# Performance Bounds of Concatenated Polar Coding 

## Schemes

Dina Goldin and David Burshtein, Senior Member, IEEE


#### Abstract

A concatenated coding scheme over binary memoryless symmetric (BMS) channels using a polarization transformation followed by outer sub-codes is analyzed. Achievable error exponents and upper bounds on the error rate are derived. The first bound is obtained using outer codes which are typical linear codes from the ensemble of parity check matrices whose elements are chosen independently and uniformly. As a byproduct of this bound, it determines the required rate split of the total rate to the rates of the outer codes. A lower bound on the error exponent that holds for all BMS channels with a given capacity is then derived. Improved bounds and approximations for finite blocklength codes using channel dispersions (normal approximation), as well as converse and approximate converse results, are also obtained. The bounds are compared to actual simulation results from the literature. For the cases considered, when transmitting over the binary input additive white Gaussian noise channel, there was only a small gap between the normal approximation prediction and the actual error rate of concatenated BCH-polar codes.


## I. Introduction

Polar coding, introduced by Arikan [1], is an exciting development in coding theory. Arikan showed that, for a sufficiently large blocklength, polar codes can be used for reliable communications at rates arbitrarily close to the symmetric capacity of an arbitrary binary-input channel. Various decoding algorithms that improve Arikan's successive cancellation (SC) decoder were shown since then. A notable example is list SC decoding [2] with the possible incorporation of CRC bits. Various architectures have been considered for parallel efficient implementation of SC and list SC decoding with improved throughput, e.g. [3], [4], [5], [6]. Those architectures involve decomposing the overall polar code into an inner code and an outer code, and using SC to decode the inner code and maximum-likelihood (ML) to decode the outer code.

Using other outer codes such as powerful algebraic codes with approximated ML decoding is also possible. Interleaved concatenation of inner polar codes with outer Reed-Solomon (RS) and Bose Chaudhuri and Hocquenghem (BCH) outer codes was studied in [7] and [8], respectively, and further studied in [9], that also proposed using convolution outer codes. In this work, we study the interleaved concatenated scheme of polar codes with good outer codes [7], [8], [9]. This scheme is described in Fig. 1. The encoding is performed from right to left as follows. First, we use $2^{\lambda}$ outer codes with rates $R_{i}$, $i=0, \ldots, 2^{\lambda}-1$, to encode the information bits, creating $2^{\lambda}$ codewords, each of length $N_{1}$. The resulting codewords are interleaved and processed by $N_{1}$ polar encoders of length $2^{\lambda}$ as shown in Fig. 1 We obtain a code with blocklength $N=N_{1} \cdot 2^{\lambda}$ and rate $R=\sum_{i=0}^{2^{\lambda}-1} R_{i} / 2^{\lambda}$.
The decoding can also be described using Fig. 1] However it is performed from left to right. As described in [9], the first information bit of each of the $N_{1}$ polar codes is decoded in parallel, using a soft-decision algorithm that produces log-likelihood ratio (LLR) values. These LLRs are used as the input of the decoder of the code $A_{0}$, and the decisions of that decoder are

[^0]

Fig. 1. Concatenation of $N_{1}$ polar codes of length $2^{\lambda}$ with $2^{\lambda}$ outer codes of length $N_{1}$
passed back to the polar decoders, and used in calculating the LLR of the second information bit of each polar code. These LLRs are then used as the input of the decoder of $A_{1}$, etc. In general, when the LLRs of the $i$-th information bits of the $N_{1}$ polar codes are calculated, the previous $i-1$ information bits of these codes are available from decoding $A_{0}, A_{1}, \ldots, A_{i-2}$. We described this coding scheme as a combination of block SC decoding and optimal (ML), or non-optimal decoding of sub-codes with blocklength $N_{1}$ in [10]. If the outer codes are polar as well, this is the scheme proposed in [4] for efficient parallel decoding of polar codes. If the outer codes are Reed-Solomon, BCH or convolutional codes, this is the scheme studied in [7], [8], [9]. Note that unlike standard polar codes, in the concatenated polar coding scheme there are no frozen bits, unless $R_{i}=0$, when the whole block $A_{i}$ is frozen. We also note that a close to ML, computationally efficient decoding of short outer codes can be realized using various algorithms such as the ordered statistics decoder (OSD) [11], the box-and-match decoder [12], used in [8], or the recent machine learning-based schemes presented in [13], [14], [15], [16] and references therein, that may be efficient for small blocklength codes (especially when using hardware implementations). The decoder of the concatenated coding scheme may possess an improved throughput compared to list SC decoding of polar codes, provided the outer codes can be decoded efficiently.

In this work we analyze the performance of the concatenated polar coding scheme. Our main interest is in the case where $\lambda$ is small (e.g., $\lambda=1,2,3$ ), and the blocklength of each outer code, $N_{1}$, is also relatively small (e.g., of order 100) such that the OSD algorithm or the other methods mentioned above, can be used to decode the outer codes with a reasonable computational effort. As a motivating example, suppose we need to design an error correcting code with blocklength which is about $N=256$, and rate which is about $1 / 2$. As will be described in detail later in the paper, in order to closely approach the error rate of the best code under these conditions, as predicted by the normal approximation in [17], one may apply the OSD algorithm to a BCH code with blocklength 255 and rate $R=131 / 255 \approx 1 / 2$. However, the resulting computational complexity would be prohibitive for actual, real time, low cost and low power applications. Alternatively, as will be shown in Section V, one may construct a BCH-polar concatenated scheme with $\lambda=1$ and $N_{1}=128$. The total blocklength is $N=256$, and the total rate is $R=1 / 2$. The first outer code is a BCH code with rate $R_{1}=36 / 128$. The second outer code is a BCH code with rate $R_{2}=92 / 128$ (these rates were determined from our analysis in Section IV. In order to decode the scheme,
one needs to apply the OSD algorithm twice, first to the lower rate code, and then to the higher rate code. As Fig. 7 shows, the computational complexity of the decoder of the first scheme ( BCH with blocklength 255 ) is larger by about three orders of magnitude compared to the BCH-polar concatenated scheme. On the other hand, as Fig. 6 shows, the resulting performance degradation, when using the BCH-polar concatenated code, for frame error rate of $10^{-3}$, is only about 0.6 dB in the signal to noise ratio. Fig. 6 also shows that our prediction to the performance of the BCH -polar concatenated scheme is very accurate.

In Section II we provide some brief background on polar codes, and fix the notations. In Section III we obtain achievable error exponents and upper bounds on the error rate for the concatenated polar coding scheme. Our first lower bound on the error exponent (upper bound on the error rate) can be achieved using outer codes which are typical linear codes from the ensemble of parity check matrices whose elements are chosen independently and uniformly (i.e., they are set to $\{0,1\}$ with probabilities $\{1 / 2,1 / 2\}$ ). As a byproduct of this bound, we determine the required rate split of the total rate to the rates of the outer codes. We then obtain a lower bound on the error exponent that holds for all binary memoryless symmetric (BMS) channels with capacity $I$. In Section IV we derive improved bounds and approximations using channel dispersions (normal approximation) for finite blocklength codes. We also derive converse and approximated converse results. In Section $V$ we compare our bounds to actual simulation results. For the cases considered, when transmitting over the binary input additive white Gaussian noise channel (BIAWGNC), there was only a small gap between the normal approximation prediction and the actual error rate of concatenated BCH-polar codes. Section VI concludes the paper.

## II. Background on Polar Codes

Consider a BMS $W: \mathcal{X} \rightarrow \mathcal{Y}$ with input alphabet $\mathcal{X}=\{0,1\}$ and output alphabe ${ }^{11} \mathcal{Y}$. The capacity of the channel, $I(W)$, i. ${ }^{2}$

$$
I(W) \triangleq \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} 0.5 W(y \mid x) \log \frac{W(y \mid x)}{\sum_{x^{\prime} \in \mathcal{X}} 0.5 W\left(y \mid x^{\prime}\right)}
$$

Channel polarization [1] is based on mapping two identical copies of the channel $W$ into the pair of BMS channels $W^{-}$: $\mathcal{X} \rightarrow \mathcal{Y}^{2}$ and $W^{+}: \mathcal{X} \rightarrow \mathcal{X} \times \mathcal{Y}^{2}$ defined as

$$
\begin{aligned}
W^{-}\left(y_{1}, y_{2} \mid u_{1}\right) & =\sum_{u_{2} \in \mathcal{X}} 0.5 W\left(y_{1} \mid u_{1} \oplus u_{2}\right) W\left(y_{2} \mid u_{1}\right) \\
W^{+}\left(y_{1}, y_{2}, u_{1} \mid u_{2}\right) & =0.5 W\left(y_{1} \mid u_{1} \oplus u_{2}\right) W\left(y_{2} \mid u_{2}\right)
\end{aligned}
$$

Recalling that the channels $W^{+}$and $W^{-}$can be defined using density evolution operators, $W^{-}=W$ 㘢 $W$ and $W^{+}=W \circledast W$ [18], [19], and applying [20, Theorem 4.141], yields $I^{2}(W) \leq I\left(W^{-}\right) \leq 1-h_{2}\left\{2 h_{2}^{-1}[1-I(W)] \cdot\left[1-h_{2}^{-1}[1-I(W)]\right]\right\}$ with $h_{2}(x) \triangleq-x \log x-(1-x) \log (1-x)$, and $h_{2}^{-1}$ is the inverse of $h_{2}$ with values in $[0,0.5]$. In addition, $I\left(W^{+}\right)+I\left(W^{-}\right)=$ $2 I(W)$ [20, Lemma 4.41]. This means $I\left(W^{+}\right)-I(W)=I(W)-I\left(W^{-}\right) \in\left[\epsilon_{l}(I(W)), \epsilon_{h}(I(W))\right]$, where $\epsilon_{h}(x)=x-x^{2}$ and $\epsilon_{l}(x)=x-1+h_{2}\left\{h_{2}^{-1}(1-x)\left[2-2 h_{2}^{-1}(1-x)\right]\right\}$.

This procedure can now be reapplied to $W^{-}$and $W^{+}$, creating $W^{--}, W^{-+}, W^{+-}$and $W^{++}$. Repeating the procedure $\lambda$ times we obtain $2^{\lambda}$ BMS sub-channels, whose average capacity is $I(W)$. It was shown [1] that these channels are polarized, i.e. for all $\delta \in(0,1)$

$$
\lim _{\lambda \rightarrow \infty}\left|\left\{\mathbf{s} \in\{+,-\}^{n}: I\left(W^{\mathbf{s}}\right) \in(0, \delta)\right\}\right| / 2^{\lambda}=1-I(W)
$$

[^1]$$
\lim _{\lambda \rightarrow \infty}\left|\left\{\mathbf{s} \in\{+,-\}^{n}: I\left(W^{\mathbf{s}}\right) \in(1-\delta, 1)\right\}\right| / 2^{\lambda}=I(W)
$$

Those $2^{\lambda}$ sub channels, denoted $W_{0}, W_{1}, \ldots, W_{2^{\lambda}-1}$, are the channels that the outer codes in the scheme described in Fig. 1 see. Each outer code $A_{i}$ sees the channel $W_{i}$, and is designed to operate over this channel.

## III. Bounds on the error exponent

Consider the concatenated, polarization based code that was described above. The blocklength of the code is $N$ and it uses outer sub-codes with blocklength $N_{1}$. The number of polarization steps is $\lambda=\log \left(N / N_{1}\right)$. We derive a lower bound on the error exponent (upper bound on the error rate) of the scheme by choosing the elements of the generating matrices of the linear sub-codes, $A_{i}$, uniformly at random, as described in [21, Section 6.2]. We can obtain an upper bound on the average error probability of the ensemble of codes that we have just defined, $P_{e}$, using the successive decoding method outlined in Section I. when transmitting over a given BMS channel, $W(y \mid x)$, as follows. We first compute the distributions (given that the zero codeword was transmitted) of the LLRs of the sub-channels after $\lambda$ polarization steps, using density evolution (DE) for polar codes as in [18], [19]. Denote these distributions by $a_{W_{i}}(x), i=0,1, \ldots, 2^{\lambda}-1$. By [21, Sections 5.6 and 6.2], when using successive cancellation to decode the polar concatenated scheme described in Fig. 1,

$$
P_{e} \leq \sum_{i=0}^{2^{\lambda}-1} e^{-N_{1} E_{r}\left(W_{i}, R_{i}\right)}
$$

where $W_{i}$ is the $i$ th sub-channel, $R_{i}$ is the rate of the corresponding outer sub-code, and $E_{r}(W, R)$ is the error exponent, given by ${ }^{3}$

$$
E_{r}(W, R)= \begin{cases}\max _{0 \leq \rho \leq 1} E_{0}(W, \rho)-\rho R \cdot \ln 2 & R \neq 0  \tag{1}\\ \infty & R=0\end{cases}
$$

where

$$
\begin{gathered}
E_{0}(W, \rho)=-\ln \sum_{y}\left(0.5\left[W(y \mid 0)^{\frac{1}{1+\rho}}+W(y \mid 1)^{\frac{1}{1+\rho}}\right]\right)^{1+\rho} \\
=(1+\rho) \ln 2-\ln \sum_{l}\left[a_{W}(l)^{\frac{1}{1+\rho}}+a_{W}(-l)^{\frac{1}{1+\rho}}\right]^{1+\rho}
\end{gathered}
$$

The second equality follows by modifying the original channel as in [20, Lemma 4.35]: We add a processing block that computes the LLR from the original channel output. The new channel is also a BMS, and is operationally equivalent to the original channel.

