Multi-Kernel Polar Codes: Proof of Polarization and Error Exponents

Meryem Benammar, Valerio Bioglio, Frédéric Gabry, Ingmar Land
Mathematical and Algorithmic Sciences Lab
Paris Research Center, Huawei Technologies France SASU
Email: {meryem.benammar,valerio.bioglio,frederic.gabry,ingmar.land}@huawei.com

Abstract—In this paper, we investigate a novel family of polar codes based on multi-kernel constructions, proving that this construction actually polarizes. To this end, we derive a new and more general proof of polarization, which gives sufficient conditions for kernels to polarize. Finally, we derive the convergence rate of the multi-kernel construction and relate it to the convergence rate of each of the constituent kernels.

I. INTRODUCTION

Channel polarization is a novel technique to create capacity-achieving codes over various channels [1]. In its original construction, a polar code is generated by a sub-matrix of the transformation matrix $T_2^\otimes n$ with the binary kernel defined by $T_2 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$. While the polarization of the Kronecker powers of the $T_2$ kernel is presented in [1], the proof has been generalized in [2] for the Kronecker power of larger binary kernels. Hereafter, various kernels have been proposed, along with their convergence rates [3].

Recently, a novel family of polar codes, where binary kernels of different sizes are mixed, has been proposed in [4]. These codes, coined as multi-kernel polar codes, make it possible to construct polar codes of any block lengths, not limited to powers of integers as when using a single kernel. In [4], authors conjecture that the proposed mixed construction polarizes; this conjecture is confirmed by the density evolution algorithm used to calculate the bit reliabilities [5], however, a rigorous convergence proof is missing.

In this paper, we fill the gap left in [4] by presenting a proof of the polarization of multi-kernel polar codes. To this end, we describe the polarization of any kernel through an inequality related to the reliability transitions in every step of the polarization. This inequality is different from the ones presented in other proofs [1, 2], however, we will show that this inequality holds for Arikan’s original kernel, along with other selected kernels. Besides, we derive the convergence rate of the obtained multi-kernel polar code based on convergence rate of each of the constituent kernels.

This paper is organized as follows. Section II introduces the information theoretic model and reviews the code construction of multi-kernel polar codes. Section III presents the main result of this paper which consists in the proof of polarization, while Section IV derives the convergence rates (or error exponents) of the resulting multi-kernel polar code.

II. SYSTEM MODEL AND CODE DEFINITION

In this section, we introduce the fundamental definitions related to polar coding and the underlying information theoretic model. Moreover, we briefly present the multi-kernel construction of polar codes.

A. Channel model

Let $W : \mathcal{X} \to \mathcal{Y}$ be a discrete input channel defined by its associated probability mass function (pmf) $W(y|x)$, and let $\mathcal{Y}^N : \mathcal{X}^N \to \mathcal{Y}^N$ be its $N$-th memoryless extension with associated pmf $W(N)(y^N_1|x^N_1)$,

$$W(N)(y^N_1|x^N_1) = \prod_{i=1}^N W(y_i|x_i).$$

Let $U^N_i \triangleq (U_1, \ldots, U_N)$ be $N$ auxiliary random variables satisfying the Markov chain $(U_1, \ldots, U_N) \leftrightarrow X_1 \leftrightarrow Y_i$ for all $i \in [1 : N]$, where $[1 : N]$ represents the set of integers from 1 to $N$. In the following, the input alphabet $\mathcal{X}$ and the respective auxiliary alphabets $U_i$ are all binary, i.e., for all $i \in [1 : N]$, $\mathcal{X} = U_i = \{0, 1\}$. The auxiliary variables, or bit components, $U_i$ are all pairwise independent $\text{Bern}(\frac{1}{2})$ variables.

B. Channel polarization

A polar code of length $N$ is a linear block code which maps the bits $u^N_1 = (u_1, \ldots, u_N)$ into the channel input array $x^N_1 = (x_1, \ldots, x_N)$ through a linear invertible mapping, i.e.,

$$x^N_1 = u^N_1 \cdot G_N,$$

where $G_N$ is depicted as the transformation matrix of the polar code. Since the bits $(U_1, \ldots, U_N)$ are independent $\text{Bern}(\frac{1}{2})$ distributed and $G_N$ is invertible, then the channel inputs $(X_1, \ldots, X_N)$ are also independent $\text{Bern}(\frac{1}{2})$ distributed. This yields to the information conservation principle

$$I(U^N_1; Y^N_1) = I(X_1^N; Y_1^N) = N \cdot I(X; Y),$$

which stems from that $W(N)$ is memoryless. In the following, since the input distribution $P_X$ is fixed, we will use $I(W) \triangleq I(X; Y)$ to denote the dependence of $I(X; Y)$ only on the channel $W$. 

These results were derived when F. Gabry was still with the Mathematical and Algorithmic Sciences Lab of Huawei.
To introduce the polarization principle, let us first use the independence of the bits $U_i$ to write the following

$$I(U_i^N; Y_1^N) = \sum_{i=1}^N I(U_i; Y_i^N | U_i^{i-1})$$  \hspace{1cm} (4)

$$= \sum_{i=1}^N I(U_i; Y_i^N, U_i^{i-1}).$$  \hspace{1cm} (5)

Let us then define the channel $W_i^{(N)}$, with output $Z_i \triangleq (Y_i^N, U_i^{i-1})$ and associated pmf

$$W_i^{(N)}(z_i | u_i) = W_i^{(N)}(y_i^N, u_i^{i-1} | u_i).$$  \hspace{1cm} (6)

To polarize the channel $W_i^{(N)}$ means to create $N$ virtual channels $W_i^{(N)}$, each being either a degraded version of $W$ or an enhanced version of it. By a proper choice of the transformation matrix $G_N$, as the one suggested by Arikan [1], it can be shown that, as the code length $N$ goes to infinity, the resulting virtual channels $W_i^{(N)}$ are either perfectly reliable channels or totally noisy channels, i.e.,

$$\forall i \in [1 : N] \quad I(U_i; Y_i^N, U_i^{i-1}) \to 0 \text{ or } 1.$$  \hspace{1cm} (7)

Arikan showed in [1] that the fraction of perfectly reliable channels among all, i.e., the fraction of bits that can be transmitted reliably, is given by the mutual information $I(X ; Y)$.

