# Polar Codes with Mixed Kernels

Noam Presman, Ofer Shapira and Simon Litsyn
School of Electrical Engineering, Tel Aviv University, Ramat Aviv 69978 Israel.
e-mails: {presmann, ofershap, litsyn}@eng.tau.ac.il.

**Abstract**

A generalization of the polar coding scheme is proposed. It exploits several homogeneous kernels over alphabets of different sizes. An analysis of the introduced scheme is undertaken. Specifically, asymptotic properties of the polarization are shown to be strongly related to the ones of the constituent kernels.

## 1 Introduction

Polar codes were introduced by Arikan in [1] and provided a scheme for achieving the symmetric capacity of binary memoryless channels (B-MC) with polynomial encoding and decoding complexity. Originally, Arikan considered a simple binary and linear 2 dimensional kernel, which is based on the $(u+v, v)$ mapping. This mapping is extended to support an arbitrary code length $N = 2^n$, by a Kronecker power of the generating matrix that defines the transformation. Multiplying a permutation of an input vector $\mathbf{u}$ by this matrix results in a vector $\mathbf{x}$, that is transmitted through $N$ independent copies of a memoryless channel, $\mathcal{W}$. As a result, $N$ dependent channels between the components of $\mathbf{u}$ and the outputs of the copies of the channel $\mathcal{W}$ are created. These channels exhibit polarization under successive cancelation (SC) decoding: as $n$ grows there is a proportion of $I(\mathcal{W})$ (the symmetric channel capacity) of the channels that have their capacity approaching 1, while the rest of the channels have their capacity approaching 0.

The exponent of the kernel as a measure of the asymptotic rate of polarization for arbitrary polar codes was introduced in [2] and generalizations were given in [3]. In [4], the authors suggested designing binary kernels based on the idea of code decomposition. However, the more natural way is to take advantage of the explicit (non-binary) code decomposition in order to design a kernel. This however, usually requires introducing additional kernels in respect to the initial binary kernel, which results in a mixed kernel structure. Our objective in this paper is to explore such constructions and analyze them.

This paper is organized as follows. In Section 2, we review the idea of code decomposition and its relation to the design of polar code kernels. This notion is the motivation for the introduction of mixed kernels. For simplicity, we decided to present the concept of mixed kernels by an example of a specific construction that is composed of a binary kernel and a quaternary kernel. This is done in Section 3. General mixed kernels are considered in Section 4. Simulations results and conclusions are given in Section 5.

Throughout we use the following notations. For $i \geq j$, let $\mathbf{u}_j^i = (u_j, ..., u_i)$ be the sub-vector of a vector $\mathbf{u}$ of length $i-j+1$ (if $i < j$ we say that $\mathbf{u}_j^i = ()$, the empty vector, and its length is 0). For a natural number $n$, we denote $[n] = \{1, 2, 3, ..., n\}$.

## 2 Preliminaries

In [4], we explored the idea of code decomposition and its relation to the construction of binary kernels for polar codes. We review these ideas here.

**Definition 1** *The set $\{T_1, ..., T_m\}$ is called a decomposition of $\{0,1\}^\ell$ , if $T_1^{()} = \{0,1\}^\ell$, and $T_i^{(\mathbf{b}_1^{i-1})}$ is partitioned into $m_i$ equally sized sets $\left\{T_{i+1}^{(\mathbf{b}_1^{i-1} b_i)}\right\}_{b_i=0,1,...,m_i-1}$, of size $\frac{2^\ell}{\prod_{j=1}^{i} m_j}$ $(i \in [m-1])$. We denote the set of sub-codes of level number $i$ by*

$$T_i = \left\{ T_i^{(\mathbf{b}_1^{i-1})} | b_j \in \{0, 1, 2, ..., m_j - 1\}, j \in [i-1] \right\}.$$

*The partition is usually described by the following chain of codes parameters*

$$(n_1, k_1, d_1) - (n_2, k_2, d_2) - ... - (n_m, k_m, d_m),$$

*if for each* $\mathcal{T} \in T_i$ *we have that* $\mathcal{T}$ *is a code of length* $n_i$*, size* $2^{k_i}$ *and minimum distance at least* $d_i$*.*

A transformation $g(\cdot)$ can be associated to a code decomposition in the following way.

**Definition 2** *Let* $\{T_1, ..., T_m\}$ *be a code decomposition of* $\{0,1\}^\ell$ *as in Definition 1, and such that* $\forall \mathcal{T} \in T_m$, $|\mathcal{T}| = 1$.

*The transformation*

$$g(v_1, v_2, ..., v_m) : \prod_{i=1}^{m} \{0,1\}^{m_i} \to \{0,1\}^\ell \quad ; \sum_{i=1}^{m} m_i = \ell \tag{1}$$

*induced by this binary code decomposition is defined as follows for* $\mathbf{v} \in \prod_{i=1}^{m} \{0,1\}^{m_i}$*.*

$$g(\mathbf{v}_1^m) = \mathbf{x}_1^\ell \quad \text{if } \mathbf{x}_1^\ell \in T_m^{(\mathbf{v}_1^m)}, \tag{2}$$

*where in the notation of* $T_m^{(\mathbf{v}_1^m)}$ *we take the decimal representation of the components of* $\mathbf{v}$*, for consistency with Definition 1. Sometimes, it is useful to denote the argument to* $g(\cdot)$ *as the vector* $\mathbf{u} \in \{\mathbf{0},\mathbf{1}\}^\ell$*, i.e. write* $g(\mathbf{u})$ *instead of* $g(\mathbf{v})$ *where* $\mathbf{v} \in \prod_{i=1}^{m} \{0,1\}^{m_i}$*. In this case, there exists the obvious correspondence between* $\mathbf{v}$ *and* $\mathbf{u}$*, that is* $v_i = \mathbf{u}_{1+\sum_{j=1}^{i-1} m_j}^{\sum_{j=1}^{i} m_j}$ $i \in [m]$*. We say that* $v_i$ *is representing* $m_i$ *bits that are "glued" together. It is convenient to denote* $v_i$ *as* $u_{s,f}$*, if* $v_i = \mathbf{u}_s^f$*.*