According to [21, Chapter 5.7], for low rates the average error probability is different from the typical error probability, since poor codes in the ensemble, although quite improbable, have a very high error probability. Using the expurgated error exponent provides a tighter estimate of the error probability of good codes than the random-coding exponent. This improved bound is [21, Theorem 5.7.1]. It asserts that the average error probability of the ensemble of typical codes with rate $R$ is upper bounded by $\exp \left[-N E_{e x}\left(W, R+\frac{2}{N}\right)\right]$, where

$$
E_{e x}(W, R)=\sup _{\rho \geq 1} E_{x}(W, \rho)-\rho R \cdot \ln 2
$$

[^2]and
\[

$$
\begin{aligned}
E_{x}(W, \rho) & =-\rho \ln \left[\frac{1+\left(\sum_{y} \sqrt{W(y \mid 0) W(y \mid 1)}\right)^{\frac{1}{\rho}}}{2}\right] \\
& =-\rho \ln \left[1+\left(\sum_{l} \sqrt{a_{W}(l) a_{W}(-l)}\right)^{\frac{1}{\rho}}\right]+\rho \ln 2
\end{aligned}
$$
\]

Therefore, in our polar concatenated scheme, we obtain the following upper bound on $P_{e}$, the ensemble average error probability when using only typical outer codes,

$$
\begin{equation*}
P_{e} \leq \sum_{i=0}^{2^{\lambda}-1} e^{-N_{1} E\left(W_{i}, R_{i}, N_{1}\right)} \tag{2}
\end{equation*}
$$

where

$$
E\left(W, R, N_{1}\right)=\max \left[E_{r}(W, R), E_{e x}\left(W, R+2 / N_{1}\right)\right]
$$

This construction is called expurgated random code. Define also

$$
\begin{equation*}
E(W, R) \triangleq \lim _{N_{1} \rightarrow \infty} E\left(W, R, N_{1}\right)=\max \left[E_{r}(W, R), E_{e x}(W, R)\right] \tag{3}
\end{equation*}
$$

For the binary symmetric channel (BSC), the true error exponent of a typical linear code with rate $R$, transmitted over $W$, is $E(W, R)[22]$. It is conjectured to be the true exponent for other BMS channels as well. By [2], the error exponent, $E_{\lambda}(W, R)$, of the polarization-based code of blocklength $N$ with $\lambda$ polarization steps followed by $2^{\lambda}$ typical random linear codes of blocklength $N_{1}=N / 2^{\lambda}$, with the best rate split, is lower bounded by, and conjectured to be equal to,

$$
\frac{1}{2^{\lambda}} \max _{R_{0}, \ldots, R_{2^{\lambda}-1}} \min _{i, R_{i} \neq 0} E\left(W_{i}, R_{i}\right)
$$

where the maximization is over all possible combinations of rates $R_{0}, \ldots, R_{2^{\lambda}-1}$ with total code rate $R$. In [10] we calculated this error exponent by searching for those values of $R_{i}$ for which $E\left(W_{i}, R_{i}\right)$ are equal. We now present an improved approach that yields an explicit recursive expression for $E_{\lambda}(W, R)$ and produces the maximizing rates as a byproduct. Denote the minimal value of the right hand side (RHS) of 2 by $\exp \left\{-N E_{\lambda}\left(W, R, N_{1}\right)\right\}$ such that (s.t.)

$$
\begin{equation*}
E_{\lambda}\left(W, R, N_{1}\right)=\frac{1}{N_{1} \cdot 2^{\lambda}} \max _{R_{0}, \ldots, R_{2} \lambda-1}-\ln \sum_{i=0}^{2^{\lambda}-1} e^{-N_{1} E\left(W_{i}, R_{i}, N_{1}\right)} \tag{4}
\end{equation*}
$$

where the maximization is over all possible combinations of rates, $R_{0}, \ldots, R_{2^{\lambda}-1}$, s.t. $\sum_{i=0}^{2^{\lambda}-1} R_{i}=2^{\lambda} \cdot R$, and for all $i$, $R_{i} \cdot N_{1}$ is an integer.

Lemma 1: Define

$$
\mathcal{A} \triangleq\left\{x \mid x \in[\max (0,2 R-1), \min (1,2 R)], x \cdot N_{1} 2^{\lambda-1} \in \mathbb{Z}\right\}
$$

For any positive integers, $\lambda, N_{1}$ and $R N_{1} 2^{\lambda}$,

$$
\begin{equation*}
E_{\lambda}\left(W, R, N_{1}\right)=\max _{R_{1} \in \mathcal{A}}-\ln \left(e^{-N_{1} 2^{\lambda-1} E_{\lambda-1}\left(W^{-}, R_{1}, N_{1}\right)}+e^{-N_{1} 2^{\lambda-1} E_{\lambda-1}\left(W^{+}, 2 R-R_{1}, N_{1}\right)}\right) /\left(N_{1} 2^{\lambda}\right) \tag{5}
\end{equation*}
$$

where $E_{0}\left(W, R, N_{1}\right)=E\left(W, R, N_{1}\right)$.
Proof: For $\lambda=1$ the claim follows immediately from (4) for $\lambda=1$. The condition $R_{1} \in[\max (0,2 R-1), \min (1,2 R)]$ follows from $R_{1}, 2 R-R_{1} \in[0,1]$.

For $\lambda>1$, note that $W_{i}$ are sub-channels of $W^{-}$for $i=0, \ldots, 2^{\lambda-1}-1$, and sub-channels of $W^{+}$for $i=2^{\lambda-1}, \ldots, 2^{\lambda}-1$. (4) can be rewritten as

$$
\begin{aligned}
& E_{\lambda}\left(W, R, N_{1}\right)=\frac{1}{N_{1} 2^{\lambda}} \max _{R^{\prime} \in \mathcal{A}} \quad \max _{R_{0} \cdots R_{2 \lambda}-1}-\ln \sum_{i=0}^{2^{\lambda}-1} e^{-N_{1} E\left(W_{i}, R_{i}, N_{1}\right)} \\
& \begin{array}{c}
\sum_{i=0}^{2^{\lambda-1}-1} R_{i}=R^{\prime} \cdot 2^{\lambda-1} \\
\sum_{i=2^{\lambda}-1}^{\lambda^{\lambda}-1} R_{i}=\left(2 R-R^{\prime}\right) \cdot 2^{\lambda-1}
\end{array} \\
& =\frac{1}{N_{1} 2^{\lambda}} \max _{R^{\prime} \in \mathcal{A}}-\ln \left(\min _{\substack{ \\
R_{0}, \ldots, R_{2} \lambda-1 \\
\sum R_{i}=R^{\prime} \cdot 2^{\lambda-1}}} \sum_{i=0}^{2^{\lambda-1}-1} e^{-N_{1} E\left(W_{i}, R_{i}, N_{1}\right)}+\sum_{\substack{R_{2 \lambda-1}, \ldots, R_{2 \lambda}-1 \\
\sum R_{i}=\left(2 R-R^{\prime}\right) \cdot 2^{\lambda-1}}} \sum_{i=2^{\lambda-1}}^{2^{\lambda}-1} e^{-N_{1} E\left(W_{i}, R_{i}, N_{1}\right)}\right) \\
& =\frac{1}{N_{1} 2^{\lambda}} \max _{R^{\prime} \in \mathcal{A}}-\ln \left(e^{-N_{1} \cdot 2^{\lambda-1} E_{\lambda-1}\left(W^{-}, R^{\prime}, N_{1}\right)}+e^{-N_{1} \cdot 2^{\lambda-1} E_{\lambda-1}\left(W^{+}, 2 R-R^{\prime}, N_{1}\right)}\right)
\end{aligned}
$$

where the first equality follows from rewriting (4), the second one follows from splitting the inner maximization into two separate ones and inserting them into the monotonic decreasing function $-\ln ()$, and the third equality follows from applying (4) for $\lambda-1$ instead of $\lambda$.

We conjecture that the condition $R_{1} \in[\max (0,2 R-1), \min (1,2 R)]$ in 5 can be replaced by $R_{1} \in[\max (0,2 R-1), R]$. This is due to the fact that, compared to $W^{-}, W^{+}$is a better channel. Hence the information rate of the sub-code corresponding to $W^{+}$should be larger than the rate of the sub-code corresponding to $W^{-}$.

Denote

$$
\mathcal{B} \triangleq[\max (0,2 R-1), \min (1,2 R)]
$$

Lemma 2: Define $E_{0}(W, R)=E(W, R)$ (defined in (3)), and for $\lambda \geq 1$ define recursively,

$$
\begin{equation*}
E_{\lambda}(W, R)=0.5 \max _{R_{1} \in \mathcal{B}} E_{m, \lambda}\left(W, R, R_{1}\right) \tag{6}
\end{equation*}
$$

where

$$
\begin{equation*}
E_{m, \lambda}\left(W, R, R_{1}\right) \triangleq \min \left[E_{\lambda-1}\left(W^{-}, R_{1}\right), E_{\lambda-1}\left(W^{+}, 2 R-R_{1}\right)\right] \tag{7}
\end{equation*}
$$

Then for $\lambda \geq 0$,

$$
E_{\lambda}(W, R)-\Theta\left(\frac{1}{N_{1}}\right) \leq E_{\lambda}\left(W, R, N_{1}\right) \leq E_{\lambda}(W, R)
$$

The bound shows that for large $N_{1}, E_{\lambda}\left(W, R, N_{1}\right)$ approaches $E_{\lambda}(W, R)$.
Proof: We prove by induction that $E_{\lambda}\left(W, R, N_{1}\right) \leq E_{\lambda}(W, R)$, with $E_{\lambda}(W, R)$ defined recursively in (6). This claim is true for $\lambda=0$ since

$$
\begin{aligned}
E_{0}\left(W, R, N_{1}\right) & =E\left(W, R, N_{1}\right)=\max \left[E_{r}(W, R), E_{e x}\left(W, R+2 / N_{1}\right)\right] \\
& \leq \max \left[E_{r}(W, R), E_{e x}(W, R)\right]=E(W, R)=E_{0}(W, R)
\end{aligned}
$$

where the inequality is due to the fact that $E_{e x}(W, R)$ is a decreasing function of $R$. For shortness of notation, define

$$
E_{m, \lambda}\left(W, R, R_{1}, N_{1}\right) \triangleq \min \left[E_{\lambda-1}\left(W^{-}, R_{1}, N_{1}\right), E_{\lambda-1}\left(W^{+}, 2 R-R_{1}, N_{1}\right)\right]
$$

Assuming the claim is true for $\lambda-1$, we have

$$
E_{\lambda}\left(W, R, N_{1}\right) \leq \frac{1}{N_{1} 2^{\lambda}} \max _{R_{1} \in \mathcal{A}}-\ln \left(e^{\left.-N_{1} \cdot 2^{\lambda-1} E_{m, \lambda}\left(W, R, R_{1}, N_{1}\right)\right]}\right)
$$

$$
\begin{aligned}
& =0.5 \max _{R_{1} \in \mathcal{A}} E_{m, \lambda}\left(W, R, R_{1}, N_{1}\right) \leq 0.5 \max _{R_{1} \in \mathcal{A}} E_{m, \lambda}\left(W, R, R_{1}\right) \\
& \leq 0.5 \max _{R_{1} \in \mathcal{B}} E_{m, \lambda}\left(W, R, R_{1}\right)=E_{\lambda}(W, R)
\end{aligned}
$$

where the first inequality follows from (5), and the induction assumption yields the second inequality.
Similarly, we can prove that $E_{\lambda}(W, R)-\Theta\left(\frac{1}{N_{1}}\right) \leq E_{\lambda}\left(W, R, N_{1}\right)$ using induction: The claim is true for $\lambda=0$ since

$$
E_{e x}\left(W, R+2 / N_{1}\right) \geq E_{e x}(W, R)-\frac{2}{N_{1}}\left|\frac{\partial E_{e x}(W, r)}{\partial r}\right|_{r=R}=E_{e x}(W, R)-\Theta\left(\frac{1}{N_{1}}\right)
$$

where the inequality follows from the fact that $E_{e x}(W, R)$ is a convex and decreasing function of $R$. This yields

$$
\begin{aligned}
E_{0}\left(W, R, N_{1}\right) & =\max \left[E_{r}(W, R), E_{\text {ex }}\left(W, R+2 / N_{1}\right)\right] \\
& \geq \max \left[E_{r}(W, R), E_{\text {ex }}(W, R)\right]-\Theta\left(\frac{1}{N_{1}}\right)=E_{0}(W, R)-\Theta\left(\frac{1}{N_{1}}\right)
\end{aligned}
$$