### C. Multi-kernel code construction

The transformation matrix $G_N$ of a polar code of length $N$ is defined by the $n$-fold Kronecker product of the binary kernel

$$T_2 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix},$$

namely $G_N = T_2^\otimes N$, where $N = 2^n$. Note that the structure of the transformation matrix renders it invertible, and that the admissible blocklengths $N$ are all powers of 2, which might be an impediment for practical applications with arbitrary blocklengths. However, in [2] authors prove that the kernel $T_3$ can be replaced by any polarizing kernel $T_i$ with dimension $l \times l$, leading to codes of blocklengths of the form $N = l^n$. An example of such kernels, which we will resort to in the document, is the $T_3$ kernel given by:

$$T_3 \triangleq \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$  \hspace{1cm} (8)

The multi-kernel polar code construction is introduced in [4], in which multiple binary kernels are used in the construction of the code. Consider to this end a collection of kernels $T_{i_1}, \ldots, T_{i_m}$ where, for $j \in [1 : m]$, $T_{i_j}$ is a binary matrix of dimension $l_j \times l_j$. A multi-kernel transformation matrix is constructed as the Kronecker product of these kernels,

$$G_N = T_{i_1} \otimes T_{i_2} \otimes \cdots \otimes T_{i_m}. $$  \hspace{1cm} (9)

Note here that the size of the resulting code is $N = \prod_{j=1}^m l_j$. An $(N, K)$ multi-kernel polar code is defined by the transformation matrix $G_N$ in (9), and the information set $\mathcal{I}$ of size $|\mathcal{I}| = K$, or conversely by the frozen set $\mathcal{F} = [1 : N] \setminus \mathcal{I}$. For the encoding, each frozen bit is set to zero, i.e., $u_j = 0$ for $i \in \mathcal{F}$, while information is stored in the remaining bits, whose indices constitute the information set $\mathcal{I}$. Then, the channel input $x_i^N$ is obtained by $x_i^N = u_i^N \cdot \mathbb{G}_N$. In [4], authors conjecture that the Kronecker product of polarizing kernels result in a polarizing transformation matrix, and they calculate the reliabilities of the bits through density evolution [5]. The information set $\mathcal{I}$ is then constituted by the $K$ bit positions with the highest reliability. In this paper, we present a prove of this conjecture, confirming the goodness of the multi-kernel construction in [4].

Decoding of multi-kernel polar codes is performed similarly to Arikan’s polar code, using successive cancellation decoding. At each step, a bit $u_i$ is decoded from the channel outputs $y_i^N$ using the previously made hard decisions on the bits $u_i^{i-1}$. In practice, the decoding is performed on the Tanner graph of the code, where each block of the decoder consists in the basic decoding operations of the kernel $T_j$ as explained in [4].

### III. POLARIZATION OF MULTI-KERNEL POLAR CODES

In this section, we prove the main result of the paper: the polarization of multi-kernel polar codes. The approach we adopt here is somewhat similar to the one proposed by Arikan in [1], but differing in the last steps, where we prove that polarization is highly kernel dependent.

#### A. Definitions

In the following, the sub-channels $W_i^{(N)}$ where $i \in [1 : N]$ are defined as

$$W_i^{(N)} \triangleq W_{b_1, \ldots, b_m} $$  \hspace{1cm} (10)

where the index $b_j$, for $j \in [1 : m]$, takes values in $[1 : l_j]$, constituting the mixed radix decomposition of $i$ in the base $(l_1, \ldots, l_m)$. Due to the Kronecker construction of the transformation matrix $G_N$ in (9), the channel polarization occurs iteratively in that, at each step $m$, each of the previous channels $W_{b_1, \ldots, b_{m-1}}$ is polarized into $l_m$ new channels $W_{b_1, \ldots, b_{m-1}, 0}, W_{b_1, \ldots, b_{m-1}, 1}, \ldots, W_{b_1, \ldots, b_{m-1}, l_m}$ as shown in Figure [1].

Let us assume that the transitions on the polarization tree occur uniformly at random, i.e. the indices $(B_1, \ldots, B_m)$ are uniformly distributed, i.e., $\forall b_j \in [1 : l_j]$, $P_{B_j}(b_j) = \frac{1}{l_j}$. The mutual information $I(W_{B_1, \ldots, B_m})$ is then a random variable...
depending on the random variables \(B_1, \ldots, B_m\), which we will denote by
\[
I_m \triangleq I(W_{B_1, \ldots, B_m})
\]
where \(I_0 \triangleq I(W)\) is a deterministic variable indicating the channel capacity. Finally, we give a formal definition of channel polarization for multi-kernel polar codes.

**Definition 1.** A multi-kernel polar code is polarizing if
\[
\forall \epsilon > 0, \lim_{m \to \infty} P(I_m \geq 1 - \epsilon) = 1 - \lim_{m \to \infty} P(I_m \leq \epsilon) = I(W).
\]

**B. Proof of polarization**

Similarly to the case of \(T_2\) kernels investigated by Arikan, the proof of polarization of multi-kernel polar codes is two-fold. First, we prove that the random process \(\{I_m\}_m\) is converging to a random variable \(I_\infty\) in probability. Then, we show that, if the kernels are selected properly, the distribution of \(I_\infty\) is a Bernoulli distribution with probability \(I(W)\), and \(I_m\) follows the Definition 1.

**Lemma 1.** The sequence \(\{I_m\}_m\) is a bounded martingale with respect to \(B_1, \ldots, B_m\), and thus converges in probability to a variable \(I_\infty\) whose expected value verifies \(E(I_\infty) = I_0\).