In [4, Example 1], we considered the decomposition into cosets described by the chain $(4,4,1) - (4,3,2) - (4,1,4)$. Using Definition 2, we introduce a kernel function

$$g_1(u_1, u_{2,3}, u_4) : \{0,1\} \times \{0,1\}^2 \times \{0,1\} \to \{0,1\}^4 \tag{3}$$

that is induced by this decomposition. The first bit $u_1$ chooses between the sub-codes $T_2^{(0)}$ and $T_2^{(1)}$. The second and the third bits are glued together, forming a binary pair, or quaternary symbol $u_{2,3}$ and they choose the correct sub-code of $T_2^{(u_1)}$. Finally, $u_4$ selects the code-word from the chosen sub-code. Note, that an easy implementation of the encoding is to multiply $\mathbf{u}$ by the proper generating matrix. Indeed, there's nothing new in this construction. The challenge is to extend this mapping to an $N = 4^n$ length mapping. The standard Arikan's construction (based on the Kronecker power) does not suffice, because of the glued bits $u_{2,3}$, that need to be jointly treated as a quaternary symbol. To facilitate this, we suggest introducing a second quaternary kernel, $g_2(\cdot)$. Because different coordinates of the input of $g_1(\cdot)$ are from different alphabet sizes, and because in order to implement this polarization scheme, we incorporate two mapping functions $g_1(\cdot)$ and $g_2(\cdot)$, we refer to the overall construction as a *mixed kernel* construction. Details on how to combine kernels $g_1(\cdot)$ and $g_2(\cdot)$ to a mixed kernel are given in Section 3.

# 3 Mixed Kernels by an Example

We begin by describing a construction of a mixed kernel by several homogeneous kernels over different alphabets. To have a comprehensive presentation of this subject, we decided to focus on a specific construction. Generalizations easily follow from this example and are given in Section 4.

## 3.1 Construction of a Mixed Kernel

Let $g_1(\cdot)$ be the mapping defined in (3). Let $g_2(\cdot) : \left(\{0,1\}^2\right)^4 \to \left(\{0,1\}^2\right)^4$ be a polarizing kernel over the quaternary alphabet. For example, $g_2(\cdot)$ can be a kernel, based on the extended Reed-Solomon code of length 4, $G_{RS}(4)$ that was shown in [3, Example 20] to be a polarizing kernel. Using $g_2(\cdot)$, we can extend the mapping of $g_1(\cdot)$ to an $N = 4^n$ length mapping. Note that $g_2(\cdot)$ is introduced in order to handle the glued bits $u_{2,3}$ in the input of $g_1(\cdot)$.

Let us first review the channel splitting principle using $g_1(\cdot)$. The output of $g_1(\cdot)$ is binary and so is the channel on which the result of the transformation is sent on. The meaning of taking two inputs and glue them together is that we want these inputs to be treated as one unit for decision making and decoding. Assume a binary vector $\mathbf{u}$ was transformed to $\mathbf{x}$ by $g_1(\cdot)$.

$$g_1(u_1, u_{2,3}, u_4) = \mathbf{x}_1^4 \quad u_1, u_4 \in \{0,1\},$$

$$u_{2,3} \in \{0,1\}^2, x_i \in \{0,1\}, i \in [4]$$

$\mathbf{x}_1^4$ is transmitted over 4 copies of the binary memoryless channel $\mathcal{W}$, and we receive the output vector $\mathbf{y}$. The channel splitting principle dictates the following channels.

$$W_4^{(1)}(\mathbf{y}_1^4|u_1) = \sum_{u_{2,3}\in\{0,1\}^2, u_4\in\{0,1\}} \frac{1}{2^3} W_4(\mathbf{y}_1^4|u_1, u_{2,3}, u_4)$$

$$W_4^{(2,3)}(\mathbf{y}_1^4, u_1|u_{2,3}) = \sum_{u_4\in\{0,1\}} \frac{1}{2^2} W_4(\mathbf{y}_1^4|u_1, u_{2,3}, u_4)$$

$$W_4^{(4)}(\mathbf{y}_1^4, u_1, u_{2,3}|u_4) = \frac{1}{2^3} W_4(\mathbf{y}_1^4|u_1, u_{2,3}, u_4).$$

Next, consider $g_2(\cdot)$, which is a quaternary input and output mapping. A binary vector $\mathbf{u} \in \{0,1\}^8$ is transformed into $\mathbf{x} \in \left\{\{0,1\}^2\right\}^4$ in the following fashion.

$$g_2(u_{1,2}, u_{3,4}, u_{5,6}, u_{7,8}) = \mathbf{x}_1^4 \quad u_{2i-1,2i}, x_i \in \{0,1\}^2 \;\; i \in [4]$$

$\mathbf{x}_1^4$ is transmitted over 4 copies of a quaternary input memoryless channel $\tilde{\mathcal{W}}$, and the output vector $\mathbf{y}$ is received. By the channel splitting principle we get the following channels for $i \in [4]$.

$$\tilde{W}_4^{(2i-1,2i)}(\mathbf{y}, \mathbf{u}_1^{2i-2}|u_{2i-1,2i}) =$$

$$= \sum_{\mathbf{u}_{2i+1}^8\in\{0,1\}^{8-2i}} \frac{1}{4^3} \tilde{W}_4(\mathbf{y}|\mathbf{u}_1^{2i-2}, u_{2i-1,2i}, \mathbf{u}_{2i+1}^8).$$

We denote $g^{(1)}(\cdot) \equiv g_1(\cdot)$. Constructing a mapping function of dimension 16, denoted by $g^{(2)}(\cdot)$, is done as follows. Let $\mathbf{u}$ be a binary vector of length 16. Define $g_1(u_1, u_{2,3}, u_4) = \mathbf{a}$, $g_2(u_{5,6}, u_{7,8}, u_{9,10}, u_{11,12}) = \mathbf{b}$ and $g_1(u_{13}, u_{14,15}, u_{16}) = \mathbf{c}$. Finally,

$$g^{(2)}(\mathbf{u}) = \big[g_1(a_1, b_1, c_1), g_1(a_2, b_2, c_2), \\ g_1(a_3, b_3, c_3), g_1(a_4, b_4, c_4)\big]. \tag{4}$$

In order to extend this construction to a general kernel $g^{(k)}(\mathbf{u}_1^{4^k})$ in which some of the inputs are glued, we suggest the following recursive algorithm.

**Mixed Kernel Construction Algorithm**

**STEP 1:** Take 4 parallel copies of $g^{(k-1)}(\cdot)$, and allocate binary inputs (that some of them will be glued) by $u_1, u_2, ...u_{4^k}$. Denote the direct inputs to $g^{(k-1)}(\cdot)$ by $\mathbf{v}_1^{4^{k-1}}$. Since we simultaneously deal with all the 4 inputs to the copies of $g^{(k-1)}(\cdot)$ having the same index, there's no need to denote them separately. Our goal is to allocate inputs of the set $u_1, u_2, ...u_{4^k}$, glue them together if necessary, perform the proper transformation (i.e. $g_1(\cdot)$ or $g_2(\cdot)$) and connect the outputs of these transformations to the inputs of $g^{(k-1)}(\cdot)$.