Assuming that the claim is true for $\lambda-1$, we have

$$
\begin{align*}
E_{\lambda} & \left(W, R, N_{1}\right) \\
& \geq \frac{1}{N_{1} 2^{\lambda}} \max _{R_{1} \in \mathcal{A}}-\ln \left(2 e^{-N_{1} \cdot 2^{\lambda-1} E_{m, \lambda}\left(W, R, R_{1}, N_{1}\right)}\right) \\
& =0.5 \max _{R_{1} \in \mathcal{A}} E_{m, \lambda}\left(W, R, R_{1}, N_{1}\right)-\ln 2 /\left(N_{1} 2^{\lambda}\right) \\
& \geq 0.5 \max _{R_{1} \in \mathcal{A}} E_{m, \lambda}\left(W, R, R_{1}\right)-\Theta\left(1 / N_{1}\right) \tag{8}
\end{align*}
$$

where the first inequality follows from (5), and the second follows from the induction assumption.
Define

$$
\hat{R_{1}} \triangleq \underset{R_{1} \in \mathcal{B}}{\operatorname{argmax}} E_{m, \lambda}\left(W, R, R_{1}\right)
$$

Since $E_{\lambda-1}\left(W^{-}, R_{1}\right)$ and $E_{\lambda-1}\left(W^{+}, 2 R-R_{1}\right)$ are convex functions of $R_{1}$ (see Lemma 3 below, Appendix A and Fig. 10,

$$
\begin{aligned}
E_{m, \lambda}\left(W, R, R_{1}\right) & =\min \left[E_{\lambda-1}\left(W^{-}, R_{1}\right), E_{\lambda-1}\left(W^{+}, 2 R-R_{1}\right)\right] \\
& \geq\left\{\begin{array}{c}
E_{\lambda-1}\left(W^{-}, \hat{R_{1}}\right)-\left|\frac{\partial E_{\lambda-1}\left(W^{-}, r\right)}{\partial r}\right|_{r=\hat{R}_{1}}\left(R_{1}-\hat{R}_{1}\right) \\
E_{\lambda-1}\left(W^{+}, 2 R-\hat{R_{1}}\right)+\left|\frac{\partial E_{\lambda-1}\left(W^{+}, 2 R-r\right)}{\partial r}\right|_{r=\hat{R_{1}}}\left(R_{1}-\hat{R_{1}}\right) \\
R_{1}>R_{1}<\hat{R_{1}}
\end{array}\right. \\
& =2 E_{\lambda}(W, R)-\left\{\begin{aligned}
\left|\frac{\partial E_{\lambda-1}\left(W^{-}, r\right)}{\partial r}\right|_{r=\hat{R_{1}}}\left(R_{1}-\hat{R_{1}}\right) & R_{1}>\hat{R}_{1} \\
\left|\frac{\partial E_{\lambda-1}\left(W^{+}, 2 R-r\right)}{\partial r}\right|_{r=\hat{R_{1}}}\left(\hat{R}_{1}-R_{1}\right) & R_{1}<\hat{R}_{1}
\end{aligned}\right.
\end{aligned}
$$

Maximizing this expression over $R_{1} \in[\max (0,2 R-1), \min (1,2 R)]$ s.t. $R_{1} \cdot N_{1} 2^{\lambda-1} \in \mathbb{Z}$ yields a result for $R_{1}-\hat{R_{1}}=$ $\Theta\left(\frac{1}{N_{1} 2^{\lambda-1}}\right)=\Theta\left(\frac{1}{N_{1}}\right)$. Therefore, the result of the maximization is $2 E_{\lambda}(W, R)-\Theta\left(\frac{1}{N_{1}}\right)$. Combining this with (8) yields $E_{\lambda}\left(W, R, N_{1}\right) \geq E_{\lambda}(W, R)-\Theta\left(\frac{1}{N_{1}}\right)$.

Lemma 3: For any integer $\lambda \geq 0$, and any $\operatorname{BMS} W, E_{\lambda}(W, R)$ is a finite, decreasing and convex function of $R$ for $R>0$. Furthermore, for $R>0,\left|\partial E_{\lambda}(W, R) / \partial R\right|<\infty$.
We prove the lemma in Appendix $A$
In Fig. 2 we plot the resulting error exponent $E_{\lambda}(W, R)$ for a code with rate $R=1 / 2$ and $\lambda=1,2,3$ as a function of the SNR when transmitting over a BIAWGNC. This is compared to a naive approach where we simply use $2^{\lambda}$ codewords of a code with blocklength $N_{1}=N / 2^{\lambda}$ (without using the polar transformation at all). Denote the error exponent of the polar-based (naive, respectively) approach by $E_{\lambda}$ and $E_{n, \lambda}$. We also plot the error exponent of an optimal code, $E$, corresponding to $\lambda=0$.


Fig. 2. Error exponents of codes created by $\lambda$ polar transforms followed by $2^{\lambda}$ codewords of outer codes with blocklength $N / 2^{\lambda}$ and optimal rate-splitting, and naively using $2^{\lambda}$ codewords of a code with blocklength $N / 2^{\lambda}$. The channel is BIAWGNC, and the total rate is $1 / 2$.

Note that $E_{n, \lambda}=E / 2^{\lambda}$. As expected, polarization is useful, i.e. $E_{\lambda} \geq E_{n, \lambda}$. All the plots in Fig. 2 have an asymptote at the SNR corresponding to the capacity $C=R=1 / 2$, which equals 0.19 dB , since for this $\mathrm{SNR}, E, E_{\lambda}$ and $E_{n, \lambda}$ approach zero. This explains why in all the discussed codes, the gap to the optimal error exponent plot is small for low SNR.

Although as the channel capacity increases (as in the case of high SNR), polarization is less effective (because for high capacity, the polarization gap, $I\left(W^{+}\right)-I(W)=I(W)-I\left(W^{-}\right)$, approaches zero, and therefore polarization has almost no affect on the channel), $E_{n, \lambda} / E_{\lambda} \nrightarrow 1$ for high SNR. This is due to the fact, that for expurgated ensembles $\lim _{S N R \rightarrow \infty} E(W, R)=\infty$, so $E_{n, \lambda}$ and $E_{\lambda}$ approach infinity for high SNR. It should be noted, that for the random coding ensemble $\lim _{S N R \rightarrow \infty} E_{r}(W, R)=\ln 2(1-R)$, so if we have not considered expurgated ensembles (i.e., ensembles with typical outer codes only), we would have obtained $E_{n, \lambda} / E_{\lambda} \rightarrow 1$ for high SNR. It should also be noted that the error exponent improvement increases with $\lambda$, i.e, for a desired error exponent, the SNR improvement by applying $\lambda$ polarization steps, compared to the naive approach, increases with $\lambda$. For example, for desired error exponent of 0.01 , the SNR gain for $\lambda=1,2,3$ is $0.2 \mathrm{~dB}, 0.5 \mathrm{~dB}$ and 0.95 dB respectively.

Fig. 2 demonstrates that polarization improves the error exponent compared to the naive approach. In the following we provide some partial theoretical justification. Define $E_{r, \lambda}(W, R)$ the same way as $E_{\lambda}(W, R)$ in Lemma 2, except that $E_{r, 0}(W, R)=E_{r}(W, R)$. That is, we only consider a random coding (non-expurgated) error exponent analysis. Similarly, define $E_{n, r, \lambda}(W, R) \triangleq E_{r}(W, R) / 2^{\lambda}$. Using the results in [23], we claim that under certain conditions, $E_{r, \lambda}(W, R) \geq E_{n, r, \lambda}(W, R)$. We demonstrate this for the case $\lambda=1$, although the argument may be extended to larger values of $\lambda$. The combination of

$$
2 E_{0}(W, \rho) \leq E_{0}\left(W^{+}, \rho\right)+E_{0}\left(W^{-}, \rho\right)
$$

[23] and (1) yields that for $R \neq 0$ and all $R_{1}$ :

$$
\begin{align*}
2 E_{r}(W, R) & =\max _{\rho \in[0,1]} 2\left[E_{0}(W, \rho)-\rho R \ln 2\right] \leq \max _{\rho \in[0,1]}\left[E_{0}\left(W^{+}, \rho\right)-\rho R_{1} \ln 2+E_{0}\left(W^{-}, \rho\right)-\rho\left(2 R-R_{1}\right) \ln 2\right]  \tag{9}\\
& \leq \max _{\rho \in[0,1]}\left[E_{0}\left(W^{+}, \rho\right)-\rho R_{1} \ln 2\right]+\max _{\rho \in[0,1]}\left[E_{0}\left(W^{-}, \rho\right)-\rho\left(2 R-R_{1}\right) \ln 2\right] \\
& =E_{r}\left(W^{-}, R_{1}\right)+E_{r}\left(W^{+}, 2 R-R_{1}\right)
\end{align*}
$$

This is true in particular for $R_{1}=\hat{R}_{1}$, defined by

$$
\hat{R}_{1} \triangleq \underset{R_{1} \in \mathcal{B}}{\operatorname{argmax}} \min \left[E_{r}\left(W^{-}, R_{1}\right), E_{r}\left(W^{+}, 2 R-R_{1}\right)\right]
$$

Now, $E_{r}\left(W^{-}, R_{1}\right)$ is decreasing in $R_{1}$, and $E_{r}\left(W^{+}, 2 R-R_{1}\right)$ is increasing in $R_{1}$. If the following condition (which parallels (25) in Appendix A)

$$
\left\{\begin{array}{l}
E_{r}\left(W^{-}, 0^{+}\right) \geq E_{r}\left(W^{+}, 2 R\right)  \tag{10}\\
E_{r}\left(W^{-}, 2 R\right) \leq E_{r}\left(W^{+}, 0^{+}\right)
\end{array}\right.
$$

holds (where for any $W$ and $R>1$ we define $E_{r}(W, R) \equiv 0$ ), then $E_{r}\left(W^{-}, R_{1}\right)$ and $E_{r}\left(W^{+}, 2 R-R_{1}\right)$ intersect at $\hat{R}_{1}$ (see Fig. 10 in Appendix A with $E_{r}\left(W^{-}, r\right)$ replaced by $E_{\lambda-1}\left(W^{-}, r\right)$ and $E_{r}\left(W^{+}, 2 R-r\right)$ replaced by $\left.E_{\lambda-1}\left(W^{+}, 2 R-r\right)\right)$. In this case

$$
E_{r}\left(W^{-}, \hat{R}_{1}\right)+E_{r}\left(W^{+}, 2 R-\hat{R}_{1}\right)=2 \max _{R_{1}} \min \left[E_{r}\left(W^{-}, R_{1}\right), E_{r}\left(W^{+}, 2 R-R_{1}\right)\right]=4 E_{r, 1}(W, R)
$$

But the LHS of (9) equals $4 E_{n, r, 1}(W, R)$. We conclude that if 10 holds then $E_{r, 1}(W, R) \geq E_{n, r, 1}(W, R)$. However, if 10) does not hold then either $\hat{R}_{1}=\max (0,2 R-1)$ or $\hat{R}_{1}=\min (1,2 R)$ and the argument fails.

We now obtain a lower bound on the achievable error exponent $E_{\lambda}(W, R)$ that depends on the channel, $W$, only through its capacity, $I(W)$. By [24, Equation (26)], for a given channel capacity $I(W)$,

$$
E_{r, B S C}(I(W), R) \leq E_{r}(W, R) \leq E_{r, B E C}(I(W), R)
$$

where $E_{r, B E C}(I(W), R)\left(E_{r, B S C}(I(W), R)\right.$, respectively) is the error exponent corresponding to random codes of rate $R$ over a binary erasure channel (binary symmetric channel) of capacity $I(W)$. By [24, Equation (27), Theorem 2] this is also true for expurgated error exponents. Therefore,

$$
E_{B S C}(I(W), R) \leq E(W, R) \leq E_{B E C}(I(W), R)
$$

is also true for expurgated random codes and their error exponents as defined in (3). Recalling that

$$
I\left(W^{+}\right)-I(W)=I(W)-I\left(W^{-}\right) \in\left[\epsilon_{l}(I(W)), \epsilon_{h}(I(W))\right]
$$

(see Section $\boxed{\square}$, we define

$$
\hat{E}_{\lambda}(I(W), R) \triangleq 0.5 \min _{\epsilon \in\left[\epsilon_{l}, \epsilon_{h}\right]} \max _{R_{1} \in \mathcal{B}} \hat{E}_{m, \lambda}\left(I(W), \epsilon, R, R_{1}\right)
$$

where $\epsilon_{l}=\epsilon_{l}(I(W)), \epsilon_{h}=\epsilon_{h}(I(W))$,

$$
\hat{E}_{m, \lambda}\left(I(W), \epsilon, R, R_{1}\right)=\min \left[\hat{E}_{\lambda-1}\left(I(W)-\epsilon, R_{1}\right), \hat{E}_{\lambda-1}\left(I(W)+\epsilon, 2 R-R_{1}\right)\right]
$$

and $\hat{E}_{0}(I(W), R)=E_{B S C}(I(W), R)$ Note that $\hat{E}_{\lambda}(I(W), R)$ does not depend on the channel $W$, but only on its capacity.
Theorem 1: For any BMS channel $W$ with capacity $I(W)$, any desired code rate $R$, and a concatenated code with $\lambda \geq 0$ polarization steps for the inner polar code and (randomly generated linear) outer codes with blocklength $N_{1} \rightarrow \infty$, the best achievable error exponent $E_{\lambda}(W, R)$ is lower bounded by $E_{\lambda}(W, R) \geq \hat{E}_{\lambda}(I(W), R)$.

