hat{F}_c/N| = 0.175, 0.207$ for blocklength $N = 1024$, and $|\hat{F}_c/N| = 0.157, 0.18$ for $N = 2048$.

We now study the behavior of the set $F_c \cap F_s^c$, which represents the deviation from perfect code nestedness, and prove that for p sufficiently small or D sufficiently large, $|F_c \cap F_s^c| = O(N^\xi)$ where $\xi > 0$ can be chosen arbitrarily small, thus improving (4) significantly for the case of sufficiently small p or sufficiently large D . In fact, we prove this result for any pair of binary memoryless symmetric (BMS) channels, $W(p)$ and $W(D)$, with Bhattacharyya parameters $Z(p)$ and $Z(D)$, without requiring degradedness of $W(D)$ with respect to $W(p)$, which is important for the generalization of the results for side information channels beyond binary DP.

Consider the random processes $Z_n(p)$ and $Z_n(D)$, $n = 0, 1, \dots, \log N$. They both follow the same sequence of Arıkan's channel transformations, defined by [2, Eq. (22)] if $B_n = 0$ and by [2, Eq. (23)] if $B_n = 1$, where $\langle B_1, B_2, \dots, B_{\log N} \rangle$ defines the index of some polar sub-channel. Initially $Z_0(p) = Z(p)$ and $Z_0(D) = Z(D)$. Denote

$$\epsilon_{1,n} \triangleq Z_n(p), \quad \epsilon_{2,n} \triangleq 1 - Z_n(D) \quad (5)$$

In particular, $\epsilon_{1,0} = Z(p)$ and $\epsilon_{2,0} = 1 - Z(D)$. Now, if $B_{n+1} = 1$ then

$$\begin{aligned} (\epsilon_{1,n+1}, \epsilon_{2,n+1}) &= (\epsilon_{1,n}^2, 1 - (Z_n(D))^2) \\ &= (\epsilon_{1,n}^2, 1 - (1 - \epsilon_{2,n})^2) = (\epsilon_{1,n}^2, 2\epsilon_{2,n} - \epsilon_{2,n}^2) \end{aligned} \quad (6)$$

If $B_{n+1} = 0$ then

$$\begin{aligned} (\epsilon_{1,n+1}, \epsilon_{2,n+1}) &\leq (2\epsilon_{1,n} - \epsilon_{1,n}^2, \psi(\epsilon_{2,n})) \\ \psi(x) &\triangleq 1 - (1 - x)\sqrt{1 + 2x - x^2} \end{aligned} \quad (7)$$

where the inequality actually denotes two inequalities, one for each term. These inequalities follow from the following well known relations, e.g. [2], [7, Eq. (13)], for $B_{n+1} = 0$,

$$Z_{n+1}(p) \leq 2Z_n(p) - Z_n(p)^2 \quad (8)$$

$$Z_{n+1}(D) \geq Z_n(D)\sqrt{2 - (Z_n(D))^2} \quad (9)$$

Lemma 1. Consider the process $(\tilde{\epsilon}_{1,n}, \tilde{\epsilon}_{2,n})$ defined by $\tilde{\epsilon}_{1,0} = Z(p)$, $\tilde{\epsilon}_{2,0} = 1 - Z(D)$, and, for $n = 1, 2, \dots, \log N$,

$$(\tilde{\epsilon}_{1,n+1}, \tilde{\epsilon}_{2,n+1}) = \begin{cases} (\tilde{\epsilon}_{1,n}^2, 2\tilde{\epsilon}_{2,n} - \tilde{\epsilon}_{2,n}^2) & \text{if } B_{n+1} = 1 \\ (2\tilde{\epsilon}_{1,n} - \tilde{\epsilon}_{1,n}^2, \psi(\tilde{\epsilon}_{2,n})) & \text{if } B_{n+1} = 0 \end{cases} \quad (10)$$

For $n = \log N$ we have N possible realizations of the process corresponding to all possible sub-channels. The number of realizations for which both $\tilde{\epsilon}_{1,\log N} \geq \delta_N(p)$ and $\tilde{\epsilon}_{2,\log N} > 1 - \delta_N(D)$ is an upper bound on $|F_c \cap F_s^c|$.

The proof follows from (6)-(7) and the fact that all the functions that appear in (10), including $\psi()$, are monotonically increasing for $\tilde{\epsilon}_{i,n} \in (0, 1)$, $i = 1, 2$.