Initialize two counters: $i \leftarrow 1$ for the inputs of $g^{(k-1)}(\cdot)$ and $j \leftarrow 1$ for the inputs of $g^{(k)}(\cdot)$.

**STEP 2:** Consider input $v_i$ of $g^{(k-1)}(\cdot)$. There are two cases here. **(a)** $v_i$ is single (i.e. it is not glued to the next input). Allocate inputs $u_j, u_{j+1}, u_{j+2}, u_{j+3}$, use them as inputs to $g_1(\cdot)$, and use the outputs of the transformation as inputs to the ordered copies of $v_i$ in the inputs of the copies of $g^{(k-1)}(\cdot)$. Note: $u_{j+1}$ and $u_{j+2}$ are now glued together (this symbol is denoted by $u_{j+1,j+2}$) and the other inputs are binary. Set

$i \leftarrow i+1$, $j \leftarrow j+4$. **(b)** $v_i$ is glued to $v_{i+1}$ (i.e. we have a quaternary symbol $v_{i,i+1}$). Allocate inputs $u_j^{j+8}$, glue them in pairs (i.e. $u_{j,j+1}, u_{j+2,j+3}, u_{j+4,j+5}, u_{j+6,j+7}$), use the four pairs as inputs to $g_2(\cdot)$, and take the outputs of the transformation as inputs to the ordered copies of $v_{i,i+1}$ in the inputs of the copies of $g^{(k-1)}(\cdot)$. Set $i \leftarrow i+2$, $j \leftarrow j+8$.

If you finished allocating the inputs of $g^{(k)}(\cdot)$ then stop, otherwise repeat **STEP 2**.

Note that the algorithm is consistent with the definition of $g^{(2)}(\cdot)$ in (4). The construction supports Arikan's analysis by the channel tree-process as we see in Section 3.2. Also note, that by this construction, successive cancelation decoding of the inputs to $g^{(k)}(\cdot)$ is actually decoding of inputs to the transformations $g_1(\cdot)$ or $g_2(\cdot)$ that use as a channel one of the synthesized channels generated by $g^{(k-1)}(\cdot)$ over $\mathcal{W}$. In other words, when decoding one bit $u_i$ (two glued bits $u_{i,i+1}$) over the channel $W_{4^k}^{(i)}(\mathbf{y}, \mathbf{u}_1^{i-1}|u_i)$ (over the channel $W_{4^k}^{(i,i+1)}(\mathbf{y}, \mathbf{u}_1^{i-1}|u_{i,i+1})$), this is manifested as decoding a bit (a pair of two glued bits) which is an input to the transformations $g_1(\cdot)$ or $g_2(\cdot)$. These transformations use as a channel the proper synthesized channel ( $\mathcal{W}_{4^{k-1}}^{(j)}$ or $\mathcal{W}_{4^{k-1}}^{(j,j+1)}$, depending on $i$).

## 3.2 The Tree Process

We now turn to describe the channel tree process corresponding to this mixed kernel construction. A random sequence $\{W_n\}_{n\geq 0}$ is defined such that $W_n \in \left\{\mathcal{W}_{4^n}^{(\tau_n(i))}\right\}_{i=1}^{\nu(n)}$, where $\nu(n)$ denotes the number of channels (where the glued channels are counted as one), and $\tau_n(i)$ denotes the index of channel number $i$ ($\tau_n(i)$ is needed because some of the channels correspond to glued bits and therefore have their indexing as a pair of integer numbers). For example, for the $\mathcal{W}_{16}$ channel, constructed using the transformation in (4), we have the number of channels $\nu(2) = 10$, where the values of $\tau_2(\cdot)$ are $[1, (2,3), 4, (5,6), (7,8), (9,10), (11,12), 13, (14,15), 16]$. We also denote by $\{N_n\}_{n\geq 0}$ the number of bits at the input of the channel, which in our case is $N_n = 1$ when we deal with a single bit channel or $N_n = 2$ when we deal with a channel of glued bits. We have the following definition of the channel processes.

$$W_{n+1} = W_n^{(B_n)} \text{ for } n \geq 0\,;\; W_0 = \mathcal{W}, N_0 = 1.$$

The probabilistic dynamics of $\{B_n\}_{n\geq 0}, \{N_n\}_{n\geq 0}$ need to be described. Let $\left\{B_n^{(1)}\right\}_{n\geq 0}$ be an i.i.d random sequence of the values $[1, (2,3), 4]$ with corresponding probabilities $[0.25, 0.5, 0.25]$, and let $\left\{B_n^{(2)}\right\}_{n\geq 0}$ be an i.i.d random sequence of the values $[(1,2), (3,4), (5,6), (7,8)]$ with uniform probabilities. Denote by the random variable $T$ the minimum non-negative $n$ such that $B_n^{(1)} = (2,3)$, and set

$$N_n = \begin{cases} 1, & n \leq T; \\ 2, & n > T. \end{cases}$$

Finally, set $B_n = B_n^{(N_n)}$. Note that $T$ is a geometric random variable with probability of success $p = 0.5$. Furthermore, given the value of $T$ the sequence of $B_n$ is of independent samples (although the distribution is not identical for all samples). Note also, that the pairs of numbers in the sequence of $B_n$ indicate channels having inputs of two glued bits.