**Proof.** To begin with, we have that \(0 \leq I_m \leq 1\) for all \(m \in \mathbb{N}\), which is due to the fact that the channel \(W_{B_1, \ldots, B_m}\) has binary inputs. Let \(m \in \mathbb{N}\), we have that
\[
\begin{align*}
E_{B_{m+1}}(I_{m+1}|B_1, \ldots, B_m) &= E_{B_{m+1}}(I(W_{B_1, \ldots, B_m,B_{m+1}})|B_1, \ldots, B_m) \\
&= \sum_{b=1}^{l_{m+1}} P(B_{m+1} = b) I(W_{B_1, \ldots, B_m,b}) \\
&= \frac{1}{l_{m+1}} \sum_{b=1}^{l_{m+1}} I(W_{B_1, \ldots, B_m,b}) \overset{(a)}{=} I(W_{B_1, \ldots, B_m})
\end{align*}
\]
where \((a)\) is a consequence the information conservation principle \((3)\) and of \((5)\). Finally, one can write that
\[
E_{B_1, \ldots, B_m,B_{m+1}}(I_{m+1}) = E_{B_1, \ldots, B_m}(E_{B_{m+1}}(I(W_{B_1, \ldots, B_m,B_{m+1}})|B_1, \ldots, B_m)) \overset{(a)}{=} E_{B_1, \ldots, B_m}(I_m) = I_m
\]
where \((a)\) is a consequence of \((15)\). Thus, \(\{E(I_m)\}_m\) is constant, and
\[
E(I_m) = E(I_{m-1}) = \cdots = E(I_0) = I_0.
\]

Thus, \(\{I_m\}_m\) is a bounded martingale, hence, uniformly integrable and thus convergent in probability to a random variable \(I_\infty\) such that \(E(I_\infty) = I_0\), which is due to \((16)\).

**Theorem 1.** If for every kernel \(T\) forming the transformation matrix of the code, and every past sequence \(\{b_1, \ldots, b_{m-1}\}\), there exist \(\alpha, \beta > 0\) such that the following inequality holds:
\[
\forall b_m \in \{1 : l_m\}, |I(W_{b_1, \ldots, b_{m-1}, b_m}) - I(W_{b_1, \ldots, b_{m-1}})| \geq I(W_{b_1, \ldots, b_{m-1}})^\alpha \cdot (1 - I(W_{b_1, \ldots, b_{m-1}}))^\beta
\]
then \(I_\infty\) is a Bernoulli distributed variable with
\[
P(I_\infty = 1) = 1 - P(I_\infty = 0) = I_0.
\]

**Proof.** Here we present a novel proof showing clearly the dependency of the polarization process on the choice of the kernel. Let \(m \in \mathbb{N}\) and let \((b_1, \ldots, b_{m-1})\) be the vector of past indices in the polarization tree. Under the assumptions of Theorem 1 we have that for all possible \(b_m \in \{1 : l_m\},
\[
|I(W_{b_1, \ldots, b_{m-1}, b_m}) - I(W_{b_1, \ldots, b_{m-1}})| \geq I(W_{b_1, \ldots, b_{m-1}})^\alpha \cdot (1 - I(W_{b_1, \ldots, b_{m-1}}))^\beta
\]
This implies that on the cylinder set \((B_1, \ldots, B_m)\), the following inequality holds with probability \(1\), i.e.,
\[
|I_m - I_{m-1}| \geq l_{m-1} \cdot (1 - I_{m-1})^\beta \geq 0.
\]
Hence, since \(\{I_m\}_m\) is a uniformly converging sequence, with limit \(I_\infty\), and since the function \(f : x \to x^\alpha(1 - x)^\beta\) is continuous, then
\[
0 \geq l_{\infty}^\beta \cdot (1 - I_{\infty})^\beta \geq 0,
\]
which in turn implies that \(I_\infty = 1\) or \(0\).

What we presented can be seen as an alternative to the original proof of polarization made by Arikan and then extended in [2] to arbitrary kernels. Even if this inequality may seem a bit restrictive, it is verified for a big family of kernels; in the following, we prove it for Arikan’s \(T_2\) binary kernel, and for the \(T_3\) kernel given in [3].

**C. Examples of kernels polarization**

In the following, we prove that the kernels \(T_2\) and \(T_3\) are polarizing according to the multi-kernel definition by showing that the constraints in \((17)\) are met for these two kernels.

For the kernel \(T_2\), consider a channel \(\mathcal{W}\) with binary inputs and finite outputs. Let \((U_1, U_2)\) be the Bernoulli auxiliary inputs and \((X_1, X_2)\) be the binary channel inputs where
\[
X_1^2 = U_1^2 \cdot T_2.
\]

The independence of \((X_1, X_2)\) implies that \((X_1, Y_1)\) and \((X_2, Y_2)\) are all pairwise independent and
\[
H(X_1|Y_1) = H(X_2|Y_2) = H(X|Y) \triangleq H_0 \leq 1.
\]

The inequality in \((17)\) implies the pair of constraints:
\[
\begin{align*}
|I(U_1; Y_1^2) - I(X; Y)| &\geq I(X; Y)^\alpha \cdot (1 - I(X; Y))^\beta \\
|I(U_2; Y_1^2 | U_1) - I(X; Y)| &\geq I(X; Y)^\alpha \cdot (1 - I(X; Y))^\beta
\end{align*}
\]
but since \(I(U_2; Y_1^2 | U_1) = 2I(X; Y) - I(U_1; Y_1^2)\), then it suffices to prove the first inequality, which amounts to
\[
H(X_1 \oplus X_2 | Y_1^2) - H_0 \geq H_0^\beta \cdot (1 - H_0)^\alpha.
\]

We prove in Appendix \(A\) that this holds with the choice \(\alpha = 1\) and \(\beta = 2\).
As for $T_3$, under similar assumptions on $(U_1, U_2, U_3)$ and the independence of $(X_1, Y_1)$, $(X_2, Y_2)$ and $(X_3, Y_3)$, and defining $H(X_1|Y_1) = H(X_2|Y_2) = H(X_3|Y_3) \triangleq H_0$, the condition in (17) amounts to proving that:

\[ I(U_1; Y_1^3) - I(X; Y) \geq I(X; Y)^\alpha(1 - I(X; Y))^\beta \]

\[ I(U_2; Y_2^3|U_1) - I(X; Y) \geq I(X; Y)^\alpha(1 - I(X; Y))^\beta \]

\[ I(U_3; Y_3^3|U_2^2) - I(X; Y) \geq I(X; Y)^\alpha(1 - I(X; Y))^\beta, \]

Again, it suffices to prove only the two first inequalities, which amounts to proving that:

\[ |H(X_1 \oplus X_2 \oplus X_3|Y_1^3) - H_0| \geq H_0^\alpha(1 - H_0)^\beta \]

\[ |H(X_2 \oplus X_3|Y_1^3, X_1 \oplus X_2 \oplus X_3) - H_0| \geq H_0^\beta(1 - H_0)^\alpha. \]

In Appendix A-B, we show that this holds with $\alpha = \beta = 2$.