We note that the bound provided by Lemma 1 on $|F_c \cap F_s^c|$ is monotonically increasing in $Z(p)$ and monotonically decreasing in $Z(D)$. We used Lemma 1 to compute bounds on $|F_c \cap F_s^c|$ for some p and D values, and compare with the actual value of $|F_c \cap F_s^c|$. $\delta_N(p)$ and $\delta_N(D)$ were set to obtain a block error probability 0.001, and average distortion D with $L_c = 8$ and $L_s = 8$. As an example, for $N = 1024$ and $p = 0.11$ ($p = 0.21$, respectively), $|F_c \cap F_s^c|$ vanishes for $D - p \geq 0.1$ ($D - p \geq 0.1$) while the bound requires $D - p \geq 0.16$ ($D - p \geq 0.14$). For $N = 2048$ and $p = 0.11$ (same results for $p = 0.21$), $|F_c \cap F_s^c|$ vanishes for $D - p \geq 0.08$ while the bound requires $D - p \geq 0.14$.

We proceed the analysis by defining $\hat{\epsilon}_{1,n}, \hat{\epsilon}_{2,n}$ by

$$\hat{\epsilon}_{1,n} \triangleq Z_n(p), \quad \hat{\epsilon}_{2,n} \triangleq 1 - (Z_n(D))^2$$

such that $\hat{\epsilon}_{1,n} = \epsilon_{1,n}$ (see (5)). Similarly to (6) we have that if $B_{n+1} = 1$ then

$$\begin{aligned} (\hat{\epsilon}_{1,n+1}, \hat{\epsilon}_{2,n+1}) &= (\hat{\epsilon}_{1,n}^2, 1 - (Z_n(D))^4) = \\ &= (\hat{\epsilon}_{1,n}^2, 1 - (1 - \hat{\epsilon}_{2,n})^2) = (\hat{\epsilon}_{1,n}^2, 2\hat{\epsilon}_{2,n} - \hat{\epsilon}_{2,n}^2) \end{aligned}$$

Similarly to (7), if $B_{n+1} = 0$ then

$$(\hat{\epsilon}_{1,n+1}, \hat{\epsilon}_{2,n+1}) \leq (2\hat{\epsilon}_{1,n} - \hat{\epsilon}_{1,n}^2, \hat{\epsilon}_{2,n}^2)$$

The inequality for the left terms is due to (8) and the inequality for the right terms is due to (9) that can be rewritten as

$$(Z_{n+1}(D))^2 \geq (Z_n(D))^2 (2 - (Z_n(D))^2)$$

Thus we have

$$(\hat{\epsilon}_{1,n+1}, \hat{\epsilon}_{2,n+1}) \begin{cases} = (\hat{\epsilon}_{1,n}^2, 2\hat{\epsilon}_{2,n} - \hat{\epsilon}_{2,n}^2) & \text{if } B_{n+1} = 1 \\ \leq (2\hat{\epsilon}_{1,n} - \hat{\epsilon}_{1,n}^2, \hat{\epsilon}_{2,n}^2) & \text{if } B_{n+1} = 0 \end{cases} \quad (11)$$

We now claim the following key lemma.

Lemma 2.

$$\hat{\epsilon}_{1,n}\hat{\epsilon}_{2,n} < \gamma \quad \forall n$$

where $\gamma = \gamma(Z(p), Z(D))$ becomes arbitrarily small for $Z(p)$ sufficiently small or $Z(D)$ sufficiently large.

Proof. For notational convenience, denote by $\hat{\epsilon}_1 \triangleq \hat{\epsilon}_{1,n}$, $\hat{\epsilon}_2 \triangleq \hat{\epsilon}_{2,n}$, and

$$R_n \triangleq \log \frac{\hat{\epsilon}_1}{1 - \hat{\epsilon}_1} + \log \frac{\hat{\epsilon}_2}{1 - \hat{\epsilon}_2}$$

Hence,

$$R_{n+1} \begin{cases} = \log \frac{\hat{\epsilon}_1^2}{1 - \hat{\epsilon}_1^2} + \log \frac{2\hat{\epsilon}_2 - \hat{\epsilon}_2^2}{(1 - \hat{\epsilon}_2)^2} & \text{if } B_{n+1} = 1 \\ \leq \log \frac{2\hat{\epsilon}_1 - \hat{\epsilon}_1^2}{(1 - \hat{\epsilon}_1)^2} + \log \frac{\hat{\epsilon}_2^2}{1 - \hat{\epsilon}_2^2} & \text{if } B_{n+1} = 0 \end{cases} \quad (12)$$