Suppose we have a certain channel $\mathcal{W}$ and binary i.i.d input vector $U_1^4$ that is transformed by $g_1(\cdot)$ to $X_1^4$, transmitted over a B-MC channel, and received as $Y_1^4$, we have

$$4I(\mathcal{W}) = I(Y_1^4; U_1^4) = I(Y_1^4; U_1) + I(Y_1^4; U_{2,3}|U_1) + \tag{5}$$

$$+I(Y_1^4, U_4|U_1^3) = I(\mathcal{W}^{(1)}) + I(\mathcal{W}^{(2,3)}) + I(\mathcal{W}^{(4)}).$$

Next, define the information random sequence corresponding to the channels as $\{I_n\}_{n\geq 0}$.

$$I_n = \frac{I(W_n)}{N_n} \quad n \geq 0. \tag{6}$$

The Bhattacharyya parameter sequence is denoted by $Z_n = Z(W_n)$, where for a $q$-ary channel $\mathcal{W}$ we have $Z(\mathcal{W}) = \frac{1}{q\cdot(q-1)} \sum_{x,x' \in \mathcal{X}^2, x \neq x'} Z_{x,x'}(\mathcal{W})$. Here, $\mathcal{X}$ is the alphabet of the channel, and

$$Z_{x,x'}(\mathcal{W}) = \sum_{y \in \mathcal{Y}} \sqrt{W(Y=y|X=x)\, W(Y=y|X=x')}.$$

The maximum and the minimum of the Bhattacharyya parameters between two symbols are defined as $Z_{\max}(\mathcal{W}) = \max_{x,x' \in \mathcal{X}, x \neq x'} Z_{x,x'}(\mathcal{W})$, and $Z_{\min}(\mathcal{W}) = \min_{x,x' \in \mathcal{X}} Z_{x,x'}(\mathcal{W})$. Observe that if $|\mathcal{X}| = 2$, then $Z_{\max}(\mathcal{W}) = Z_{\min}(\mathcal{W}) = Z(\mathcal{W})$.

Note that $I_n \in [0,1]$, and so is also $Z_n$. By using [5, Proposition 3], it can be shown that $Z_n \to 1 \iff I_n \to 0$, and that $Z_n \to 0 \iff I_n \to 1$.

**Proposition 1** *The process $\{I_n\}_{n \geq 0}$ is a bounded martingale which is uniformly integrable. As a result, it converges almost surely to $I_\infty$.*

**Proof** By the definition of the information sequence (6) we have

$$\mathbb{E}\left[I_{n+1} | I_n, N_n = 1\right] = \frac{1}{4} \frac{I\left(W_n^{(1)}\right)}{1} + \frac{2}{4} \frac{I\left(W_n^{(2,3)}\right)}{2} + \frac{1}{4} \frac{I\left(W_n^{(4)}\right)}{1} \tag{7}$$

Using (5) we have that

$$\begin{aligned} &\mathbb{E}\left[I_{n+1} | I_n, N_n = 1\right] = \\ &= \frac{1}{4}\left(I\left(W_n^{(1)}\right) + I\left(W_n^{(2,3)}\right) + I\left(W_n^{(4)}\right)\right) = \frac{I(W_n)}{N_n} = I_n \end{aligned} \tag{8}$$

On the other hand

$$\begin{aligned} &\mathbb{E}\left[I_{n+1} | I_n, N_n = 2\right] = \\ &= \frac{1}{4} \frac{I\left(W_n^{(1,2)}\right)}{2} + \frac{1}{4} \frac{I\left(W_n^{(3,4)}\right)}{2} + \frac{1}{4} \frac{I\left(W_n^{(5,6)}\right)}{2} + \frac{1}{4} \frac{I\left(W_n^{(7,8)}\right)}{2} \end{aligned} \tag{9}$$

which is

$$\begin{aligned} \mathbb{E}\left[I_{n+1} | I_n, N_n = 2\right] = \frac{1}{2} \cdot \frac{1}{4}\left(I\left(W_n^{(1,2)}\right) + I\left(W_n^{(3,4)}\right) + \right. \\ \left. + I\left(W_n^{(5,6)}\right) + I\left(W_n^{(7,8)}\right)\right) = \frac{I(W_n)}{N_n} = I_n \end{aligned} \tag{10}$$

So, by taking (8) and (10) we have

$$\mathbb{E}\left[I_{n+1} | I_n\right] = I_n, \tag{11}$$

which means that the sequence $\{I_n\}_{n \geq 0}$ is a martingale. Furthermore, it is uniformly integrable (see for example, [6, Theorem 4.5.3]) and therefore it converges almost surely to $I_\infty$. $\diamondsuit$

Note that

$$\Pr\left(I_n \in S\right) = \frac{1}{4^n} \sum_{i \in [\nu(n)] \ \ s.t. \ \ I(\mathcal{W}_{4^n}^{(\tau_n(i))}) \in S} \#\left(\tau_n(i)\right), \tag{12}$$

where $\#\left(\tau_n(i)\right)$ counts the number of bits at the input of channel $\tau_n(i)$, which is 1 for a single bit channel, and 2 for a glued 2 bits channel. A similar expression to (12) can be stated for the process $Z_n$. This probabilistic method gives the two bits of the glued bits pair, the same behavior in terms of probability of decoding error and mutual information, and as such they are counted. Note that $\mathbb{E}\left[I_n\right] = \mathbb{E}\left[I_\infty\right] = I(\mathcal{W})$. Thus, by showing that the mixed kernel is polarizing, i.e. $I_\infty \in \{0,1\}$, we may infer that the proportion of clean channels (created by the transformation and successive cancelation decoding) is $I(\mathcal{W})$ by (12).

Also note that for $g^{(n)}(\cdot)$ we can easily count the number of glued 2 bits input channels (denoted here by $\gamma_n$) as $\gamma_n = 4^n \cdot \frac{1}{2} \cdot \Pr(N_n = 2) = \frac{4^n}{2} \cdot \left(1 - \frac{1}{2^n}\right)$. The proportion of the glued 2 bits channel goes to 1 as $n$ grows,

and so is the number of uses of $g_2(\cdot)$ kernel. Because of this, the properties of $g_2(\cdot)$ dominate the construction asymptotically. Specifically, we show in the sequel, that if the kernel $g_2(\cdot)$ is polarizing, so is the mixed kernel construction we propose. Moreover, if the kernel $g_2(\cdot)$ has a lower bound and an upper bound on the exponent, $E_1(g_2)$ and $E_2(g_2)$ respectively, then $E_1(g_2)$ and $E_2(g_2)$ serve also as a lower bound and an upper bound on the rate of polarization of the mixed configuration.

### 3.3 Polarization and Polarization Rate

In this part, we study the polarization property of the mixed kernel and its rate of polarization. We show that $g_2(\cdot)$'s properties determine the asymptotic mixed kernel properties.