IV. RATE OF CONVERGENCE OF MULTI-KERNEL POLAR CODES

The rate of convergence of the sequence $\{I_m\}_m$, which is related to the error exponent of the generator matrix $G_N = T_1 \otimes \cdots \otimes T_{m}$, is the asymptotic convergence rate of the probability of error. In this section, we show how to derive the rate of convergence of a multi-kernel polar code based on the rate of convergence of each of the constituent kernels $T_1, \ldots, T_{m}$.

A. Definitions

Let us first extend the definition of the Bhattacharyya parameter, a key measure in the rate of convergence of polar codes, to the case of multi-kernel polar codes. We recall that the Bhattacharyya parameter associated with a binary input channel $W$ is

\[ Z(W) \triangleq \sum_{y \in \mathcal{Y}} \sqrt{W(y|1)W(y|0)}. \]

Accordingly to the notation $W_i^{(N)} = W_{b_1,\ldots,b_m}$, where $(b_1, \ldots, b_m)$ are the mixed radix decomposition of $i$ in the basis $l_1, \ldots, l_m$, we define a random Bhattacharyya parameter associated to the random realization of $(B_1, \ldots, B_m)$ as follows:

Definition 2 (Bhattacharyya parameter). The random Bhattacharyya parameter associated to the random realization of $(B_1, \ldots, B_m)$ is given by

\[ Z_m \triangleq Z(W_{B_1,\ldots,B_m}) = \sum_{z \in Z} \sqrt{W_{B_1,\ldots,B_m}(z|1)W_{B_1,\ldots,B_m}(z|0)} \]

where $z$ is defined in (6).

The Bhattacharyya parameter is particularly useful since it yields bounds on the block error probability $P_e(N)$ of a length $N$ code, see [2].

Next, we define the rate of convergence of a multi-kernel polar code based on the convergence of the sequence $\{Z_m\}_m$.

Definition 3. A multi-kernel polar code has a rate of convergence $E_i$ if and only if the following properties hold:

1) For all $\gamma \geq E_i$,

\[ \lim_{m \to \infty} P(Z_m \geq 2^{-N^\gamma}) = 1; \]

2) for all $0 < \gamma \leq E_i$,

\[ \lim_{m \to \infty} P(Z_m \leq 2^{-N^\gamma}) = I(W). \]

The convergence rate relates directly to the error exponent, i.e. the rate of convergence of the block error probability to 0. As such, if a polar code has a convergence rate of $E_i$, then the block error probability satisfies

\[ P_e(N) \sim 2^{-N^E_i}. \]

For polar codes based on Arikan’s kernels, i.e. $G_N = T_{2^m}$, the convergence rate was shown in [6] to be equal to $E = 0.5$. For polar codes formed by larger kernels, authors in [2] derive the rate of convergence through the partial distances of the given kernel $T_i$ of length $l$ yielding, for instance, for kernel $T_3$ proposed in [5] a rate of convergence of $E = 0.42$.

B. Calculation of the rate of convergence

The rate of convergence of a polar code can be derived for Arikan’s $T_2$ kernels, and more generally, for arbitrary kernels $T_i$, through the following result.

Lemma 2. [5, Theorem 14]

Let $T_i$ be an arbitrary kernel with size $l \times l$, then the rate of convergence of the associated polar code is given by

\[ E_i \triangleq \frac{1}{l} \sum_{i=1}^{l} \log_2(D_i) \]

where $D_i$ is the partial distance of the $i$-th row of the transformation matrix $T_i = (t_{i1}, \ldots, t_{i1}, \ldots, t_{i1})$, defined as

\[ D_i \triangleq \text{dist}(t_{ij}, t_{ij+1}, \ldots, t_{ij}) \]

where $< t_{i+1}, \ldots, t_i >$ is the linear code spanned by the remaining rows of $T_i$ and $A^t$ the transpose of the matrix $A$.

In the following, we show that the rate of convergence of a multi-kernel polar code can be derived on the basis of the rate of convergence of each of the constituent kernels.

Theorem 2. Consider a multi-kernel polar code in which each of the $s$ distinct constituent kernels $T_{ij}$ has an error exponent $E_{ij}$ for $j \in [1 : s]$ and is used with frequency $p_{ij}$ in the Kronecker composition $G_N$ as $N \to \infty$. Then, the rate of convergence of the resulting mixed kernel polar code is given by

\[ E = \sum_{j=1}^{s} p_{ij} \log_2(l_{ij}) \cdot E_{ij} \]

Proof. Let $m$ be the index of the current iteration in the Kronecker product, let $i \in [1 : N]$ and let $(b_1, \ldots, b_m)$ be its corresponding mixed radix decomposition. The proof of this theorem follows from the following inequality, which is a result of ([2, Lemma 13]),

\[ Z(W_{b_1,\ldots,b_{m-1}})^{D_{b_m}} \leq Z(W_{b_1,\ldots,b_m}) \leq 2^{l_{m-b_m}} Z(W_{b_1,\ldots,b_{m-1}})^{D_{b_m}} \]
where \( D_{m\ast} \), with \( b_m \in [1 : l_m] \), is the partial distance of the row \( t_{b_m} \) in the kernel \( T_{m\ast} \). The proof that items 1) and 2) in definition \( \mathbb{3} \) holds, follows by analytic calculus and is presented in Appendix \( \mathbb{B} \).

As a result, the rate of convergence of a multi-kernel polar code consists in a weighted sum of the error exponents of each of the constituent kernels, where the weights are related to the frequency of occurrence of a kernel in the construction.