Proof: We will prove the theorem using induction. The claim is trivial for $\lambda=0$. We will prove it for $\lambda$ assuming it is true for $\lambda-1$.

$$
\begin{aligned}
E_{\lambda}(W, R) & =0.5 \max _{R_{1} \in \mathcal{B}} E_{m, \lambda}\left(W, R, R_{1}\right) \\
& \geq 0.5 \max _{R_{1} \in \mathcal{B}} \min \left[\hat{E}_{\lambda-1}\left(I\left(W^{-}\right), R_{1}\right), \hat{E}_{\lambda-1}\left(I\left(W^{+}\right), 2 R-R_{1}\right)\right]
\end{aligned}
$$



Fig. 3. Lower bounds for error exponents of codes created by $\lambda$ polar transforms followed by $2^{\lambda}$ outer codes of blocklength $N / 2^{\lambda}$ and rate $R=1 / 2$ with optimal rate-splitting, and naively using $2^{\lambda}$ outer codes of blocklength $N / 2^{\lambda}$, for a general BMS channel with capacity $I(W)$ and total rate $R=1 / 2$.

$$
\geq 0.5 \min _{\epsilon \in\left[\epsilon_{l}, \epsilon_{h}\right]} \max _{R_{1} \in \mathcal{B}} \hat{E}_{m, \lambda}\left(I(W), \epsilon, R, R_{1}\right)=\hat{E}_{\lambda}(I(W), R)
$$

where the first equality follows from (6), the first inequality follows from the induction assumption, and the second one follows from the fact that there exists $\epsilon_{l}(I(W)) \leq \epsilon \leq \epsilon_{h}(I(W))$ s.t. $I(W)-I\left(W^{-}\right)=I\left(W^{+}\right)-I(W)=\epsilon$.
We thus showed that $\hat{E}_{\lambda}(I(W), R)$ is a lower bound on the error exponent $E_{\lambda}(W, R)$ for all BMS channels with a given capacity $I(W)$. In Fig. 3, we plot the asymptotic lower bound for the error exponent, $\hat{E}_{\lambda}(I(W), R)$, for a code with rate $R=1 / 2$ and $\lambda=1,2,3$ as a function of $I(W)$. This is compared to the lower bound for the naive approach, where we use $2^{\lambda}$ codewords with blocklength $N_{1}=N / 2^{\lambda}$, without using the polar transformation. This lower bound is $E_{B S C}(I(W), R) / 2^{\lambda}$. As expected, the lower bound for the scheme with polarization is higher, i.e. $\hat{E}_{\lambda} \geq \hat{E}_{n, \lambda}$. All the plots in Fig. 3 have an asymptote at $I(W)=R=1 / 2$, since when the rate approaches capacity, $E_{B S C}, \hat{E}_{\lambda}$ and $\hat{E}_{n, \lambda}$ approach zero. We also see that $\hat{E}_{\lambda}$ and $\hat{E}_{n, \lambda}$ have an asymptote at $I(W)=1$. This asymptote follows from using expurgated codes, and since it is a lower bound on the error exponent, it follows that $E_{\lambda}$ and $E_{n, \lambda}$ approach infinity as $\mathrm{SNR} \rightarrow \infty(I(W) \rightarrow 1)$ in Fig. 2 for the BIAWGNC. It should also be noted that the difference between the lower bounds on the error exponents of the polarization-based and naive schemes increases with $\lambda$.

## IV. IMPROVED BOUNDS AND APPROXIMATIONS

## A. Achievable bound for BEC

[17. Theorem 37] provides an upper bound on the achievable frame error rate (FER) of an error-correcting code with blocklength $N$ and rate $R$ for the binary erasure channel (BEC) with erasure probability $\epsilon$. Combining this upper bound with (2), we obtain that for a polar-concatenated code with outer codes of length $N_{1}$, and rates $R_{i}, i=0, \ldots, 2^{\lambda}-1$, an achievable upper bound is

$$
\begin{align*}
P_{e} & \leq \sum_{i=0}^{2^{\lambda}-1} \min \left\{\sum_{t=0}^{N_{1}}\binom{N_{1}}{t}\left(1-I\left(W_{i}\right)\right)^{t} I\left(W_{i}\right)^{N_{1}-t} 2^{-\left[N_{1}-t-\log \left(\frac{2^{N_{1} R_{i-1}}}{2}\right)\right]^{+}}, e^{-N_{1} E\left(W_{i}, R_{i}, N_{1}\right)}\right\}  \tag{11}\\
& \triangleq \sum_{i=0}^{2^{\lambda}-1} \gamma_{i}\left(R_{i}\right)
\end{align*}
$$

where $x^{+} \triangleq \max (x, 0)$. We compute an upper bound on the achievable FER for polar codes of length $2^{\lambda}$ concatenated with codes of length $N_{1}$ and the best rate division between sub-codes given the total code rate $R$, by minimizing the RHS of (11) given $\sum_{i=0}^{2^{\lambda}-1} R_{i}=2^{\lambda} R$ (which can also be written as $\sum_{i=0}^{2^{\lambda}-1} R_{i} N_{1}=R N$ ) using the following efficient dynamic programming algorithm. Denote

$$
\delta_{l}(\rho) \triangleq \min _{R_{0}, \ldots, R_{l}} \sum_{i=0}^{l} \gamma_{i}\left(R_{i}\right)
$$

for integer $\rho \in[0, R N]$. The minimization is under the constraint that $N_{1} R_{i}$ are all positive integers, satisfying $\sum_{i=0}^{l} R_{i} N_{1}=\rho$. For each $l=0,1, \ldots, 2^{\lambda}-1$, and each integer value of $\rho \in[0, R N]$, the algorithm computes $\delta_{l}(\rho)$ recursively using

$$
\delta_{l}(\rho)=\min _{R_{l}}\left\{\delta_{l-1}\left(\rho-N_{1} R_{l}\right)+\gamma_{l}\left(R_{l}\right)\right\}
$$

subject to the constraint that $N_{1} R_{l}$ is an integer and $N_{1} R_{l} \in[0, \rho]$. The recursion is initialized using

$$
\delta_{0}(\rho)=\gamma_{0}\left(\rho / N_{1}\right)
$$

The output of the algorithm is $\delta_{2^{\lambda}-1}(R N)$, which is the minimum of the RHS of (11). The minimizing rates, $R_{0}, \ldots, R_{2^{\lambda}-1}$, can be easily obtained as a byproduct of this recursive algorithm.

## B. Dispersion-based (normal) approximation

Consider transmission over a BMS channel, $W$, with capacity $I(W)$ and dispersion $V(W)$, defined as $V(W) \triangleq \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \frac{1}{2} W(y \mid x)\left(\log \frac{W(y \mid x)}{\sum_{x^{\prime} \in \mathcal{X}} \frac{1}{2} W\left(y \mid x^{\prime}\right)}\right)^{2}-I(W)^{2}$. By [17], the maximal rate of transmission at error probability $\epsilon$ and blocklength $N$ is closely approximated by $I(W)-\sqrt{\frac{V(W)}{N}} Q^{-1}(\epsilon)$ where $Q()$ is the complementary Gaussian cumulative distribution function. The approximation improves as $N$ gets larger, but is known to be tight already for $N$ as short as about 100. The error probability of the best code with blocklength $N$ and rate $R$ can be approximated by $Q\left\{\sqrt{\frac{N}{V(W)}}\left[I(W)-R+O\left(\frac{\log N}{N}\right)\right]\right\}[17]$ (this is sometimes referred to as the normal approximation in the literature). However, for some channels this expression can be improved. For BIAWGNC [25] and BSC [17, Theorem 52], $P_{e} \approx$ $Q\left\{\sqrt{\frac{N}{V(W)}}\left[I(W)-R+\frac{\log N}{2 N}+O\left(N^{-1}\right)\right]\right\}$, and for BEC it is the same expression without $0.5 \log N / N$ [17, Theorem 53]. Therefore, for the BEC and for general channels we approximate the error probability as $Q\left(\sqrt{\frac{N}{V(W)}}(I(W)-R)\right)$, while for the BIAWGNC and BSC we use

$$
\begin{equation*}
Q\left[\sqrt{\frac{N}{V(W)}}\left(I(W)-R+\frac{\log N}{2 N}\right)\right] \tag{12}
\end{equation*}
$$

The smallest achievable error probability of our concatenated polar coding scheme is thus approximated by

$$
\begin{align*}
& \left.P_{e} \approx \min _{\substack{R_{0} \ldots R_{2 \lambda}-1 \\
R_{i} \in\left[0, I\left(W_{i}\right)\right]}} \sum_{i=0}^{\sum_{i=0}^{\lambda^{\lambda}-1} R_{i}=2^{\lambda} R}\right\}  \tag{13}\\
& \leq 2^{\lambda} \min _{\substack{R_{0} \ldots R_{2} \lambda-1 \\
R_{i} \in\left[0, I\left(W_{i}\right)\right] \\
\sum_{i=0}^{2^{\lambda}-1} R_{i}=2^{\lambda} R}} \max _{i} Q\left(\left(I\left(W_{i}\right)-R_{i}+C\left(N_{1}\right)\right) \sqrt{\frac{N_{1}}{V\left(W_{i}\right)}}\right)  \tag{14}\\
& \left.\max ^{2} Q\left(\left(W_{i}\right)-R_{i}+C\left(N_{1}\right)\right) \sqrt{\frac{N_{1}}{V\left(W_{i}\right)}}\right)
\end{align*}
$$

where $C\left(N_{1}\right)$ is a correction term. For the BIAWGNC and BSC, $C\left(N_{1}\right)=\log N_{1} /\left(2 N_{1}\right)$, and for other channels $C\left(N_{1}\right)=0$. Note that $P_{e}$ is also approximately lower bounded by the same expression in without the multiplying $2^{\lambda}$ term.

In the simulations section it will be shown that provides a tight approximation to the actual performance of concatenated BCH -polar codes over the BIAWGNC. The minimization in 13 is computed efficiently using a dynamic programming


Fig. 4. The minimal rate for which holds compared to the channel capacity for the BIAWGNC
algorithm, similarly to the algorithm that was described in Section IV-A The algorithm also provides the rates $R_{0}, \ldots, R_{2^{\lambda}-1}$ as a byproduct.

To gain insight on the min-max problem (14), we first look at the simpler problem

$$
\begin{equation*}
\min _{\substack{R_{0} \ldots R_{2^{\lambda}-1} \\ \sum_{i=0}^{2^{\lambda}-1} R_{i}=2^{\lambda} R}} \max _{i} Q\left(\left(I\left(W_{i}\right)-R_{i}\right) \sqrt{\frac{N_{1}}{V\left(W_{i}\right)}}\right) \tag{15}
\end{equation*}
$$

This problem is solved when $\frac{I\left(W_{i}\right)-R_{i}}{\sqrt{V\left(W_{i}\right)}}, i=0, \ldots, 2^{\lambda}-1$ are equal and $\sum_{i=0}^{2^{\lambda}-1} R_{i}=2^{\lambda} R$. Hence, $I\left(W_{i}\right)-R_{i}=$ $\frac{2^{\lambda}[I(W)-R] \sqrt{V\left(W_{i}\right)}}{\sum_{j=0}^{2 \lambda-1} \sqrt{V\left(W_{j}\right)}}$, and the solution of 15$]$ is $Q\left([I(W)-R] \sqrt{\frac{N}{V_{\lambda}(W)}}\right)$ where

$$
\begin{equation*}
V_{\lambda}(W) \triangleq \frac{\left(\sum_{i=0}^{2^{\lambda}-1} \sqrt{V\left(W_{i}\right)}\right)^{2}}{2^{\lambda}} \tag{16}
\end{equation*}
$$

Since $\sqrt{15]}$ is a relaxed version of the min-max problem in $\sqrt[14]{ }$, if the solution of $\sqrt[15]{ }$ ) obeys all the constraints of the min-max problem in 14, then it is the solution of this problem as well. Therefore, if $\forall i=0, \ldots, 2^{\lambda}-1$

$$
\begin{equation*}
R_{i}=I\left(W_{i}\right)-\frac{2^{\lambda}[I(W)-R] \sqrt{V\left(W_{i}\right)}}{\sum_{j=0}^{2^{\lambda}-1} \sqrt{V\left(W_{j}\right)}} \geq 0 \tag{17}
\end{equation*}
$$

then

$$
\begin{equation*}
P_{e} \approx 2^{\lambda} Q\left([I(W)-R] \sqrt{\frac{N}{V_{\lambda}(W)}}\right) \tag{18}
\end{equation*}
$$

For $R$ sufficiently close to $I(W)$ the condition (17) is guaranteed to hold. The minimal $R$, for which (17) holds is

$$
\begin{equation*}
R_{\min }=I(W)-\min _{i}\left(I\left(W_{i}\right) \sqrt{\frac{V_{\lambda}(W)}{2^{\lambda} V\left(W_{i}\right)}}\right) \tag{19}
\end{equation*}
$$

Fig. 4 shows $R_{\min }$ for different values of $\lambda$ for the BIAWGNC. Fig. 4 suggests that $\lim _{S N R \rightarrow \infty} R_{\min }=1-1 / 2^{\lambda}$. This will be proved theoretically later. In this entire derivation, we ignore the requirement that $R_{i} \cdot N_{1} \in \mathbb{Z}$, since this requirement becomes redundant as $N_{1}$ increases. For the same reason, we also ignore the correction term. This is compared to a naive


Fig. 5. $V_{\lambda}(W)$ and $V_{n, \lambda}(W)$ for a BIAWGNC
approach where we simply use $2^{\lambda}$ codewords with blocklength $N_{1}=N / 2^{\lambda}$ (without using the polar transformation at all). In this approach,

$$
\begin{equation*}
P_{e} \approx 2^{\lambda} Q\left([I(W)-R] \sqrt{\frac{N_{1}}{V(W)}}\right)=2^{\lambda} Q\left([I(W)-R] \sqrt{\frac{N}{V_{n, \lambda}(W)}}\right) \tag{20}
\end{equation*}
$$

where $V_{n, \lambda}(W) \triangleq 2^{\lambda} V(W)$. Showing that $V_{\lambda}(W) \leq V_{n, \lambda}(W)$ would mean that polarization is helpful. As can be seen in Fig. 5 for codes of rate $R>R_{\min }$, transmitted over the BIAWGNC, $V_{\lambda}(W)<V_{n, \lambda}(W)$, so polarization helps. For high SNR, Fig. 5 suggests that $V_{\lambda}(W) \approx V(W)$. This can be proved theoretically, but it requires the following lemma first.