**Proposition 2** *Assume that $g_2(\cdot)$ is a polarizing kernel, i.e. for a construction that is based only on $g_2(\cdot)$ we have that*

$$\lim_{n\to\infty} \Pr\left(I\left(\tilde{W}_n\right)/2 \in (\delta, 1-\delta)\right) = 0, \quad \forall \delta \in (0, 0.5) \tag{13}$$

*As a result, the mixed kernel construction is also polarizing, i.e.*

$$\lim_{n\to\infty} \Pr\left(I_n \in (\delta, 1-\delta)\right) = 0, \quad \forall \delta \in (0, 0.5) \tag{14}$$

**Proof** We prove that for a given $\delta$ for each $\epsilon > 0$ there exists an $n_0 = n_0(\delta, \epsilon)$, such that for all $n > n_0$

$$\Pr\left(I_n \in (\delta, 1-\delta)\right) < \epsilon.$$

Let $n_1$ be chosen such that $\Pr\left(N_n = 2\right) \geq 1 - \frac{\epsilon}{2}$ for each $n \geq n_1$. Now, for $n = n_1$ consider all the glued bits channels $\mathcal{W}_{4^{n_1}}^{(i,j)}$. When $n$ grows further, each one of them undergoes polarization, that is each one of the $\gamma_{n_1}$ glued channels has an index $n_2(i,j)$ such that when $n \geq n_1 + n_2$

$$\Pr\left(I(W_n)/2 \in (\delta, 1-\delta) | W_{n_1} = \mathcal{W}_{4^{n_1}}^{(i,j)}\right) < \frac{\epsilon}{2}.$$

Denote by $n_2^*$ the maximum over these $n_2(i,j)$, and by $n_0 = n_1 + n_2^*$. We have that for $n \geq n_0$

$$\Pr\left(I_n \in (\delta, 1-\delta)\right) =$$

$$= \underbrace{\Pr\left(I_n \in (\delta, 1-\delta) | N_n = 1\right)}_{\leq 1} \underbrace{\Pr\left(N_n = 1\right)}_{<\epsilon/2} + \underbrace{\Pr\left(I_n \in (\delta, 1-\delta) | N_n = 2\right)}_{<\epsilon/2} \underbrace{Pr\left(N_n = 2\right)}_{\leq 1} < \epsilon. \tag{15}$$

♢

We now turn to discuss the polarization rate. To do this, we need to consider the partial distances of the kernels. We use the notations of [3]. For a given kernel $g(v_1, v_2, \ldots, v_m)$ as defined in (1), we give the following definitions.

$$D_{x,x'}^{(i)}\left(\mathbf{v}_1^{i-1}\right) = \min_{\mathbf{w}_{i+1}^m, \tilde{\mathbf{w}}_{i+1}^m} d_H\left(g\left(\mathbf{v}_1^{i-1}, x, \mathbf{w}_{i+1}^m\right), g\left(\mathbf{v}_1^{i-1}, x', \tilde{\mathbf{w}}_{i+1}^m\right)\right)$$

$$D_{x,x'}^{(i)} = \min_{\mathbf{v}_1^{i-1}} D_{x,x'}^{(i)}\left(\mathbf{v}_1^{i-1}\right) \quad x, x' \in \{0,1\}^{m_i}$$

$$D_{\max}^{(i)} = \max_{x,x' \in \{0,1\}^{m_i}} D_{x,x'}^{(i)} \; ; \; D_{\min}^{(i)} = \min_{x,x' \in \{0,1\}^{m_i}, x \neq x'} D_{x,x'}^{(i)}$$

In order to distinguish between the partial distances of the two kernels, $g_1(\cdot)$ and $g_2(\cdot)$, we add an additional subscript to these parameters to indicate the kernel. For example, $D_{1,\min}^{(i)}$ and $D_{2,\min}^{(i)}$ denote the $i^{\text{th}}$ minimum partial distance of kernel $g_1(\cdot)$ and kernel $g_2(\cdot)$ respectively. We note here that for the binary kernels, we have $D_{\max}^{(i)} = D_{\min}^{(i)}$.

**Proposition 3** *There exist positive constants $c_1, c_2$ such that*

$$Z_{max}(W_{n+1}) \leq c_1 \cdot Z_{max}(W_n)^{\hat{D}_n} \tag{16}$$

$$Z_{min}(W_{n+1}) \geq c_2 \cdot Z_{min}(W_n)^{\breve{D}_n} \qquad n \geq 0, \tag{17}$$

*where $Z_{max}(W_n), Z_{min}(W_n)$ are the maximum and the minimum Bhattacharyya parameters of the channel $W_n$. $\left\{\hat{D}_n\right\}_{n\geq 0}, \left\{\breve{D}_n\right\}_{n\geq 0}$ sequences are defined as follows*

$$\hat{D}_n = D_{t,min}^{\tilde{\tau}(B_n)} \quad \breve{D}_n = D_{t,max}^{\tilde{\tau}(B_n)},$$

*where the parameter $t$, that indicates the kernel to which the partial distances refer to, equals 1, if $N_n = 1$, and otherwise equals 2. $\tilde{\tau}(\cdot)$ maps between the names of the channels and their ordinal numbers. For example, for $t = 1$, it gives $\tilde{\tau}(1) = 1$, $\tilde{\tau}((2,3)) = 2$ and $\tilde{\tau}(4) = 3$.*

**Proof** First, we note that for the quaternary input channel we have that [3, Corollary 18]

$$Z_{\max}\left(\tilde{\mathcal{W}}^{(\tilde{\tau}(i))}\right) \leq 4^{4-i} Z_{\max}\left(\tilde{\mathcal{W}}\right)^{D_{2,\min}^{(i)}} \quad i \in [4] \tag{18}$$

$$\frac{1}{4^{7-i}} Z_{\min}\left(\tilde{\mathcal{W}}\right)^{D_{2,\max}^{(i)}} \leq Z_{\min}\left(\tilde{\mathcal{W}}^{(\tilde{\tau}(i))}\right), \quad i \in [4], \tag{19}$$

where $\tilde{\tau}(i) = (2i-1, 2i)$, $Z_{\min}(\mathcal{W}) = \min_{x,x' \in \mathcal{X}, x \neq x'} Z_{x,x'}(\mathcal{W})$, and $Z_{\max}(\mathcal{W}) = \max_{x,x' \in \mathcal{X}} Z_{x,x'}(\mathcal{W})$. Note that if $|\mathcal{X}| = 2$, then $Z_{\max}(\mathcal{W}) = Z_{\min}(\mathcal{W}) = Z(\mathcal{W})$. Also note that for the binary kernel we have from [3, Corollary 18]

$$\frac{1}{2^6} Z(\mathcal{W})^{D_{1,\min}^{(1)}} \leq Z\left(\mathcal{W}^{(1)}\right) \leq 2^3 Z(\mathcal{W})^{D_{1,\min}^{(1)}}$$

$$\frac{1}{2^3} Z(\mathcal{W})^{D_{1,\min}^{(3)}} \leq Z\left(\mathcal{W}^{(4)}\right) \leq Z(\mathcal{W})^{D_{1,\min}^{(3)}} \tag{20}$$

Note that here because we deal with binary inputs $u_1, u_4$ to $g_1(\cdot)$ we have $D_{1,\min}^{(1)} = D_{1,\max}^{(1)}$ and $D_{1,\min}^{(3)} = D_{1,\max}^{(3)}$. Also note, that the difference between the indices of the channels and the distance parameter in (20) is because of the glued bits $u_{2,3}$.