We will first show that for all n , if $R_n \leq A$, where $A < 0$, then $R_{n+1} \leq A$. For that, it is sufficient to consider the first case in (12) ($B_{n+1} = 1$), since the same proof holds for the other case, $B_{n+1} = 0$. Now, the first case in (12) can be written as

$$R_{n+1} = R_n + \log \frac{\hat{\epsilon}_1(2 - \hat{\epsilon}_2)}{(1 + \hat{\epsilon}_1)(1 - \hat{\epsilon}_2)} \quad (13)$$

Since $R_n < A < 0$, we have

$$\frac{\hat{\epsilon}_1}{1 - \hat{\epsilon}_1} \cdot \frac{\hat{\epsilon}_2}{1 - \hat{\epsilon}_2} < 1$$

Hence, $\hat{\epsilon}_1 + \hat{\epsilon}_2 < 1$. Therefore,

$$\frac{2 - \hat{\epsilon}_2}{1 - \hat{\epsilon}_2} = 1 + \frac{1}{1 - \hat{\epsilon}_2} < 1 + \frac{1}{\hat{\epsilon}_1} = \frac{\hat{\epsilon}_1 + 1}{\hat{\epsilon}_1}$$

Using this inequality in (13) yields $R_{n+1} < R_n \leq A$ as claimed. We conclude that if $R_0 = A < 0$ then $R_n \leq R_0$ for all n , so that

$$\begin{aligned} \log \hat{\epsilon}_{1,n} + \log \hat{\epsilon}_{2,n} &< R_n \leq R_0 = \\ &= \log \frac{Z(p)}{1 - Z(p)} + \log \frac{1 - Z^2(D)}{Z^2(D)} \triangleq \log \gamma(Z(p), Z(D)) \end{aligned}$$

where the first inequality follows from the fact that $1 - \hat{\epsilon}_{i,n} < 1$, for $i = 1, 2$. Note that $\gamma(Z(p), Z(D))$ can be made

arbitrarily small by choosing $Z(p)$ sufficiently small or $Z(D)$ sufficiently large. \square

We can now state and prove our main result for the nested codes property.

Theorem 1. Consider the case where $W(p)$ and $W(D)$ are BMS channels with Bhattacharyya parameters $Z(p)$ and $Z(D)$. Given $0 < Z(p) < Z(D) < 1$, suppose either a small enough $Z(p)$ or a large enough $Z(D)$. Then, $|F_c \cap F_s^c| = O(N^\xi)$ where $\xi > 0$ can be set arbitrarily small.

Proof. By (11) we have

$$\hat{\epsilon}_{1,n+1}\hat{\epsilon}_{2,n+1} \leq \begin{cases} \hat{\epsilon}_{1,n}^2\hat{\epsilon}_{2,n}(2 - \hat{\epsilon}_{2,n}) & \text{if } B_{n+1} = 1 \\ \hat{\epsilon}_{1,n}(2 - \hat{\epsilon}_{1,n})\hat{\epsilon}_{2,n}^2 & \text{if } B_{n+1} = 0 \end{cases}$$

Hence,

$$\rho_n \triangleq \frac{\hat{\epsilon}_{1,n+1}\hat{\epsilon}_{2,n+1}}{\hat{\epsilon}_{1,n}\hat{\epsilon}_{2,n}} \leq \begin{cases} \hat{\epsilon}_{1,n}(2 - \hat{\epsilon}_{2,n}) & \text{if } B_{n+1} = 1 \\ \hat{\epsilon}_{2,n}(2 - \hat{\epsilon}_{1,n}) & \text{if } B_{n+1} = 0 \end{cases}$$

Now, since by Lemma 2 either $\hat{\epsilon}_{1,n} < \sqrt{\gamma}$ or $\hat{\epsilon}_{2,n} < \sqrt{\gamma}$, we conclude that w.p. $1/2$, $\rho_n \leq 2$ and w.p. $1/2$, $\rho_n \leq 2\sqrt{\gamma}$. That is,

$$\frac{\hat{\epsilon}_{1,n+1}\hat{\epsilon}_{2,n+1}}{\hat{\epsilon}_{1,n}\hat{\epsilon}_{2,n}} \leq \begin{cases} 2 & \text{if } \tilde{B}_{n+1} = 0 \\ 2\sqrt{\gamma} & \text{if } \tilde{B}_{n+1} = 1 \end{cases}$$

where, similarly to $\{B_i\}$, the random variables $\{\tilde{B}_i\}$ are independent, binary, uniformly distributed (i.e., $\tilde{B}_i = (0, 1)$ w.p. $(1/2, 1/2)$).