Lemma 4: If $R$ is sufficiently large, s.t. (17) holds, i.e. $R>R_{\min }$, then, for $\lambda \geq 1, V_{\lambda}(W)$ can be calculated recursively using $V_{\lambda}(W)=0.5\left(\sqrt{V_{\lambda-1}\left(W^{-}\right)}+\sqrt{V_{\lambda-1}\left(W^{+}\right)}\right)^{2}$ where $V_{0}(W) \triangleq V(W)$.

Proof: For $\lambda>1$, note that $W_{i}$ are sub-channels of $W^{-}$for $i=0, \ldots, 2^{\lambda-1}-1$, and sub-channels of $W^{+}$for $i=$ $2^{\lambda-1}, \ldots, 2^{\lambda}-1$.

$$
\begin{aligned}
V_{\lambda}(W) & =0.5\left(\frac{\sum_{i=0}^{2^{\lambda-1}-1} \sqrt{V\left(W_{i}\right)}}{\sqrt{2^{\lambda-1}}}+\frac{\sum_{i=2^{\lambda-1}}^{2^{\lambda}-1} \sqrt{V\left(W_{i}\right)}}{\sqrt{2^{\lambda-1}}}\right)^{2} \\
& =0.5\left(\sqrt{V_{\lambda-1}\left(W^{-}\right)}+\sqrt{V_{\lambda-1}\left(W^{+}\right)}\right)^{2}
\end{aligned}
$$

where the first equality follows from rewriting (16), and the second one follows from applying (16) for $\lambda-1$ instead of $\lambda$. We now consider the case of $I(W)$ close to 1 (the high SNR case) and claim the following.

Lemma 5: Consider a BMS channel $W$ and suppose that $V(W)$ can be linearly approximated as ${ }^{4} V(W)=\alpha \delta(W)+o(\delta(W))$ where $\alpha$ is some constant and $\delta(W) \triangleq 1-I(W)$. Also assume that $W^{+}$and $W^{-}$satisfy the same property (with the same value of $\alpha$ ). Then $V_{\lambda}(W)=V(W)+o(\delta(W))$ for $\lambda \geq 1$. Furthermore, under our assumptions, $\lim _{I(W) \rightarrow 1} R_{\min }=1-1 / 2^{\lambda}$.

We prove the Lemma in Appendix B
The conditions of Lemma 5 hold in particular for the BEC. The condition regarding $W$ holds with $\alpha=1$ since $V(W)=$ $I(W)[1-I(W)]$. The conditions regarding $W^{+}$and $W^{-}$hold since by [1, Proposition 6], if $W$ is a BEC, so are $W^{+}$and $W^{-}$. For the BIAWGNC, [24, Fig. 4] suggests that the condition on $W$ is valid. Furthermore, if $W$ is a BIAWGNC, then $W^{+}$

[^3]and $W^{-}$can be approximated as BIAWGNCs as well [8]. These arguments suggest that the conditions of Lemma 5 are also valid for the BIAWGNC. Fig. 5 supports this conjecture.

Consider transmission over a BMS channel $W$ with code rate $R>R_{\min }$, where $I(W)$ is sufficiently close to 1 . One could think that under the conditions of Lemma 5 on $W, W^{-}$and $W^{+}$, and the normal approximation to the error probability, the best achievable error probability of the polar concatenated coding scheme, 18 , is larger only by a factor of $2^{\lambda}$ compared to the best achievable error probability of an arbitrary code with the same total blocklength and rate, i.e. $Q(x / \sqrt{V(W)}) / Q\left(x / \sqrt{V_{\lambda}(W)}\right) \approx$ 1 for $x \triangleq \sqrt{N}[I(W)-R]$ and $I(W)$ close to 1 .

Unfortunately, this is not true. Since $Q(u) \sim \exp \left(-0.5 u^{2}\right)$ for large $u$, and, according to the conditions and claims of Lemma 5, $\lim _{I(W) \rightarrow 1} V(W), V_{\lambda}(W)=0$, we have

$$
\begin{equation*}
Q(x / \sqrt{V(W)}) / Q\left(x / \sqrt{V_{\lambda}(W)}\right) \approx \exp \left[-\frac{x^{2}}{2}\left(\frac{1}{V(W)}-\frac{1}{V_{\lambda}(W)}\right)\right] \tag{21}
\end{equation*}
$$

for $I(W)$ sufficiently close to 1 . The condition

$$
\lim _{I(W) \rightarrow 1} \frac{1}{V(W)}-\frac{1}{V_{\lambda}(W)}=0
$$

is required for claiming that the RHS of 21) approaches one for $I(W) \rightarrow 1$. This condition is stronger than the one proved in Lemma 5. and it does not necessarily hold, since $1 / V(W)-1 / V_{\lambda}(W)=\frac{o(\delta)}{\alpha^{2} \delta^{2}+\alpha o\left(\delta^{2}\right)}$ does not necessarily approach zero for $\delta \rightarrow 0$. I removed the example illustrated in the previous Fig. 6 in order to focus on our main contributions. As an example, it is easy to prove using Lemma 4, that for the BEC and $I(W)=1-\delta, V(W)=\delta(1-\delta)$, but $V_{1}(W)=\delta+\Theta\left(\delta^{1.5}\right)$, so $1 / V(W)-1 / V_{1}(W)=\Theta\left(\delta^{-0.5}\right)$ and the RHS of 21) approaches zero for $I(W) \rightarrow 1$.

Suppose that the condition $R>R_{\min }$ holds. Then the error rate of the polar concatenated (naive) scheme is given by (18) ( (20), respectively). Hence, comparing $V_{\lambda}(W)$ to $V_{n, \lambda}(W)$, we can assess the usefulness of polarization compared to the naive approach with the same value of $\lambda$. The following result shows that for large $N, V_{\lambda}(W)$ and $V_{n, \lambda}(W)$ can be used to compare between schemes with different values of $\lambda$. Define $\hat{V}_{\lambda}(N, W)$ as the solution to

$$
\begin{equation*}
2^{\lambda} Q\left([I(W)-R] \sqrt{\frac{N}{V_{\lambda}(W)}}\right)=Q\left([I(W)-R] \sqrt{\frac{N}{\hat{V}_{\lambda}(N, W)}}\right) \tag{22}
\end{equation*}
$$

for given $I(W)-R, \lambda$ and $N$. Also define $\hat{V}_{n, \lambda}(N, W)$ as the solution of

$$
2^{\lambda} Q\left([I(W)-R] \sqrt{\frac{N}{V_{n, \lambda}(W)}}\right)=Q\left([I(W)-R] \sqrt{\frac{N}{\hat{V}_{n, \lambda}(N, W)}}\right)
$$

Lemma 6: $\hat{V}_{\lambda}(N, W)-V_{\lambda}(W)=\Theta\left(\frac{1}{N}\right)$, so $\lim _{N \rightarrow \infty} \hat{V}_{\lambda}(N, W)=V_{\lambda}(W)$. The same claim also holds for $\hat{V}_{n, \lambda}(N, W)$ and $V_{n, \lambda}(W)$.
We prove the lemma in Appendix C
For asymptotically large blocklengths the error probabilities predicted by Gallager's error exponents are more accurate (the error exponents are conjectured to be the correct ones under our sub-optimal decoding scheme). However, in the finite blocklength regime, the normal approximation is better. In our setting, we target the case of $N_{1}$ values which are not very large (e.g., of order 100). In the following section we show a very good match between the normal approximation and simulation results, using close to ML decoded, powerful outer algebraic codes.

## C. Approximated converse for BIAWGNC

In [26], a lower bound on the optimal FER of optimal spherical codes over additive white Gaussian noise channels (AWGNCs) is provided. Since the BIAWGNC is a constrained version of a AWGNC with the same SNR, this bound can be treated as a lower bound on the FER for a BIAWGNC as well, i.e. for the same $\mathrm{SNR}=1 / \sigma^{2}$, rate $R$ and block size $N, P_{e, B I A W G N C}(\sigma, R, N) \geq$ $P_{e, A W G N C}(\sigma, R, N)$. We obtain the following approximated lower bound on the achievable frame error rate after $\lambda$ polarization steps.

$$
\begin{align*}
P_{e} & \geq 1-\prod_{i=0}^{2^{\lambda}-1}\left[1-P_{e}\left(W_{i}, R_{i}, N_{1}\right)\right]  \tag{23}\\
& \approx 1-\prod_{i=0}^{2^{\lambda}-1}\left[1-P_{e, \text { BIAWGN }}\left(\sigma_{i}, R_{i}, N_{1}\right)\right] \\
& \geq 1-\prod_{i=0}^{2^{\lambda}-1}\left[1-P_{e, A W G N}\left(\sigma_{i}, R_{i}, N_{1}\right)\right] \\
& \geq \min _{R_{i}}\left[1-\prod_{i=0}^{2^{\lambda}-1}\left[1-P_{e, A W G N}\left(\sigma_{i}, R_{i}, N_{1}\right)\right]\right]
\end{align*}
$$

The first inequality is under the assumption that each outer code $A_{i}$ is ML decoded given the channel observations and the previously decoded codewords of the outer codes $A_{0}, \ldots, A_{i-1}$. We bound the error rate from below by assuming a genie aided decoder: the genie informs us what was the actual transmitted codeword of the outer code, $A_{i}$, immediately after we decode it (so that it can be used for decoding the codewords of the following outer codes, $A_{i+1}, A_{i+2}, \ldots$ ). The approximation in the second line follows from the approximation of the sub-channels $W_{i}$ as BIAWGNCs, when $W$ is a BIAWGNC [8, Equations (7)-(8)]. The second inequality follows from the explanation above. The term $P_{e, A W G N}(\sigma, R, N)$ is calculated using [26]

$$
P_{e, A W G N}(\sigma, R, N) \approx \frac{1}{\sqrt{N \pi}} \frac{1}{\sqrt{1+G^{2}} \sin \theta} \cdot \frac{\left[G \sin \theta \exp \left(-\frac{1}{2 \sigma^{2}}+\frac{G \cos \theta}{2 \sigma}\right)\right]^{N}}{G \sin ^{2} \theta / \sigma-\cos \theta}
$$

where $G=0.5\left(\cos \theta / \sigma+\sqrt{\cos ^{2} \theta / \sigma^{2}+4}\right)$ and $\theta$ is computed by solving $2^{N R} \approx \frac{\sqrt{2 \pi N} \sin \theta \cos \theta}{\sin ^{N} \theta}$. The two approximations above are extremely accurate for $N \geq 100$.