We now need to consider the glued bits channel, which is the result of $u_{2,3}$, the second input of $g_1(\cdot)$. To do so we take similar derivations to [3, Lemma 21]. We assume that the input to this channel is quaternary, although the direct input for the channel is binary. Denote by $\mathcal{W}_{u_1}^{(2,3)}$ the channel assuming that $u_1$ was transmitted. We have

$$Z_{x,x'}\left(\mathcal{W}_{u_1}^{(2,3)}\right) = \sum_{\mathbf{y}} \sqrt{W_{u_1}^{(2,3)}(\mathbf{y}|x) W_{u_1}^{(2,3)}(\mathbf{y}|x')} =$$

$$= \frac{1}{2} \sum_{\mathbf{y}} \sqrt{\sum_{u_4 \in \{0,1\}} W(\mathbf{y}|g_1(u_1, x, u_4)) \sum_{v_4 \in \{0,1\}} W(\mathbf{y}|g_1(u_1, x', v_4))} \leq$$

$$= \frac{1}{2} \sum_{u_4 \in \{0,1\}} \sum_{v_4 \in \{0,1\}} \left( \sum_{\mathbf{y}} \sqrt{W(\mathbf{y}|g_1(u_1, x, u_4)) W(\mathbf{y}|g_1(u_1, x', v_4))} \right) \leq$$

$$\leq \frac{1}{2} \cdot 4 \cdot Z(\mathcal{W})^{D_{1,\min}^{(2)}}.$$

On the other hand,

$$Z_{x,x'}\left(\mathcal{W}_{u_1}^{(2,3)}\right) =$$

$$= \frac{1}{2} \sum_{\mathbf{y}} \sqrt{\sum_{u_4 \in \{0,1\}} W(\mathbf{y}|g_1(u_1, x, u_4)) \sum_{v_4 \in \{0,1\}} W(\mathbf{y}|g_1(u_1, x', v_4))} \geq$$

$$\geq \frac{1}{2} \max_{u_4, v_4 \in \{0,1\}} \left\{ \sum_{\mathbf{y}} \sqrt{W(\mathbf{y}|g_1(u_1, x, u_4)) W(\mathbf{y}|g_1(u_1, x', v_4))} \right\} \geq$$

$$\geq \frac{1}{2} Z(\mathcal{W})^{D_{\max}^{(2)}}.$$

Therefore, taking $c_1 = 4^3$ and $c_2 = 4^{-6}$ gives (16) and (17). ♢

Proposition 3 enables us to derive the asymptotic rate of polarization in the way that was done in [2] and [3].

**Proposition 4** *If $g_2(\cdot)$ is a polarizing kernel and $Z(\mathcal{W}) \neq 0$ then*

$$\lim_{n\to\infty} \Pr\left(Z_n \leq 2^{-4^{\beta n}}\right) = I(W), \quad \forall \beta < E_1(g_2). \tag{21}$$

$$\lim_{n\to\infty} \Pr\left(Z_n \geq 2^{-4^{\beta n}}\right) = 1, \quad \forall \beta > E_2(g_2), \tag{22}$$

*where $E_1(g_2) = 1/4 \sum_{i=1}^{4} \log_4\left(D_{2,min}^{(i)}\right)$ and $E_2(g_2) = 1/4 \sum_{i=1}^{4} \log_4\left(D_{2,max}^{(i)}\right)$. $E_1(g_2)$ and $E_2(g_2)$ are the lower bound and the upper bound (respectively) on the exponent of the kernel $g_2(\cdot)$.*

**Proof Idea** Taking the path of [7, Section E] enables us to prove (21), using the following two observations. (a) $\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \log\left(\hat{D}_i\right) = E_1(g_2)$ almost surely. (b) Conditioning on the value of the random variable $T$, the sequence $\left\{\hat{D}_n\right\}_{n\geq 1}$ is of independent samples. Adjusting the proof in [8, Section III] while using the fact that, almost surely, $\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \log\left(\breve{D}_i\right) = E_2(g_2)$ results in (22). For details, see the Appendix. $\diamondsuit$

# 4 General Mixed Kernels

The analysis that was done in Section 3 was for a specific dimension $\ell = 4$ and alphabet sizes 2 and 4. This technique can be generalized quite easily to general mixed kernel schemes. Let $g_1(v_1, v_2, ..., v_m)$ be equal to $g(\cdot)$ in (1). Denote the set of indices of the glued bits by $\mathcal{B} = \{i \in [m] | m_i \geq 2\}$. For each $i \in \mathcal{B}$ we supply a kernel $g_{i+1} : \left(\{0,1\}^{m_i}\right)^{\ell} \to \left(\{0,1\}^{m_i}\right)^{\ell}$ (by convention, if $m_i = m_j$ we usually take $g_{i+1}(\cdot) \equiv g_{j+1}(\cdot)$). We note that in [9, Table 5], the author gives a list of code decompositions that can be used for the definition of $g_1(\cdot)$. For the auxiliary kernels $g_{i+1}(\cdot)$, $i \in \mathcal{B}$ one can use non-binary kernels from [10].

The construction of larger dimension transform, $g^{(k)}\left(\mathbf{u}_1^{\ell^k}\right)$, can be done by a proper adjustment of the algorithm we suggested in Section 3, using the auxiliary kernels $g_i(\cdot)$ $i \in \mathcal{B}$ for the glued bits inputs of $g^{(k-1)}(\cdot)$.