**APPENDIX A**

**PROOF OF POLARIZATION OF \( T_2 \) AND \( T_3 \)**

A. Proof of polarization of \( T_2 \)

To prove (22), note that

\[
H(X_1 \circ X_2 | Y_1^2) = \sum_{y_1, y_2} P_{Y_1, Y_2}(y_1, y_2) H(X_1 \circ X_2 | y_1, y_2) \tag{33}
\]

\[
= \sum_{y_1, y_2} P_{Y_1}(y_1) P_{Y_2}(y_2) H(X_1 \circ X_2 | y_1, y_2) \tag{34}
\]

\[
= \sum_{y_1, y_2} P_{Y_1}(y_1) P_{Y_2}(y_2) h_2(q_{y_1} \ast q_{y_2}) \tag{35}
\]

where (a) is a consequence from the independence of \( Y_1 \) and \( Y_2 \), and (b) follows since, conditioned on \( (y_1, y_2) \), \( X_1 \circ X_2 \) is a binary variable with probability

\[
P(X_1 \circ X_2 = 0 | y_1, y_2) = P(X_1 = 0 | y_1, y_2) \ast P(X_2 = 0 | y_1, y_2).
\]

Then, resolting to Mrs Gerber’s Lemma (7), i.e., convexity of \( h_2(p \ast h_2^{-1}(x)) \) in \( x \), we can write

\[
H(X_1 \circ X_2 | Y_1^2) \geq \sum_{y_1} P_{Y_1}(y_1) h_2(q_{y_1} \ast h_2^{-1}(H_0)) \tag{36}
\]

\[
\geq h_2 \left( h_2^{-1}(H_0) \ast h_2^{-1}(H_0) \right) \tag{37}
\]

Finally, it can be proved that, \( \forall a \in [0 : 1/2] \),

\[
h_2(a \ast a) - h_2(a) \geq h_2^2(a) \cdot (1 - h_2(a)) \geq 0. \tag{38}
\]

which, when replacing \( a = h_2^{-1}(H_0) \), yields the desired result.

B. Proof of polarization of \( T_3 \)

To prove (23), we first note that similar to (37),

\[
H(X_1 \circ X_2 \circ X_3 | Y_1^3) \geq h_2 \left( h_2^{-1}(H_0) \ast h_2^{-1}(H_0) \ast h_2^{-1}(H_0) \right) \tag{39}
\]

\[
\geq h_0 + h_2 \left( h_2^{-1}(H_0) \ast h_2^{-1}(H_0) \right) \cdot (1 - H_0) \tag{40}
\]

\[
= H_0 + H_0^2 \cdot (1 - H_0) \geq H_0 + H_0^2 \cdot (1 - H_0)^2 \tag{41}
\]

where, (a) and (b) are consequences of (38).

Next, to prove (24), we write the following:

\[
H(X_2 \circ X_3 | Y_1^3) = H(X_2 \circ X_3 | Y_1^3) - H(X_1 \circ X_2 \circ X_3 | Y_1^3)
\]

Next, we note that we can write the following upper bound on \( H(X_2 \circ X_3 | Y_1^3) \).

\[
H(X_2 \circ X_3 | Y_1^3) = \sum_{y_2, y_3} P_{Y_2}(y_2) P_{Y_3}(y_3) h_2(q_{y_2} \ast q_{y_3}) \tag{42}
\]

\[
\leq H(X_2 | Y_2) + (1 - H(X_3 | Y_3)) \cdot H(X_2 | Y_2) \tag{43}
\]

\[
= H_0 + H_0 \cdot (1 - H_0) \tag{44}
\]

where (a) is due to a consequence of Mrs Gerber’s lemma (7).

Thus, we can finally write that

\[
H(X_2 \circ X_3 | Y_1^3) = H(X_2 \circ X_3 | Y_1^3) - H_0 \leq H_0 \cdot (1 - H_0) \leq 0
\]

where (a) follows similarly from (37) and (38) by leaving \( X_2 \circ X_3 \) grouped.

**APPENDIX B**

**ERROR EXPONENTS: PROOF OF THEOREM 2**

In order to prove that the error exponent of a multi-kernel polar code is given by

\[
E = \sum_{j=1}^s P_{t_j} \log_2(l_j) \cdot E_{t_j} \tag{46}
\]

we would need to prove that \( E \) satisfies both conditions in definition \( \mathbb{3} \) namely

- Condition 1): For all \( \gamma \geq E \),

\[
\lim_{m \to \infty} \mathbb{P}(Z_m \geq 2^{-N^\gamma}) = 1; \tag{47}
\]

- Condition 2) for all \( 0 < \gamma \leq E \)

\[
\lim_{m \to \infty} \mathbb{P}(Z_m \leq 2^{-N^\gamma}) = I(W). \tag{48}
\]

Before proving that both these conditions hold, we list some preliminaries which are essential to the proof.
A. Preliminaries

Let \( m \) be the current iteration in the multi-kernel construction given by \( G_N = \bigotimes_{k=1}^{m} T_k \) where the kernels \( T_k \) can take values in a set of distinct kernels \( \{T_1, \ldots, T_s\} \) with a probability \( p_j \) for \( j \in [1:s] \). Let \( b_1, \ldots, b_m \) the realization of the random variables \( B_1, \ldots, B_m \) up to this iteration. We will use the shorthand notation \( Z_m \) to denote the realization of the random Bhattacharyya parameter \( Z(W_{b_1, \ldots, b_m}) \) defined in Definition 2. Let \( D_{b_1, \ldots, b_m} \) denote the corresponding partial distances and \( N = \prod_{k=1}^{m} l_j \) be the blocklength at iteration \( m \).