## D. Converse for the BEC

Combining (23) with [17, Theorem 38] and the fact that the sub-channels of a BEC are also BEC, yields that for all polar concatenated codes over the BEC with capacity $I(W), \lambda$ polarization steps, and blocklength $N_{1}$

$$
\begin{align*}
P_{e} & \geq 1-\prod_{i=0}^{2^{\lambda}-1}\left[1-P_{e}\left(W_{i}, R_{i}, N_{1}\right)\right] \\
& \geq 1-\prod_{i=0}^{2^{\lambda}-1}\left[1-P_{c, B E C}\left(I\left(W_{i}\right), R_{i}, N_{1}\right)\right] \\
& \geq \min _{R_{i}}\left[1-\prod_{i=0}^{2^{\lambda}-1}\left[1-P_{c, B E C}\left(I\left(W_{i}\right), R_{i}, N_{1}\right)\right]\right] \tag{24}
\end{align*}
$$

where

$$
P_{c, B E C}(I(W), R, N) \triangleq \sum_{l=\lfloor N(1-R)\rfloor+1}^{N}\binom{N}{l}[1-I(W)]^{l} I(W)^{N-l}\left(1-2^{N(1-R)-l}\right)
$$

TABLE I
Rates of outer codes in the $\left(\lambda, N_{1}, R\right)=(1,64,1 / 2)$ SCHEME FOR THE BIAWGNC

| SNR[dB] | $N_{1} \cdot R_{0}$ | $N_{1} \cdot R_{1}$ |
| :---: | :---: | :---: |
| 1 | 17 | 47 |
| 1.5 | 16 | 48 |
| 2 | 16 | 48 |
| 2.5 | 16 | 48 |
| 3 | 15 | 49 |
| 3.5 | 15 | 49 |
| 4 | 14 | 50 |

TABLE II
Rates of outer codes in the $\left(\lambda, N_{1}, R\right)=(1,128,1 / 2)$ SCHEME FOR THE BIAWGNC

| $\mathrm{SNR}[\mathrm{dB}]$ | $N_{1} \cdot R_{0}$ | $N_{1} \cdot R_{1}$ |
| :---: | :---: | :---: |
| 0.5 | 34 | 94 |
| 1 | 33 | 95 |
| 1.5 | 33 | 95 |
| 2 | 32 | 96 |
| 2.5 | 31 | 97 |
| 3 | 31 | 97 |
| 3.5 | 30 | 98 |

## V. Comparison with simulation results

We start by comparing the actual performance of BCH-polar codes, for short blocklengths and one polarization step $(\lambda=1)$ transmitted over a BIAWGNC, with the normal approximation-based expression (13), for the best polar-concatenated code with these parameters using the dynamic programming algorithm described in Section IV. The total code rate was $R=1 / 2$. We used two setups. In the first, $N_{1}=64$ and $N=128$. In the second setup, $N_{1}=128$ and $N=256$. We also calculated the normal approximation $\widehat{12}$ to the best achievable error probability when using a $(128,64)((256,128))$ code. For each SNR, the outer code rates that minimize (13) were calculated as a by-product of the dynamic programming algorithm, and are shown in Tables $\square$ and $\square$

Due to the results in Table I, we used $(64,18,22)$ and $(64,45,8)$ extended BCH codes, whose generator matrices appear in [27], as outer codes in the simulated BCH-polar coding scheme for $N=128$, and decoded them using OSDs of order 5 [27]. The total rate of the scheme is $R=63 / 128$, which is close to the planned rate. The $(128,64,22)$ BCH code was decoded using OSD of order 5 as well. The consumed processor time (measured in clock ticks) normalized by the number of information bits of the decoders of BCH -polar and BCH codes were compared as well. The error rates and decoding times of the various schemes are shown in Figs. 6 and 7 Fig. 6 suggests that the normal approximation we use is accurate in the SNR range examined. We also see that the BCH -polar code suffers a loss of 0.7 dB compared to the BCH code (Fig. 6), but is about 25 to 1000 times faster, depending on the SNR (Fig. 7). Compared to the $(128,64)$ list SC decoder with CRC (with list size 32), whose FER results were taken from [28], the BCH-polar code suffers a loss of only 0.25 dB .

For extended BCH with $N=128$ and $R=1 / 2$ we see in Fig. 6 that the normal approximation is accurate for low SNR, but for high SNR it slightly underestimates the FER. We conjecture the same behavior for other blocklengths and rates. Therefore,


Fig. 6. FER normal approximations and simulation results for BCH -polar codes with $\lambda=1$, and BCH codes with different blocklengths and rates close to $1 / 2$ for the BIAWGNC, compared with $(128,64)$ List decoder with CRC from [28]
using (13) to estimate the optimal rates for BCH-polar codes would be accurate for low SNR (considering that not all BCH rates are feasible), but for high SNR it would overestimate the rates of the good sub-channels. That's why we pick slightly lower $R_{1}$ and slightly higher $R_{0}$ than the ones in Table $\square$ while designing the BCH-polar codes.

The results for $N=256$ show a similar trend. Due to the rates in Table $\Pi$, and since $R_{1}$ should be slightly lowered (and $R_{0}$ slightly increased) the chosen extended BCH codes were $(128,36,32)$ and $(128,92,12)$, whose generator matrices appear in [27]. For comparison, we have taken the FER results of the $(255,131) \mathrm{BCH}$ code from [29], and measured the average processing time of the OSD of order 5 of this code. This time, the BCH-polar code suffers a loss of only 0.5 dB compared to the BCH code (Fig. 6), but is about 1000 times faster (Fig. 7. Figs. 6 and 7 also show the following:

1) For $\mathrm{SNR}<2 \mathrm{~dB}$, the $(256,128) \mathrm{BCH}$-polar code has a higher FER compared to the $(128,64) \mathrm{BCH}$ code, but it requires a lower processing time per information bit.
2) For $2 \mathrm{~dB}<\mathrm{SNR}<3 \mathrm{~dB}$, the $(256,128) \mathrm{BCH}$-polar code has a lower FER and lower processing time per information bit compared to the $(128,64) \mathrm{BCH}$ code.
3) For $\mathrm{SNR}>3 \mathrm{~dB}$, the $(256,128) \mathrm{BCH}$-polar code has a lower FER than the $(128,64) \mathrm{BCH}$ code, but it requires higher processing time per information bits.

Note that when decoding a $(256,128)$ BCH-polar code, we need to decode two BCH codes of blocklength 128. However, the rate of the first is lower than $1 / 2$ while the rate of the second is higher than $1 / 2$. Also, the required complexity of the OSD algorithm is maximal for rate $1 / 2$ codes. Therefore, it is usually more efficient to decode a $(256,128)$ BCH-polar code compared to the decoding of two $(128,64) \mathrm{BCH}$ codes. However, the decoding of the $(256,128) \mathrm{BCH}$-polar code also requires some handling of soft information due to the polar transformation. We did not attempt to implement this part in our algorithm efficiently.

We also compared our results to simulation results from the literature. First, we considered the setup in [9, Section VIII.A], where $\lambda=3, N_{1}=128, N=1024$, and the channel is a BEC. The overall rate of the code is $R=0.4$. We calculated the achievable upper bound on the FER, with typical random (the expurgated random ensemble) outer codes, using the recursive algorithm from Lemma 1 We then computed an upper bound on the achievable FER by finding the best rate division between


Fig. 7. Comparison of processor time per information bits of BCH -polar codes with $\lambda=1$, and BCH codes with different lengths and rate close to $1 / 2$ for the BIAWGNC.


Fig. 8. FER upper bounds for concatenated polar codes with outer code length $N_{1}=128$ and inner polar code length 8, compared with BCH-polar code results from 9 Fig. 5].
sub-codes given the total code rate $2^{\lambda} R$, that brings the RHS of (11) to a minimum as described in Section IV-A Our upper bounds were computed twice. Once by optimizing the rate division for a fixed BEC, $W$, with erasure rate 0.4 as in [9], and once by optimizing the rate division for the actual BEC we are transmitting over, for each point in the graph. The corresponding graphs are denoted by " $\operatorname{BEC}(0.4)$ " and "opt." in Fig. 8. We have also plotted the normal approximation (13) to the best achievable error probability, and the converse bound in Section IV-D.

The graphs show small gaps between the achievable bound, Equation 11), and the actual results with BCH codes. The normal approximation (13) has a somewhat lower FER, while the bound based on Lemma 1 is less tight. The converse (24) is close to the normal approximation. For comparison, note that a standard SC decoded polar code of length $N=1024$, yields $\mathrm{FER} \approx 2 \cdot 10^{-3}$ for $I(W)=0.6$ [9].

The second setup we considered is taken from [8, Section IV], where $\lambda=3, N_{1}=127, N=1016$, and the channel is a


Fig. 9. FER upper bounds for concatenated polar codes with outer random code length $N_{1}=127$, inner polar code length 8 , compared with BCH-polar code from [8]

BIAWGNC. The overall rate of the code is $R=1 / 2$. We first calculated the achievable upper bound on the FER from Lemma 1. Then we calculated the normal approximation (13) to the best achievable error probability. Both graphs were obtained by optimizing over the best rate split for each SNR point in the graph. We also plot the normal approximation given optimal rate division for fixed $\mathrm{SNR}=3 \mathrm{~dB}$. We compared these bounds with a simulation of the BCH-polar code in [8] using an outer OSD of order 5. Note that in [8] the outer code rates were optimized only for $\mathrm{SNR}=3 \mathrm{~dB}$. As can be seen, the normal approximation is close to the performance of the scheme with outer BCH codes and almost not affected by using a fixed rate division. The figure also shows the performance of a standard polar code with $N=1024$ and $R=1 / 2$. Finally, we have plotted the converse for polar concatenated scheme as was described in Section IV-C, Equation (23). We see that for a desired FER of $\approx 2 \cdot 10^{-4}$ the BCH-polar code is only 0.75 dB worse than the converse, so under the given $\lambda, N_{1}$ and $R$, we cannot gain more than 0.75 dB by smartly choosing the outer codes. Finally note that the normal approximation for the best $(1016,508)$ code at $\mathrm{SNR}=1.25 \mathrm{~dB}$ shows only 1 dB improvement compared to the BCH -polar code.

## VI. CONCLUSION

We studied the properties of a concatenated scheme of polar codes with good outer codes. We obtained an upper bound on the error exponent using the corresponding expurgated random coding ensemble, and calculated a lower bound on the error exponent, which is valid for all channels with a given capacity. We obtained converse and approximated converse results, as well as a dispersion-based normal approximation to the performance for finite length codes, which can also be used to determine the required rate split between the outer codes. We showed good agreement between this prediction and simulation results for BCH-polar codes, when transmitting over the BIAWGNC.

## Appendix A

## Proof of Lemma 3

We first prove by induction that $E_{\lambda}(W, R)$ is finite for all integer $\lambda$ and $R>0 . E_{\lambda}(W, R)$ is finite for $\lambda=0$ and $R>0$, since both $E_{r}(W, R)$ and $E_{e x}(W, R)$ are finite: $E_{r}(W, R) \leq E_{0}(W, 1)$, and for binary-input channels (except for a perfect channel for which the two inputs can never be confused at the receiver), $E_{e x}(W, R)$ is finite for $R>0$, which can be derived from [21,

Equations (5.7.14)-(5.7.16)]. Now, assume $E_{\lambda-1}(W, R)$ is a finite function of $R$ for $R>0$. Then, $E_{m, \lambda}\left(W, R, R_{1}\right)$ defined in (7) is finite, since $E_{\lambda-1}\left(W^{-}, R_{1}\right)$ can be infinite only for $R_{1}=0$, but then $2 R-R_{1}=2 R>0$ and $E_{\lambda-1}\left(W^{+}, 2 R-R_{1}\right)$ is finite, so $E_{m, \lambda}\left(W, R, R_{1}\right)$ is finite too. This also means that $E_{\lambda}(W, R)=0.5 \max _{R_{1}} E_{m, \lambda}\left(W, R, R_{1}\right)$ is finite. This shows that $E_{\lambda}(W, R)$ is finite for all integer $\lambda$ and $R>0$. Note however that $E_{\lambda}(W, 0)=\infty$ for all integer $\lambda$ and any channel $W$. This can be shown by induction, since for $R=0$ yields $E_{\lambda}(W, 0)=0.5 \min \left[E_{\lambda-1}\left(W^{+}, 0\right), E_{\lambda-1}\left(W^{-}, 0\right)\right]$ and $E_{\lambda}(W, 0)=\infty$ for $\lambda=0$ due to (1).

Next, we prove by induction that $E_{\lambda}(W, R)$ is decreasing in $R$. The claim is trivial for $\lambda=0$ since $E_{0}(W, R)=E(W, R)$ is defined in (3) as a maximum between two continuous, decreasing functions. Hence it is decreasing whether it equals $E_{e x}(W, R)$ or $E_{r}(W, R)$. Now assume that $E_{\lambda-1}(W, R)$ is decreasing in $R$ for any $W$. First consider rates $R$ for which

$$
\left\{\begin{array}{l}
E_{\lambda-1}\left(W^{-}, 0^{+}\right) \geq E_{\lambda-1}\left(W^{+}, 2 R\right)  \tag{25}\\
E_{\lambda-1}\left(W^{-}, 2 R\right) \leq E_{\lambda-1}\left(W^{+}, 0^{+}\right)
\end{array}\right.
$$

(for rates $R>1$ we define for all $\lambda, E_{\lambda}(W, R) \equiv 0$ ). This case is depicted in Fig. 10 . By (6) and Fig. 10 . $2 E_{\lambda}(W, R)$ is the height of the intersection point of $E_{\lambda-1}\left(W^{-}, R_{1}\right)$, which is a decreasing function of $R_{1}$, and $E_{\lambda-1}\left(W^{+}, 2 R-R_{1}\right)$, which is an increasing function of $R_{1}$. Therefore, increasing $R$ would move $E_{\lambda-1}\left(W^{+}, 2 R-R_{1}\right)$ to the right, as can be seen in Fig. 10, thus decreasing the intersection point height, $2 E_{\lambda}(W, R)$, and increasing $\hat{R}_{1}=\operatorname{argmax}_{R_{1}} \min \left[E_{\lambda-1}\left(W^{-}, R_{1}\right), E_{\lambda-1}\left(W^{+}, 2 R-R_{1}\right)\right]$. This means $E_{\lambda}(W, R)$ is decreasing, and $\hat{R}_{1}=\hat{R}_{1}(R)$ is an increasing function of $R$, as can be seen in Fig. $10 \hat{R}_{2} \triangleq \hat{R}_{1}(R+\Delta R)>\hat{R}_{1}(R)$ for $\Delta R>0$.