A tree process, can also be defined in a similar fashion to Section 3.2. The probabilities for the choice of descendent channels for the first kernel are $\frac{m_i}{\ell}$ $i \in [m]$, and the probabilities for the channels induced by the kernels $g_i(\cdot)$ for $i \in \mathcal{B}$ are uniform. The random variable $T$ indicates the transition from the initial kernel $g_1(\cdot)$ to one of the $q$-ary kernels, where $q > 2$. Finally, the information sequence $I_n = \frac{I(W_n)}{N_n}$ is a bounded martingale. Generalizing Section 3.3 we are able to show that

**Proposition 5** *Assume that for all $i \in \mathcal{B}$, $g_{i+1}(\cdot)$ is a polarizing kernel, i.e. for a construction that is based only on $g_{i+1}(\cdot)$ we have that*

$$\lim_{n\to\infty} Pr\left(I\left(\tilde{W}_n\right)/m_i \in (\delta, 1-\delta)\right) = 0, \ \forall \delta \in (0, 0.5).$$

*As a result, the mixed kernel construction is also polarizing, i.e.*

$$\lim_{n\to\infty} \Pr\left(I_n \in (\delta, 1-\delta)\right) = 0, \quad \forall \delta \in (0, 0.5).$$

Let $E_1(g) = \min_{i\in\mathcal{B}} E_1\left(g_{i+1}\right)$ and $E_2(g) = \max_{i\in\mathcal{B}} E_2\left(g_{i+1}\right)$, where $E_1\left(g_{i+1}\right)$ and $E_2\left(g_{i+1}\right)$ are, respectively, the lower-bound and the upper-bound on the exponent assuming that we use only the kernel $g_{i+1}$.

**Proposition 6** *If for all $i \in \mathcal{B}$, $g_{i+1}(\cdot)$ is a polarizing kernel and $Z(\mathcal{W}) \neq 0$, then*

$$\lim_{n\to\infty} \Pr\left(Z_n \leq 2^{-\ell^{\beta n}}\right) = I(\mathcal{W}), \quad \beta < E_1(g) \tag{23}$$

$$\lim_{n\to\infty} \Pr\left(Z_n \geq 2^{-\ell^{\beta n}}\right) = 1, \quad \beta > E_2(g). \tag{24}$$

See the Appendix for details of the proof of Proposition 6.

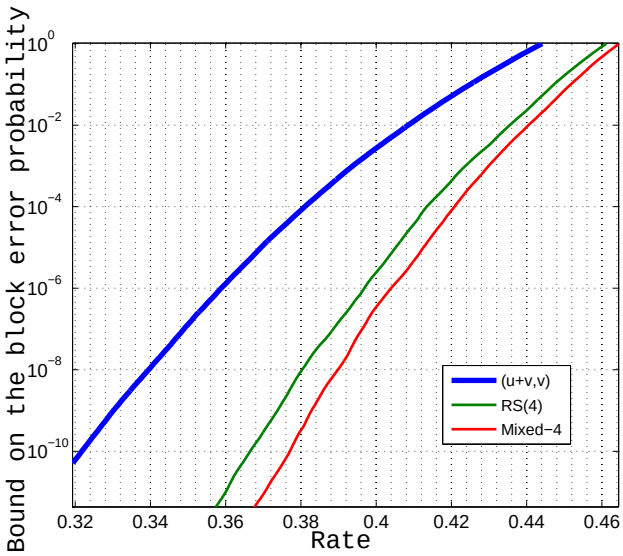


Figure 1: Upper bounds on the block error probability versus rate for three polar codes structures and SC decoding at block length $2^{14}$ bits on the BEC with erasure probability 0.5.

## 5 Simulations and Concluding Remarks

Proposition 6 implies that when considering the exponent as a measure of the polarization rate, the behavior of the mixed kernel is the same as the weakest from the auxiliary kernels. However, the exponent is an asymptotic measure and it may fail capturing the performance of a polar coding scheme for a finite block length $N$.

In Figure 1, we give results of density evolution computation over the Binary Erasure Channel (BEC) with erasure probability 0.5. Three polar codes with the same block length of $2^{14}$ bits are considered: $(u+v, v)$ is Arikan's binary polar code [1], $RS(4)$ is the extended Reed-Solomon construction considered in [3, Example 20] and *Mixed*-4, is the mixed kernel example from Section 3 (for the second kernel, $g_2(\cdot)$, we used $RS(4)$). To allow the $RS(4)$ scheme to have the same length of the other schemes, we took two $RS(4)$ transformations of length $2^{13}$ bits and applied on their outputs the quaternary $(u+v, v)$ transformation. The curves represent upper-bounds on the block error probability versus rate under SC decoding. The upper-bound here is a summation of the error probabilities of the split channels corresponding to the information set of the code. The information set for each curve point was determined using the technique of [1, Section V.D]. The figure demonstrates an advantage of the mixed kernel code in respect to the other candidates. We note, that as the theory predicts, the gap between the *Mixed*-4 and $RS(4)$ curves decreases for codes of lengths $4^n$ bits as $n$ grows.

We further note, that the mixed kernel scheme has an advantage over the $RS(4)$ in terms of decoding complexity. Computing the marginal probabilities of $g_2(\cdot)$ requires much more multiplications and summations in comparisons to the $g_1(\cdot)$ kernel. Therefore, decoding of a code based on the *Mixed*-4 kernel requires less multiplications and summations than would have been required for decoding a code of the same length based on the $RS(4)$ kernel.

## References

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

[2] S. B. Korada, E. Sasoglu, and R. Urbanke, "Polar codes: Characterization of exponent, bounds, and constructions," Jan. 2009. [Online]. Available: http://arxiv.com/abs/0901.0536

[3] R. Mori and T. Tanaka, "Performance and construction of polar codes on symmetric binary-input memoryless channels," in *Proc. IEEE Int. Symp. Information Theory ISIT 2009*, 2009, pp. 1496–1500.

[4] N. Presman, O. Shapira, and S. Litsyn, "Binary polar code kernels from code decompositions," Jan. 2011. [Online]. Available: http://arxiv.org/abs/1101.0764

[5] E. Sasoglu, E. Telatar, and E. Arikan, "Polarization for arbitrary discrete memoryless channels," Aug. 2009. [Online]. Available: http://arxiv.org/abs/0908.0302

[6] K. L. Chung, *A Course in Probability Theory*, 3rd ed. Academic Press, 2001.

[7] T. Tanaka and R. Mori, "Refined rate of channel polarization," Jan. 2010. [Online]. Available: http://arxiv.org/abs/1001.2067

[8] E. Arikan and E. Telatar, "On the rate of channel polarization," Jul. 2008. [Online]. Available: http://arxiv.com/abs/0807.3806

[9] S. Litsyn, *Handbook of Coding Theory*. Eds., Elsevier, The Netherlands, 1998, ch. An Updated Table of the Best Binary Codes Known.

[10] R. Mori and T. Tanaka, "Non-binary polar codes using reed-solomon codes and algebraic geometry codes," Jul. 2010. [Online]. Available: http://arxiv.org/abs/1007.3661

# A Appendix

## A.1 Proof of Proposition 4

We begin by proving the following auxiliary lemma.