**Definition 4.** We define the occurrences set of a kernel \( T_i \) in \( G_N \) as

\[
\mathcal{N}_j \triangleq \{ k \in [1:m], T_k = T_i \}
\]

and the occurrences set of an index \( i \in [1:l_j] \) in the kernel \( T_j \) as

\[
\mathcal{N}^{m}_{i,j} \triangleq \{ k \in \mathcal{N}_j, b_k = b_i \}
\]

with cardinalities

\[
n_j = |\mathcal{N}_j| \quad \text{(51)}
\]

\[
n_{i,j} = |\mathcal{N}^{m}_{i,j}| \quad \text{(52)}
\]

By the law of large numbers, and since the variables \( (B_1, \ldots, B_m) \) are each uniformly distributed, the cardinalities of these sets verify

\[
\lim_{m \to \infty} \frac{n_j}{m} = p_j, \quad \lim_{m \to \infty} \frac{n_{i,j}}{m} = \frac{1}{l_j}
\]

\[
\implies \lim_{m \to \infty} \frac{n_{i,j}}{m} = \frac{p_j}{l_j}
\]

As such,\( \forall \epsilon > 0, \exists M \geq 0, \text{ s.t. } \left| \frac{n_{i,j}}{m} - \frac{p_j}{l_j} \right| \leq \epsilon \) for all \( m \geq M \). (55)

In the following, we list some key properties verified by the sequence of Bhattacharyya parameters \( (Z_m)_{m \geq 0} \).

**Lemma 3.** There exists \( K \geq 1 \), such that, for all \( m \geq 1 \),

\[
Z_{m-1} \leq Z_m \leq KZ_{m-1}
\]

**Proof.** This inequality follows from simple analytic manipulations of the result of ([2 Lemma 13]),

\[
Z_{m-1} \leq Z_m \leq 2^{l\cdot(Z_{m-1})} Z_{m-1} 
\]

\[
\leq 2^{l^*} Z_{m-1},
\]

where

\[
l^* \triangleq \max_{j \in [1:s]} l_j
\]

(59)

The results follows then by defining \( K \triangleq 2^{l^*} \).

**Lemma 4.** For any polarizing kernel \( T_i \), the partial distances \( D_i \) for \( i \in [1:l_j] \), satisfy the following

1) All partial distances are greater than 1 and bounded up by \( l_j \) \( \forall i \in [1:l_j] \), \( 1 \leq D_i \leq l_j \);

2) there exists at least a partial distance \( D_{i,j} \), of a row \( i_j \) of the kernel’s matrix \( T_{i,j} \), such that \( D_{i,j} \geq 2 \).

**Proof.** The first claim follows from the invertibility of a polarizing kernel matrix and the definition of partial distances. Concerning the second claim, if all partial distances \( D_i \) are equal to 1 for a Kernel \( T_i \), (50) becomes

\[
Z_{m-1} \leq Z_m, \text{ for all } m \geq 1.
\]

Thus, \( (Z_m)_{m \geq 0} \) is a non-decreasing positive sequence, which can converge to 0 only if it’s constant and equal to 0, contradicting the polarization property.

The following Lemma will be instrumental in the second part of the proof.

**Lemma 5.** For all \( m_1 \) and \( m_2 \) such that \( m_2 \geq m_1 \), we have that

\[
Z_{m_2} \leq \left( K^{(m_2-m_1)} Z_{m_1} \right) \prod_{k=m_1+1}^{m_2} D_{b_k}
\]

**Proof.** Let \( m_2 \geq m_1 \geq 0 \) be two integers. We have from the right hand side of inequality (55) that

\[
Z_{m_2} \leq K Z_{m_2-1}
\]

\[
\leq K K_{D_{m_2}} Z_{D_{m_2} D_{m_2-1}}
\]

\[
\leq K \left[ \prod_{m=m_1+1}^{m_2} D_{b_k} \prod_{k=m_1+1}^{m_2} D_{b_k} \right]
\]

\[
\leq K \left( \prod_{k=m_1+1}^{m_2} D_{b_k} \right) \prod_{k=m_1+1}^{m_2} D_{b_k}
\]

\[
= \left( K^{(m_2-m_1)} Z_{m_1} \right) \prod_{k=m_1+1}^{m_2} D_{b_k}
\]

where (a) and (b) result from both that \( K \geq 1 \), and from claim 2) Lemma 2

B. Proof of condition 1)

To show condition 1), we rely on the left hand side of (56). Given an \( \epsilon \geq 0 \) with \( m \geq M \) satisfying (55), we have that

\[
\prod_{k=1}^{m} D_{b_k} \geq \prod_{k=1}^{m} Z_{b_k} \geq Z_{m-1}^{Z_{m-1}}
\]

The exponent of \( Z_0 \) can be upper bounded as

\[
\prod_{k=1}^{m} D_{b_k} = \prod_{j=1}^{s} \prod_{k \in X_j} D_{b_k}
\]

\[
= \prod_{j=1}^{s} l_j \prod_{k \in X_j} D_{b_k}^{l_j}
\]

(70)
\[
\begin{align*}
&= \prod_{j=1}^{s} \left[ \sum_{i=1}^{l_j} n_{i,j} \log_{l_j}(D_{b_i}) \right] \\
&= \prod_{j=1}^{s} \left[ \sum_{i=1}^{l_j} n_{i,j} \log_{l_j}(D_{b_i}) \right]
\end{align*}
\]

Lemma 6. Proof of condition 2)

Proof. The proof of this Lemma is a generalization of the proof of \[\text{Lemma 5.11}.\] Defining the event

\[
\Omega \triangleq \{ w : \lim_{m \to \infty} Z_m(w) = 0 \}
\]

\[
= \{ w : \forall k > 1, \exists m_0 \text{ s.t. } Z_m(w) \leq \frac{1}{k} \forall m \geq m_0 \}
\]

\[
= \bigcap_{k>1} \bigcup_{m_0 \geq 0} A_{m_0,k}.
\]

where

\[
A_{m_0,k} \triangleq \{ w : Z_m(w) \leq \frac{1}{k} \forall m \geq m_0 \}.
\]

Since the multi-kernel polar code defined by \( G_N \) polarizes, i.e. the sequence \((Z_m)_{m>0}\) converges to a Bern(1-I0) random variable, we have that

\[
P(\Omega) = I_0.
\]

In the following, the notation \( Z_m(w) \) will be shortened to \( Z_m \) since the dependence on \( w \) is implicit. Let \( k \geq 1 \) be fixed. The sequence of sets \((A_{m_0,k})_{m_0 \geq 0}\) is increasing in \( m_0 \), thus we can write

\[
A_{m_0,k} = \bigcup_{n \leq m_0} A_{n,k}.
\]