Following [2, Section IV.B], define, for $\eta \in (0, 1/2)$, the event

$$\mathcal{U}_n(\eta) \triangleq \left\{ \sum_{i=1}^n \tilde{B}_i > \left(\frac{1}{2} - \eta\right)n \right\}$$

Using the same argument as in [2, Section IV.B] we know that if the event $\mathcal{U}_n(\eta)$ holds then

$$\hat{\epsilon}_{1,n}\hat{\epsilon}_{2,n} \leq \zeta \cdot [2^{0.5+\eta}\zeta^{0.5-\eta}]^n \quad (14)$$

for $\zeta = 2\sqrt{\gamma}$. It is also known [2, Section IV.B], by Chernoff's bound, that

$$P(\mathcal{U}_n(\eta)) \geq 1 - 2^{-n[1-h_2(0.5-\eta)]}$$

Now, (14) implies that

$$\min(\hat{\epsilon}_{1,n}, \hat{\epsilon}_{2,n}) \leq \sqrt{\zeta} \cdot [2^{0.5+\eta}\zeta^{0.5-\eta}]^{n/2}$$

Setting $n = \log N$ we obtain

$$P(\min(Z_{\log N}(p), 1 - Z_{\log N}^2(D)) \leq N^{-a}) \geq 1 - N^{-1+\xi}$$

where $\xi = h_2(0.5 - \eta)$. Furthermore, $\xi > 0$ can be made arbitrarily small by setting $\eta \rightarrow 0.5^-$. Recall that if either $Z(p)$ is small enough or $Z(D)$ large enough, then γ can be made as small as desired. Hence a can be set as large as desired (any $a > 2$ is sufficient to prove the theorem). Recalling the connection between the random process $Z_n(p)$ ($Z_n(D)$, respectively) for $n = \log N$ and the values of the sub-channels, $Z_N^{(i)}(p)$ ($Z_N^{(i)}(D)$) [2], we obtain

$$\left| \left\{ i : Z_N^{(i)}(p) > N^{-a} \text{ and } Z_N^{(i)}(D) < \sqrt{1 - N^{-a}} \right\} \right| < N^\xi$$

Combining this with (3) concludes the proof (since we can take $a > 2$). \square

IV. SIMULATION RESULTS

We now present results for our polar SCL scheme to the binary DP problem. All the results presented here were achieved without the need to use retransmission, i.e., $F_s \cap F_c^s = \emptyset$ in all the reported cases.

Figure 2 presents our results for $p = 0.11$ and $D \in \{0.21, 0.31, 0.41\}$. We used $L_c = 8$ lists and CRC of size 8 in the decoder, and $L_s \in \{1, 50\}$ in the encoder. For each experiment the figure shows the maximum polar SCL rate under the constraint of error rate below $\epsilon_p = 0.001$ and average distortion below D . It can be seen how increasing the number of lists in the source encoder increases the achievable rate. The

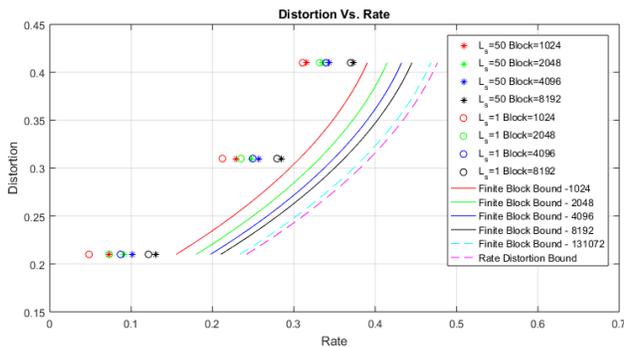


Fig. 2. Distortion vs. rate for the binary DP problem with $p = 0.11$ and $D \in \{0.21, 0.31, 0.41\}$. Simulation used $L_c = 8$ and various blocklengths and L_s values. Block error rate is at most 10^{-3} for all experiments.