If (25) does not hold, i.e., $E_{\lambda-1}\left(W^{-}, 0^{+}\right)<E_{\lambda-1}\left(W^{+}, 2 R\right)\left(E_{\lambda-1}\left(W^{+}, 0^{+}\right)<E_{\lambda-1}\left(W^{-}, 2 R\right)\right.$, respectively), $\hat{R_{1}}=$ $\max (0,2 R-1)\left(\hat{R}_{1}=\min (1,2 R)\right)$ and $E_{\lambda}(W, R)=0.5 E_{\lambda-1}\left(W^{-}, \max \left(0^{+}, 2 R-1\right)\right)\left(0.5 E_{\lambda-1}\left(W^{+}, 2 R-\min (1,2 R)\right)\right)$. Trivially, $E_{\lambda}(W, R)$ is decreasing in this case as well. Thus we have shown that $E_{\lambda}(W, R)$ is decreasing in $R$ for all integer $\lambda$, and that $\hat{R}_{1}$ is an increasing (but not strictly increasing) function of $R$.

We proceed by proving by induction that $E_{\lambda}(W, R)$ is convex. The claim is trivial for $\lambda=0$ since $E_{0}(W, R)=E(W, R)$ is defined in (3) as a maximum between two continuous, convex functions, $E_{r}(W, R)$ and $E_{e x}(W, R)$, and since maximization preserves convexity. Now assume that $E_{\lambda-1}(W, R)$ is convex in $R$ for all BMS channels $W$. Since it was already shown that $E_{\lambda-1}(W, R)$ is also decreasing for all BMS channels $W$, it follows that $\left|\partial E_{\lambda-1}\left(W^{-}, r\right) / \partial r\right|_{r=\hat{R}_{1}(R)}$ is decreasing in $R$ (recalling that $\hat{R}_{1}(R)$ is an increasing function of $R$ ). Since $E_{\lambda-1}\left(W^{+}, 2 R-\hat{R_{1}}\right)=2 E_{\lambda}(W, R)$, and $E_{\lambda}(W, R)$ is a decreasing function of $R, E_{\lambda-1}\left(W^{+}, 2 R-\hat{R}_{1}(R)\right)$ is also decreasing in $R$. But $E_{\lambda-1}\left(W^{+}, r\right)$ is decreasing in $r$. Thus $2 R-\hat{R}_{1}(R)$ is increasing in $R$. Since in addition $E_{\lambda-1}\left(W^{+}, r\right)$ is convex and decreasing, it follows that $\left|\partial E_{\lambda-1}\left(W^{+}, 2 R-r\right) / \partial r\right|_{r=\hat{R}_{1}(R)}$ is decreasing in $R$. We proceed by claiming that

$$
\begin{equation*}
\left|\partial E_{\lambda}(W, R) / \partial R\right|=\left(\frac{1}{\left|\partial E_{\lambda-1}\left(W^{-}, r\right) / \partial r\right|_{r=\hat{R}_{1}(R)}}+\frac{1}{\left|\partial E_{\lambda-1}\left(W^{+}, 2 R-r\right) / \partial r\right|_{r=\hat{R}_{1}(R)}}\right)^{-1} \tag{26}
\end{equation*}
$$

Assume that 26) holds. Since both $\left|\partial E_{\lambda-1}\left(W^{-}, r\right) / \partial r\right|_{r=\hat{R}_{1}(R)}$ and $\left|\partial E_{\lambda-1}\left(W^{+}, 2 R-r\right) / \partial r\right|_{r=\hat{R}_{1}(R)}$ decrease with $R$, it follows that $\left|\partial E_{\lambda}(W, R) / \partial R\right|$ is also decreasing in $R$. Since, in addition, $\partial E_{\lambda}(W, R) / \partial R<0$ (due to the fact that $E_{\lambda}(W, R)$ is decreasing in $R$ ), this means $E_{\lambda}(W, R)$ is convex.

It remains to prove 26. First assume that 25) holds. Consider Fig. 10, which shows $E_{\lambda-1}\left(W^{-}, r\right), E_{\lambda-1}\left(W^{+}, 2 R-r\right)$ and $E_{\lambda-1}\left(W^{+}, 2(R+\Delta R)-r\right)(\Delta R$ is small fixed value) as a function of $r$, together with intersection points standing for $2 E_{\lambda}(W, R)$ and $2 E_{\lambda}(W, R+\Delta R)$. Now, $2 \Delta E \triangleq 2 E_{\lambda}(W, R)-2 E_{\lambda}(W, R+\Delta R)$ is the height of the triangle formed by the


Fig. 10. Functions required to calculate $E_{\lambda}(W, R)$ and its derivative.
vertices $\left(\hat{R}_{1}, 2 E_{\lambda}(W, R)\right),\left(\hat{R}_{2}, 2 E_{\lambda}(W, R+\Delta R)\right)$ and $\left(\hat{R}_{1}+2 \Delta R, 2 E_{\lambda}(W, R)\right)$ in Fig. 10. For $\Delta R \rightarrow 0$, the slopes of this triangle's edges are approximated as

$$
\begin{aligned}
& a \triangleq 2 \Delta E /\left(\hat{R}_{2}-\hat{R}_{1}\right) \approx\left|\partial E_{\lambda-1}\left(W^{-}, r\right) / \partial r\right|_{r=\hat{R}_{1}} \\
& b \triangleq 2 \Delta E /\left(\hat{R}_{1}+2 \Delta R-\hat{R}_{2}\right) \approx\left|\partial E_{\lambda-1}\left(W^{+}, 2 R-r\right) / \partial r\right|_{r=\hat{R}_{1}}
\end{aligned}
$$

It can be easily verified that $2 \Delta E=2 \Delta R(1 / a+1 / b)^{-1}$, and together with $\Delta E / \Delta R \rightarrow\left|\partial E_{\lambda}(W, R) / \partial R\right|$, the approximations above yield (26. Now, if 25) does not hold, then as was noted above, either $E_{\lambda}(W, R)=$ $0.5 E_{\lambda-1}\left(W^{-}, \max \left(0^{+}, 2 R-1\right)\right)$ or $E_{\lambda}(W, R)=0.5 E_{\lambda-1}\left(W^{+}, 2 R-\min (1,2 R)\right)$. In both cases $E_{\lambda}(W, R)$ is convex. Finally, the claim that for $R>0,\left|\partial E_{\lambda}(W, R) / \partial R\right|<\infty$ is due to the fact that $E_{\lambda}(W, R)$ is a finite, convex, decreasing function for all $R>0$.

## Appendix B

## Proof of Lemma 5

We first prove by induction on $\lambda$ that $V_{\lambda}(W)=V(W)+o(\delta(W))$. This claim is trivial for $\lambda=0$. Now assume it holds for $\lambda-1$. By the induction assumption,

$$
V_{\lambda-1}(W)=V(W)+o(\delta(W))=\alpha \delta(W)+o(\delta(W))
$$

Now, $\epsilon_{l}()$ and $\epsilon_{h}()$, defined in Section II, satisfy

$$
\epsilon_{l}(I(W))=\delta(W)+o(\delta(W)) \quad, \quad \epsilon_{h}(I(W))=\delta(W)-\delta(W)^{2}
$$

(the claim on $\epsilon_{l}(I(W))$ follows by a first order Taylor expansion of $\epsilon_{l}(1-\delta)$ for small $\delta$ ). Hence, recalling Section II.

$$
I(W)-I\left(W^{-}\right)=I\left(W^{+}\right)-I(W)=\delta(W)+o(\delta(W))
$$

so that

$$
I\left(W^{-}\right)=1-2 \delta(W)+o(\delta(W)) \quad, \quad I\left(W^{+}\right)=1+o(\delta(W))
$$

By the assumption of the theorem regarding $W^{-}$and $W^{+}$,

$$
V\left(W^{-}\right)=\alpha \delta\left(W^{-}\right)+o\left(\delta\left(W^{-}\right)\right)=2 \alpha \delta(W)+o(\delta(W)) \quad, \quad V\left(W^{+}\right)=\alpha \delta\left(W^{+}\right)+o\left(\delta\left(W^{+}\right)\right)=o(\delta(W))
$$

Applying the induction assumption to $W^{-}$and $W^{+}$, we know that

$$
V_{\lambda-1}\left(W^{-}\right)=V\left(W^{-}\right)+o\left(\delta\left(W^{-}\right)\right)=2 \alpha \delta(W)+o(\delta(W)), V_{\lambda-1}\left(W^{+}\right)=V\left(W^{+}\right)+o\left(\delta\left(W^{+}\right)\right)=o(\delta(W))
$$

By Lemma 4 ,

$$
V_{\lambda}(W)=0.5(\sqrt{2 \alpha \delta(W)+o(\delta(W))}+\sqrt{o(\delta(W))})^{2}=\alpha \delta(W)+o(\delta(W))=V(W)+o(\delta(W))
$$

This concludes the proof of the first part of the lemma.
We proceed by proving by induction on $\lambda$ that after $2^{\lambda}$ polarization steps of the channel $W$,

$$
\begin{equation*}
I\left(W_{0}\right)=I\left(W^{-\cdots-}\right)=1-2^{\lambda} \delta(W)+o(\delta(W)) \tag{27}
\end{equation*}
$$

and

$$
\begin{equation*}
I\left(W_{i}\right)=1+o(\delta(W)) \quad i \neq 0 \tag{28}
\end{equation*}
$$

This claim is trivial for $\lambda=0$. Assuming it holds for $\lambda-1$, we note that if $I\left(W_{i}\right)=1+o(\delta(W))$, then, using essentially the same arguments that were used in the beginning of the proof

$$
I\left(W_{i}\right)-I\left(W_{i}^{-}\right)=I\left(W_{i}^{+}\right)-I\left(W_{i}\right)=o(\delta(W)) \quad i \neq 0
$$

Therefore, for $i \neq 0$,

$$
I\left(W_{i}^{-}\right)=1+o(\delta(W)) \quad, \quad I\left(W_{i}^{+}\right)=1+o(\delta(W))
$$

In addition, by the induction assumption,

$$
I\left(W_{0}\right)-I\left(W_{0}^{-}\right)=I\left(W_{0}^{+}\right)-I\left(W_{0}\right)=2^{\lambda-1} \delta(W)+o(\delta(W))
$$

Hence, reapplying the induction assumption,

$$
I\left(W_{0}^{-}\right)=1-2^{\lambda} \delta(W)+o(\delta(W)) \quad, \quad I\left(W_{0}^{+}\right)=1+o(\delta(W))
$$

This concludes the proof of (27) and 28). Thus, after $\lambda$ polarization steps, $V\left(W_{0}\right)=2^{\lambda} \alpha \delta(W)+o(\delta(W))$ and $V\left(W_{i}\right)=$ $o(\delta(W))$ for $i \neq 0$. We conclude that

$$
\lim _{I(W) \rightarrow 1} I\left(W_{0}\right) \sqrt{\frac{V_{\lambda}(W)}{2^{\lambda} V\left(W_{0}\right)}}=\lim _{\delta(W) \rightarrow 0}\left[1-2^{\lambda} \delta(W)+o(\delta(W))\right] \sqrt{\frac{\alpha \delta(W)+o(\delta(W))}{2^{2 \lambda} \alpha \delta(W)+o(\delta(W))}}=1 / 2^{\lambda}
$$

and for $i \neq 0$,

$$
\lim _{I(W) \rightarrow 1} I\left(W_{i}\right) \sqrt{\frac{V_{\lambda}(W)}{2^{\lambda} V\left(W_{i}\right)}}=\lim _{\delta(W) \rightarrow 0}[1+o(\delta(W))] \sqrt{\frac{\alpha \delta(W)+o(\delta(W))}{2^{\lambda} o(\delta(W))}}=\infty
$$

Therefore,

$$
\lim _{I(W) \rightarrow 1} \min _{i}\left(I\left(W_{i}\right) \sqrt{\frac{V_{\lambda}(W)}{2^{\lambda} V\left(W_{i}\right)}}\right)=1 / 2^{\lambda}
$$

Thus, by (19), $\lim _{I(W) \rightarrow 1} R_{\min }=1-1 / 2^{\lambda}$.