If we define

\[
A_{\infty,k} \triangleq \bigcup_{m_0 \geq 0} A_{m_0,k},
\]

we can write that

\[
\lim_{m_0 \to \infty} P(A_{m_0,k}) = \lim_{m_0 \to \infty} P \left( \bigcup_{n \leq m_0} A_{n,k} \right)
\]

\[
= P \left( \lim_{m_0 \to \infty} \bigcup_{n \leq m_0} A_{n,k} \right)
\]

\[
= P(A_{\infty,k}).
\]

Thus, for all \( k \geq 1 \), for all \( \epsilon > 0 \) there exists \( M_0 \) such that for all \( m_0 \geq M_0 \)

\[
P(A_{m_0,k}) \geq P(A_{\infty,k}) - \epsilon.
\]

More specifically, for \( k = K^{l^{*}+1} \), for all \( \epsilon > 0 \) there exists \( M_0 \) such that

\[
P(A_{M_0,K^{l^{*}+1}}) \geq P(A_{\infty,K^{l^{*}+1}}) - \epsilon
\]

\[
\geq P \left( \bigcap_{k \geq 1} A_{\infty,k} \right) - \epsilon
\]

\[
= P(\Omega) - \epsilon
\]

\[
= I_0 - \epsilon,
\]

which concludes the proof.

\[ \Box \]

The following shows a key inequality on the exponential decay of the sequence \((Z_m)_{m>0}\).
Lemma 7. For every $\epsilon > 0$, there exists $M_1$ such that for all $m \geq M_1$

$$P\left(Z_m \leq K^{-\frac{m}{2^*}}\right) > I_0 - \epsilon. \quad (103)$$

Proof. Fix an $\epsilon > 0$. Following the result of the previous lemma, let $M_0 > 0$ such that

$$P(A_{M_0,K^{l*+1}}) \geq I_0 - \epsilon. \quad (104)$$

Besides, from the right-hand side of inequality (56), we have also that

$$Z_{m+1} \leq KZ_m D_{m+1}^{b_m} \leq KZ_m a_m M \quad (105)$$

for every $w \in A_{M_0,K^{l*+1}}$, for all $m \geq M_0$. Now, given the result of Lemma 4, we can bound $Z_{m+1}$ as

$$Z_{m+1} \leq Z_m K^{l*} \quad \text{if } b_{m+1} = i_{m+1}$$

and

$$Z_{m+1} \leq Z_m K \quad \text{if } b_{m+1} \neq i_{m+1} \quad (110)$$

where $i_{m+1}$ is the index for which the partial distance satisfies $D_{i_{m+1}} \geq 2$.

Now, combining (110) and (119), for all $\epsilon > 0$ there exists an $M_1 \geq M_0$ such that, for all $m_1 \geq M_1$,

$$P(T^{m_1}_{m_0}) \geq 1 - \epsilon. \quad (117)$$

Next, since $T^{m_1}_{m_0} \subset T^{m_1}_{m_0}$, we have that

$$P(T^{m_1}_{m_0}) \geq P(T^{m_1}_{m_0}). \quad (119)$$

Now we define the typical set $T^{m_1}_{m_0}$ as the set of events for which each index $b_j \in [1 : l_j]$ appears with a probability at least equal to $\frac{\epsilon}{l_j}$.

We now define the typical set $T^{m_1}_{m_0}$ as the set of events for which each index $b_j \in [1 : l_j]$ appears with a probability at least equal to $\frac{\epsilon}{l_j}$.

Since every index $b_j$ occurs asymptotically with a probability $\frac{\epsilon}{l_j}$, then

$$\lim_{m_1 \to \infty} P(T^{m_1}_{m_0}) = 1, \quad (117)$$

hence, and for all $\epsilon > 0$ there exists an $M_1 \geq M_0$ such that, for all $m_1 \geq M_1$,

$$P(T^{m_1}_{m_0}) \geq 1 - \epsilon. \quad (118)$$

By combining (116) and (119), for all $\epsilon > 0$ there exists an $M_1 \geq M_0$ such that, for all $m_1 \geq M_1$,

$$P(T^{m_1}_{m_0}) \geq 1 - \epsilon. \quad (120)$$

For such $\epsilon$, let us choose $M_1$ such that $M_1 \geq 2M_0$, so that, for all $m_1 \geq M_1$, inside the set $T^{m_1}_{m_0}$, the partial distance satisfies $D_{i_{m+1}} \geq 2$.

If we define the typical set $T^{m_1}_{m_0}$ as the set of events for which each index $b_j \in [1 : l_j]$ appears at least $\alpha l_j$ fraction of the time between $m_0$ and $m_1$, we have that

$$\alpha^{m_1}_{m_0} \geq \sum_{j=1}^s p_j \alpha = \alpha \quad (113)$$

inside $T^{m_1}_{m_0}$. Note that the typical sets $T^{m_1}_{m_0}$ are non-increasing in $\alpha$ for a fixed $m_0$ and $m_1$, and increasing in $m_1$ for a fixed $\alpha$ and $m_0$. Given

$$\alpha^* = \frac{2l* + 1}{2(l* + 1)} \quad (114)$$

for $\alpha^*$ it holds that

$$\alpha^* \leq \frac{1}{l_j} \quad \text{for all } j \in [1 : s]. \quad (115)$$

Thus, since the sets $T^{m_1}_{m_0}$ are non-increasing in $\alpha$, we can write that

$$\lim_{m_1 \to \infty} P(T^{m_1}_{m_0}) \geq \lim_{m_1 \to \infty} P(T^{m_1}_{m_0}). \quad (116)$$

Therefore, we can bound $Z_{m+1}$ as

$$Z_{m+1} \leq Z_m K^{l*} + 1 \quad (111)$$

where

$$\alpha^{m_1}_{m_0} \geq \frac{1}{l_j} \quad (113)$$

inside $T^{m_1}_{m_0}$. Note that the typical sets $T^{m_1}_{m_0}$ are non-increasing in $\alpha$ for a fixed $m_0$ and $m_1$, and increasing in $m_1$ for a fixed $\alpha$ and $m_0$. Given

$$\alpha^* = \frac{2l_1 + 1}{2(l_1 + 1)} \quad (114)$$

for $\alpha^*$ it holds that

$$\alpha^* \leq \frac{1}{l_j} \quad \text{for all } j \in [1 : s]. \quad (115)$$

Thus, since the sets $T^{m_1}_{m_0}$ are non-increasing in $\alpha$, we can write that

$$\lim_{m_1 \to \infty} P(T^{m_1}_{m_0}) \geq \lim_{m_1 \to \infty} P(T^{m_1}_{m_0}). \quad (116)$$