figure also shows the approximated maximum achievable rate of any nested coding scheme for binary DP. It was obtained using the approximated maximum achievable channel coding rate, and minimum achievable lossy compression rate in a finite blocklength regime, in [9, Theorem 52] and [10, Eqs. (1), (11) and (93)] respectively. Since the code is nested, its approximated binary DP maximum achievable rate is obtained by subtracting these two approximations from [9] and [10]. Let N be the blocklength. Denote by ϵ_p the block error rate, and by ϵ_D the distortion constraint violation rate. Then

$$R_{GP}(N, D, p, \epsilon_p, \epsilon_D) = \bar{R}_D - \bar{R}_p$$

$$\begin{aligned} \bar{R}_D &\approx 1 - h_2(D) + \sqrt{\frac{V(D)}{N}} Q^{-1}(\epsilon_D) \\ V(D) &= D \log^2 D + (1 - D) \log^2(1 - D) - h_2^2(D) \\ \bar{R}_p &\approx 1 - h_2(p) + \frac{\log N}{2N} \\ &\quad - \sqrt{\frac{p(1-p)}{N}} \log\left(\frac{1-p}{p}\right) Q^{-1}(\epsilon_p) \end{aligned}$$

where $Q^{-1}(\cdot)$ is the inverse of the standard Gaussian complementary cumulative distribution function. The approximated bounds in Fig. 2 were obtained by setting $\epsilon_p = 0.001$ and $\epsilon_D = 0.5$, which means that we are taking the standard rate distortion bound for the lossy source coding part (since in this experiment we only set an average distortion constraint, as in

[1]). Repeating the same experiment with $\epsilon_D = 0.01$ yields an even smaller gap between the achievable rates using the SCL polar coding scheme and the bounds.

We have also compared our polar SCL coding scheme to the superposition coding scheme in [11] for long blocklength codes, using LDPC codes for channel coding and convolutional codes for source coding. In [11] the blocklength was $N=100,000$. In our experiments we used both $N = 2^{17} = 131,072$ and $N = 2^{16} = 65,536$. The results, shown in Table I, show comparable results for long blocklength codes. The SCL results were tested 100 times as in [11]. We note that the method in [11] required considerable computational resources (150–200 belief propagation iterations and 10–15 BCJR iterations with 1024 states in the decoding trellis). Results for shorter blocklengths are not reported in [11].

TABLE I
LDPC V.S. SCL BINARY DP PERFORMANCE, $p = 0.1$, $D = 0.3$

Description	Block size	DP Rate	Bit Error Rate
LDPC scheme	100K	0.36	1.28×10^{-5}
SCL, $L_c = 4$, $L_s = 1$	130K	0.353	5.9×10^{-6}
SCL, $L_c = 8$, $L_s = 8$	130K	0.357	5.6×10^{-6}
SCL, $L_c = 16$, $L_s = 50$	130K	0.362	5.0×10^{-6}
SCL, $L_c = 16$, $L_s = 50$	65K	0.356	1.2×10^{-5}
Capacity		0.42	

ACKNOWLEDGMENT

This research was supported by the Israel Science Foundation (grant no. 1868/18).

REFERENCES

- [1] S. B. Korada and R. L. Urbanke, "Polar codes are optimal for lossy source coding," *IEEE Transactions on Information Theory*, vol. 56, no. 4, pp. 1751–1768, 2010.
- [2] E. Arikan, "Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," *IEEE Transactions on Information Theory*, vol. 55, no. 7, pp. 3051–3073, 2009.
- [3] D. Burshtein and A. Strugatski, "Polar write once memory codes," *IEEE Transactions on Information Theory*, vol. 59, no. 8, pp. 5088–5101, August 2013.
- [4] S. H. Hassani and R. Urbanke, "Universal polar codes," in *IEEE International Symposium on Information Theory (ISIT)*, 2014, pp. 1451–1455.
- [5] M. Mondelli, S. H. Hassani, I. Sason, and R. L. Urbanke, "Achieving Marton's region for broadcast channels using polar codes," *IEEE Transactions on Information Theory*, vol. 61, no. 2, pp. 783–800, February 2015.
- [6] I. Tal and A. Vardy, "List decoding of polar codes," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2213–2226, 2015.
- [7] S. H. Hassani, K. Alishahi, and R. Urbanke, "Finite-length scaling for polar codes," *IEEE Transactions on Information Theory*, vol. 60, no. 10, pp. 5875–5898, 2014.
- [8] D. Goldin and D. Burshtein, "Improved bounds on the finite length scaling of polar codes," *IEEE Transactions on Information Theory*, vol. 60, no. 11, pp. 6966–6978, November 2014.
- [9] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.
- [10] V. Kostina and S. Verdú, "Fixed-length lossy compression in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 58, no. 6, pp. 3309–3338, 2012.
- [11] A. Benmatan, D. Burshtein, G. Caire, and S. Shamai, "Superposition coding for side-information channels," *IEEE Transactions on Information Theory*, vol. 52, no. 5, pp. 1872–1889, May 2006.