## Appendix C

## Proof of Lemma 6

For shortness of notation, define $x \triangleq I(W)-R$. Also denote $\hat{V}=\hat{V}_{\lambda}(N, W)$ and $V=V_{\lambda}(W)$. First, we will find a sufficient condition for

$$
\begin{equation*}
2^{\lambda} Q\left(x \sqrt{\frac{N}{V}}\right) \leq Q\left(x \sqrt{\frac{N}{\hat{V}}}\right) \tag{29}
\end{equation*}
$$

Recall that for $u>0$ the complementary Gaussian cumulative distribution function, $Q()$, satisfies,

$$
\begin{equation*}
\frac{1}{\sqrt{2 \pi} u}\left(1-\frac{1}{u^{2}}\right) e^{-u^{2} / 2} \leq Q(u) \leq \frac{1}{\sqrt{2 \pi} u} e^{-u^{2} / 2} \tag{30}
\end{equation*}
$$

The left hand side (LHS) of 30) implies that

$$
\frac{\sqrt{\hat{V}}}{x \sqrt{2 \pi N}}\left(1-\frac{\hat{V}}{x^{2} N}\right) e^{-\frac{x^{2} N}{2 \hat{V}}} \leq Q\left(x \sqrt{\frac{N}{\hat{V}}}\right)
$$

while the RHS of (30) implies that

$$
2^{\lambda} Q\left(x \sqrt{\frac{N}{V}}\right) \leq \frac{2^{\lambda} \sqrt{V}}{x \sqrt{2 \pi N}} e^{-\frac{x^{2} N}{2 V}}
$$

Therefore,

$$
\begin{equation*}
\sqrt{\hat{V}}\left(1-\frac{\hat{V}}{x^{2} N}\right) e^{-\frac{x^{2} N}{2 \hat{V}}} \geq 2^{\lambda} \sqrt{V} e^{-\frac{x^{2} N}{2 V}} \tag{31}
\end{equation*}
$$

is a sufficient condition for (29). But (31) is equivalent to

$$
\begin{equation*}
\frac{\ln \hat{V}}{2}+\ln \left(1-\frac{\hat{V}}{x^{2} N}\right)-\frac{x^{2} N}{2 \hat{V}} \geq \lambda \ln 2+\frac{\ln V}{2}-\frac{x^{2} N}{2 V} \tag{32}
\end{equation*}
$$

Since

$$
\ln \left(1-\frac{\hat{V}}{x^{2} N}\right)=-\frac{\hat{V}}{x^{2} N}+\Theta\left(\frac{1}{N^{2}}\right)
$$

the inequality (32) can be written as

$$
\frac{\ln \hat{V}}{2 N}-\frac{\hat{V}}{x^{2} N^{2}}+\Theta\left(\frac{1}{N^{3}}\right)-\frac{x^{2}}{2 \hat{V}} \geq \frac{\lambda \ln 2}{N}+\frac{\ln V}{2 N}-\frac{x^{2}}{2 V}
$$

That is,

$$
\frac{1}{V}-\frac{1}{\hat{V}} \geq \Theta\left(\frac{1}{N}\right)
$$

which leads to

$$
\hat{V}-V \geq \Theta\left(\frac{1}{N}\right)
$$

Now, we will find a sufficient condition for

$$
\begin{equation*}
2^{\lambda} Q\left(x \sqrt{\frac{N}{V}}\right) \geq Q\left(x \sqrt{\frac{N}{\hat{V}}}\right) \tag{33}
\end{equation*}
$$

The LHS of (30) implies that

$$
2^{\lambda} Q\left(x \sqrt{\frac{N}{V}}\right) \geq \frac{2^{\lambda} \sqrt{V}}{x \sqrt{2 \pi N}}\left(1-\frac{V}{x^{2} N}\right) e^{-\frac{x^{2} N}{2 V}}
$$

while the RHS of (30) implies that

$$
Q\left(x \sqrt{\frac{N}{\hat{V}}}\right) \leq \frac{\sqrt{\hat{V}}}{x \sqrt{2 \pi N}} e^{-\frac{x^{2} N}{2 \hat{V}}}
$$

Therefore,

$$
\frac{\sqrt{\hat{V}}}{x \sqrt{2 \pi N}} e^{-\frac{x^{2} N}{2 \dot{V}}} \leq \frac{2^{\lambda} \sqrt{V}}{x \sqrt{2 \pi N}}\left(1-\frac{V}{x^{2} N}\right) e^{-\frac{x^{2} N}{2 V}}
$$

is a sufficient condition for (33). This condition is equivalent to

$$
\frac{\ln \hat{V}}{2}-\frac{x^{2} N}{2 \hat{V}} \leq \frac{\ln V}{2}+\lambda \ln 2+\ln \left(1-\frac{V}{x^{2} N}\right)-\frac{x^{2} N}{2 V}
$$

which means

$$
\frac{\ln \hat{V}}{2 N}-\frac{x^{2}}{2 \hat{V}} \leq \frac{\ln V}{2 N}+\frac{\lambda \ln 2}{N}-\frac{V}{x^{2} N^{2}}+\frac{V^{2}}{x^{4}} \Theta\left(\frac{1}{N^{3}}\right)-\frac{x^{2}}{2 V}
$$

Therefore,

$$
\frac{1}{V}-\frac{1}{\hat{V}} \leq \Theta\left(\frac{1}{N}\right)
$$

and

$$
\hat{V}-V \leq \Theta\left(\frac{1}{N}\right)
$$

Summarizing, for some constants $\kappa_{1}$ and $\kappa_{2}$ (both independent of $N$ ) and $N$ sufficiently large, we have shown that if $\hat{V}_{\lambda}(N, W)-V_{\lambda}(W) \geq \kappa_{1} / N\left(\hat{V}_{\lambda}(N, W)-V_{\lambda}(W) \leq \kappa_{2} / N\right.$, respectively) then the LHS of 22) is smaller (larger) than the RHS. In addition, note that the RHS of (22), that defines $\hat{V}_{\lambda}(N, W)$, is monotonically increasing in $\hat{V}_{\lambda}(N, W)$. This proves our claim.

The proof for $V_{n, \lambda}(W)$ and $\hat{V}_{n, \lambda}(N, W)$ is identical to the proof for $V_{\lambda}(W)$ and $\hat{V}_{\lambda}(N, W)$, and is therefore omitted.

## REFERENCES

[1] E. Arikan, "Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," IEEE Transactions on Information Theory, vol. 55, no. 7, pp. 3051-3073, 2009.
[2] I. Tal and A. Vardy, "List decoding of polar codes," IEEE Transactions on Information Theory, vol. 61, no. 5, pp. 2213-2226, 2015.
[3] C. Leroux, A. J. Raymond, G. Sarkis, and W. J. Gross, "A semi-parallel successive-cancellation decoder for polar codes," IEEE Transactions on Signal Processing, vol. 61, no. 2, pp. 289-299, 2013.
[4] B. Li, H. Shen, D. Tse, and W. Tong, "Low-latency polar codes via hybrid decoding," in Proc. 8th Int. Symp. Turbo Codes and Iterative Inf. Processing (ISTC), August 2014, pp. 223-227.
[5] B. Yuan and K. Parhi, "Low-latency successive-cancellation list decoders for polar codes with multibit decision," IEEE Trans. Very Large Scale Integr. (VLSI) Syst., vol. 23, no. 10, pp. 2268-2280, 2015.
[6] C. Xiong, J. Lin, and Z. Yan, "Symbol-decision successive cancellation list decoder for polar codes," IEEE Transactions on Signal Processing, vol. 64, no. 3, pp. 675-687, February 2016.
[7] H. Mahdavifar, M. El-Khamy, J. Lee, and I. Kang, "Performance limits and practical decoding of interleaved reed-solomon polar concatenated codes," Communications, IEEE Transactions on, vol. 62, no. 5, pp. 1406-1417, May 2014.
[8] P. Trifonov and P. Semenov, "Generalized concatenated codes based on polar codes," in Proc. 8th International Symposium on Wireless Communication Systems (ISWCS), Nov 2011, pp. 442-446.
[9] Y. Wang, K. R. Narayanan, and Y. C. Huang, "Interleaved concatenations of polar codes with BCH and convolutional codes," IEEE Journal on Selected Areas in Communications, vol. 34, no. 2, pp. 267-277, Feb 2016.
[10] D. Goldin and D. Burshtein, "Block successive cancellation decoding of polarization-based codes," in Proc. 9th Int. Symp. Turbo Codes and Iterative Inf. Processing (ISTC), September 2016, pp. 256-260.
[11] M. P. Fossorier and S. Lin, "Soft-decision decoding of linear block codes based on ordered statistics," IEEE Transactions on Information Theory, vol. 41, no. 5, pp. 1379-1396, 1995.
[12] A. Valembois and M. Fossorier, "Box and match techniques applied to soft-decision decoding," IEEE Transactions on Information Theory, vol. 50, no. 5, pp. 796-810, 2004.
[13] E. Nachmani, Y. Be'ery, and D. Burshtein, "Learning to decode linear codes using deep learning," in Communication, Control, and Computing (Allerton), 2016 54th Annual Allerton Conference on. IEEE, 2016, pp. 341-346.
[14] E. Nachmani, E. Marciano, L. Lugosch, W. J. Gross, D. Burshtein, and Y. Beery, "Deep learning methods for improved decoding of linear codes," arXiv preprint arXiv:1706.07043, 2017.
[15] T. Gruber, S. Cammerer, J. Hoydis, and S. t. Brink, "On deep learning-based channel decoding," accepted for CISS 2017, arXiv preprint arXiv:1701.07738, 2017.
[16] S. Cammerer, T. Gruber, J. Hoydis, and S. t. Brink, "Scaling deep learning-based decoding of polar codes via partitioning," arXiv preprint arXiv:1702.06901, 2017.
[17] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," IEEE Transactions on Information Theory, vol. 56, no. 5, pp. 2307-2359, May 2010.
[18] S. B. Korada, "Polar codes for channel and source coding," Ph.D. dissertation, EPFL, Lausanne, Switzerland, 2009.
[19] R. Mori and T. Tanaka, "Performance and construction of polar codes on symmetric binary-input memoryless channels," in Proc. IEEE International Symposium on Information Theory (ISIT), Seoul, Korea, June 2009, pp. 1496 - 1500.
[20] T. Richardson and R. Urbanke, Modern Coding Theory. Cambridge, UK: Cambridge University Press, 2008.
[21] R. G. Gallager, Information Theory and Reliable Communication. New York: Wiley, 1968.
[22] A. Barg and G. D. Forney, "Random codes: minimum distances and error exponents," IEEE Transactions on Information Theory, vol. 48, no. 9, pp. 2568-2573, Sep 2002.
[23] M. Alsan and E. Telatar, "Polarization improves $E_{0}$," IEEE Transactions on Information Theory, vol. 60, no. 5, pp. 2714-2719, May 2014.
[24] A. G. i Fabregas, I. Land, and A. Martinez, "Extremes of error exponents," IEEE Transactions on Information Theory, vol. 59, no. 4, pp. 2201-2207, April 2013.
[25] T. Erseghe, "Coding in the finite-blocklength regime: bounds based on Laplace integrals and their asymptotic approximations," IEEE Transactions on Information Theory, vol. 62, no. 12, pp. 6854-6883, Dec 2016.
[26] C. E. Shannon, "Probability of error for optimal codes in a gaussian channel," The Bell System Technical Journal, vol. 38, no. 3, pp. 611-656, May 1959.
[27] R. H. Morelos-Zaragoza, The Art of Error Correcting Coding, 2nd ed. New York: Wiley, 2006, web site: http://www.the-art-of-ecc.com
[28] G. Liva, L. Gaudio, T. Ninacs, and T. Jerkovits, "Code design for short blocks: a survey," arXiv preprint arXiv:1610.00873, 2016.
[29] J. V. Wonterghem, A. Alloumf, J. J. Boutros, and M. Moeneclaey, "Performance comparison of short-length error-correcting codes," in 2016 Symposium on Communications and Vehicular Technologies (SCVT), Nov 2016, pp. 1-6.


[^0]:    This research was supported by the Israel Science Foundation, grant no. 1082/13.
    D. Goldin is with the school of Electrical Engineering, Tel-Aviv University, Tel-Aviv 6997801, Israel (email: dinagold@post.tau.ac.il).
    D. Burshtein is with the school of Electrical Engineering, Tel-Aviv University, Tel-Aviv 6997801, Israel (email: burstyn@eng.tau.ac.il).

    This material in this paper was presented in part in ISIT 2017.

[^1]:    ${ }^{1}$ The assumption that the channel is discrete is made for notational convenience only. For continuous output channels, sums should be replaced by integrals.
    ${ }^{2}$ The default basis of logarithms in this paper is 2

[^2]:    ${ }^{3}$ By [21 Section 5.6] $E_{r}(W, R)=\max _{0 \leq \rho \leq 1} E_{0}(W, \rho)-\rho R$, where the rate $R$ is defined using natural logarithms, and measured in nats per channel use [21 Paragraph after Equation (5.1.2)]. In order to use the more widespread notation, by which $R$ is the number of bits per channel use, we substitute $R$ with $R \ln 2$. In order to further improve the bound on $P_{e}$, we define $E_{r}(W, 0)=\infty$ instead of the actual error exponent, since there cannot be errors if no information bits were transmitted.

[^3]:    ${ }^{4} o(\delta)$ denotes a term which is negligible in its absolute value compared to $\delta$ for sufficiently small $\delta$