Now we are ready to prove condition 2) of definition 3. Let $\epsilon > 0$ and let $\alpha < \frac{1}{l_j}$, and let $\gamma < 1$ such that

$$\forall j \in [1 : s], \quad \alpha l_j \gamma > 1 - \epsilon. \quad (128)$$

We define the constant $C$, independent of the blocklength, as

$$C \triangleq \sum_{j=1}^s p_j l_j, \quad (129)$$

Let $m_3$ be an integer large enough so that, by defining

$$m_1 \triangleq \log(2m_3)8l^* \quad (130)$$

and

$$m_2 \triangleq \left(1 + \frac{1}{8l^*}\right)m_1 \quad (131)$$

we have

$$\lim_{m_1 \to \infty} P(T^{m_1}_{m_0}) = 1, \quad (117)$$

hence, and for all $\epsilon > 0$ there exists an $M_1 \geq M_0$ such that, for all $m_1 \geq M_1$,
it holds that
\[
m_1 \geq \max(M_0, 8l^*) \tag{132}
\]
\[
m_3 - m_2 \geq \gamma m_3, \tag{133}
\]
while the typical sets \(T_{\alpha,m_1}, T_{\alpha,m_2}\) verify
\[
\mathbb{P}(T_{\alpha,m_1}^{m_2}) > 1 - \epsilon \quad \text{and} \quad \mathbb{P}(T_{\alpha,m_2}^{m_3}) > 1 - \epsilon. \tag{134}
\]
We start the proof by bounding \(Z_{m_2}\):
\[
Z_{m_2} \leq (K^{m_2-m_1}Z_{m_1}) \prod_{i=m_1+1}^{m_2} D_{b_i} \tag{135}
\]
\[
\leq (K^{m_2-m_1}Z_{m_1}) \prod_{i=m_1+1}^{m_2} D_{b_i}, \tag{136}
\]
\[
= (K^{m_2-m_1}Z_{m_1}) \prod_{i=m_1+1}^{m_2} D_{b_i}, \tag{137}
\]
\[
\leq K \prod_{i=m_1+1}^{m_2} D_{b_i} \tag{138}
\]
\[
= K \prod_{i=m_1+1}^{m_2} D_{b_i}, \tag{139}
\]
\[
= K \prod_{i=m_1+1}^{m_2} D_{b_i}, \tag{140}
\]
\[
= K^{c^{m_2-m_1}} \tag{141}
\]
\[
= K^{c^{m_2-m_1}}. \tag{142}
\]
where (a) stems from Lemma [5], (b) is the result of Lemma [7], (c) results from the definition of \(m_2\) in (131), (d) is a consequence of the constraint (132), (e) stems from the definition of \(T_{\alpha,m_1}\), and (f) is a result of (133). Let us now bound the term \(Z_{m_3}\) as
\[
Z_{m_3} \leq (K^{m_3-m_2}Z_{m_2}) \prod_{i=m_2+1}^{m_3} D_{b_i} \tag{143}
\]
\[
\leq (K^{m_3-m_2}Z_{m_2}) \prod_{i=m_2+1}^{m_3} D_{b_i}, \tag{144}
\]
\[
= (K^{m_3-m_2}Z_{m_2}) \prod_{i=m_2+1}^{m_3} D_{b_i}, \tag{145}
\]
\[
\leq K \prod_{i=m_2+1}^{m_3} D_{b_i} \tag{146}
\]
\[
= K \prod_{i=m_2+1}^{m_3} D_{b_i}, \tag{147}
\]
\[
= K \prod_{i=m_2+1}^{m_3} D_{b_i}, \tag{148}
\]
where (a) is a consequence of (142), (b) stems from the definition of \(m_3\) in (130), (c) results from the not-limiting assumption \(m_3 \geq 2\), (d) follows from the definition of \(T_{\alpha,m_2}\), (e) results from the assumption in (133), (f) is a consequence of (128), while (g) follows by noting that the block length at iteration \(m_3\) is \(N_3 = \prod_{i=1}^{m_3} l_i\). The proof is concluded using similar calculations as in (81), and noticing that
\[
\mathbb{P}\{w : Z_m \leq K^{-\frac{m}{m^*}} \cap T_{\alpha,m_1}^{m_2} \cap T_{\alpha,m_2}^{m_3} \geq I_0 - \epsilon. \tag{152}
\]
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, July 2009.
[2] S. B. Korada, E. Sasoglu, and R. Urbanke, “Polar codes: Characterization of exponent, bounds, and constructions,” IEEE Transactions on Information Theory, vol. 56, no. 12, pp. 6253–6264, Dec. 2010.
[3] N. Pressman, O. Shapira, S. Litsyn, T. Etzion, and A. Vardy, “Binary polarization kernels from code decompositions,” IEEE Transactions on Information Theory, vol. 61, no. 5, pp. 2227–2239, May 2015.
[4] E. Gabry, V. Bioglio, I. Land, and J.-C. Belfiore, “Multi-kernel construction of polar codes,” in IEEE International Conference on Communications (ICC), Paris, France, May 2017, preprint available on arXiv:1612.06099.
[5] R. Mori and T. Tanaka, “Performance of polar codes with the construction using density evolution,” IEEE Communications Letters, vol. 13, no. 7, pp. 519–521, July 2009.
[6] E. Arikan and E. Telatar, “On the rate of channel polarization,” in Information Theory, 2009. ISIT 2009. IEEE International Symposium on. IEEE, 2009, pp. 1493–1495.
[7] A. Wyner, “A theorem on the entropy of certain binary sequences and applications – part I,” IEEE Transactions on Information Theory, vol. 19, no. 6, pp. 769–772, Nov 1973.
[8] E. Sasoglu, “Polarization and polar codes,” Foundations and Trends in Communications and Information Theory, vol. 8, no. 4, pp. 259–381, 2012.