**Lemma 1** *Let $\{Y_n\}_{n\geq 0}$ be a bounded sequence defined by*

$$Y_n = \begin{cases} Y_n^{(1)}, & n \leq T; \\ Y_n^{(2)}, & n > T. \end{cases}$$

*where $\left\{Y_n^{(1)}\right\}_{n\geq 0}$ and $\left\{Y_n^{(2)}\right\}_{n\geq 0}$ are i.i.d sequences, $T$ is a r.v. with the property*

$$\lim_{t\to\infty} \Pr(T > t) = 0, \tag{25}$$

*and $T$ is statistically independent of the sequence $Y_n^{(2)}$. Then,*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} Y_i = \underbrace{\mathbb{E}\left[Y_n^{(2)}\right]}_{\mu_2} \quad \textit{almost surely.}$$

**Proof** Using the strong law of large numbers we know that

$$\Pr\left(\omega \in \Omega;\, \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} Y_n^{(2)}(\omega) = \mu_2\right) = 1. \tag{26}$$

Let $\omega \in \Omega$ be such that $\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} Y_n^{(2)}(\omega) = \mu_2$. Assume that $T = t$, then

$$\frac{1}{n}\sum_{i=1}^{n} Y_i(\omega) = \frac{1}{n}\sum_{i=1}^{t} Y_i^{(1)}(\omega) + \frac{n-t}{n}\frac{1}{n-t}\sum_{i=t+1}^{n} Y_i^{(2)}(\omega), \tag{27}$$

Now, $\frac{1}{n}\sum_{i=1}^{t} Y_i^{(1)}(\omega)$ surely goes to zero as $n$ grows, and $\frac{1}{n-t}\sum_{i=t}^{n} Y_i^{(2)}(\omega)$ goes to $\mu_2$ because of the choice of $\omega$ (remember that $T$ and the sequence $Y_n^{(2)}$ are independent). This means that

$$\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} Y_n(\omega) = \mu_2. \tag{28}$$

Therefore,

$$\left\{\omega\in\Omega;\ \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} Y_n^{(2)}(\omega)=\mu_2\right\} \subseteq \left\{\omega\in\Omega;\ \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} Y_n(\omega)=\mu_2\right\},$$

so

$$\Pr\left(\omega\in\Omega;\ \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} Y_n(\omega)=\mu_2\right) \geq \Pr\left(\omega\in\Omega;\ \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} Y_n^{(2)}(\omega)=\mu_2\right) = 1.$$

♢

### A.1.1 Proof of (21)

To prove (21) we make the following adjustments to the proofs of Mori and Tanaka in [3, Proposition 15]. We attempted to give a comprehensive version of their proofs at Subsection A.3 ([3] doesn't contain proofs for this statement). There are two parts of the proof. First we should prove that for an arbitrary fixed $\rho\in(0,1)$,

$$\lim_{n\to\infty}\Pr\left(Z_n\leq\rho^n\right)=\Pr(Z_\infty=0).$$

The key method in the proof is using the law of large numbers, on the sequence $\hat{D}_n$. This is applicable here because of Lemma 1. This leads to the observation that for any $0\leq\beta<1/4\sum_{i=1}^{4}\log_4\left(D_{2,\min}^{(i)}\right)$, for the proper choice of $\rho,\gamma$ and large enough $n$ (see Subsection A.3 below for the definitions of $D_n(\beta)$, $\mathcal{G}_{m,n}(\gamma)$ and $\mathcal{C}_m(\rho)$)

$$\Pr\left(D_n(\beta)\right)\geq\Pr\left(\mathcal{G}_{\alpha n,n}(\gamma)\bigcap\mathcal{C}_{\alpha n}(\rho)\right).$$

In the original proof, the next step is to claim that the two events at the right side of the equation are independent which results in (53). But this doesn't apply here, because the sequence $\left\{\hat{D}_n\right\}_{n\geq0}$ is not of independent samples. However, conditioning on the event $T<\alpha\cdot n$ the two events are independent. We have that

$$\Pr\left(D_n(\beta)\right)\geq$$

$$\Pr\left(\mathcal{G}_{\alpha n,n}(\gamma)\,|T<\alpha n\right)\cdot\Pr\left(\mathcal{C}_{\alpha n}(\rho)|T<\alpha n\right)\Pr\left(T<\alpha n\right). \tag{29}$$

This leads us to

$$\lim_{n\to\infty}\Pr\left(\mathcal{D}_n(\beta)\right)\geq$$

$$\lim_{n\to\infty}\underbrace{\Pr\left(\mathcal{G}_{\alpha n,n}(\gamma)\,|T<\alpha n\right)}_{\to1}\cdot\underbrace{\Pr\left(\mathcal{C}_{\alpha n}(\rho)|T<\alpha n\right)\Pr\left(T<\alpha n\right)}_{\to\Pr(Z_\infty=0)}=$$

$$=\Pr(Z_\infty=0)$$

♢

### A.1.2 Proof of (22)

We take the path of the proof of the converse part of [8, Theorem 3]. We consider the $Z_{\min}(W_n)$ sequence and define the sequence $\left\{\tilde{Z}_n\right\}_{n\geq0}$ in the following way

$$\tilde{Z}_0=Z_{\min}(W)=Z(\mathcal{W})$$

$$\tilde{Z}_{n+1} = c_2 \cdot \tilde{Z}_n^{\breve{D}_n} \quad n \geq 0.$$

Note that $Z_{\min}(W_n) \geq \tilde{Z}_n$, and therefore

$$\Pr\left(Z_n \geq \delta_n\right) \geq \Pr\left(Z_{\min}(W_n) \geq \delta_n\right) \geq \Pr\left(\tilde{Z}_n \geq \delta_n\right). \tag{30}$$

By the definition of $\tilde{Z}_n$ we have that,

$$\tilde{Z}_n = (c_2)^n \, \tilde{Z}_0^{\prod_{i=0}^{n-1} \breve{D}_i} \tag{31}$$

Assume that $c_2 < 1$, otherwise we can replace it by $c_3 < c_2$ such that $c_3 < 1$.

$$\log_2\left(\tilde{Z}_n\right) = n \log_2\left(c_2\right) + \log_2\left(\tilde{Z}_0\right) \cdot \prod_{i=0}^{n-1} \breve{D}_i, \tag{32}$$

For large enough $n$ we have

$$\log_l\left(-\log_2\left(\tilde{Z}_n\right)\right) =$$

$$= \log_l\left(n \log_2\left(c_2^{-1}\right) + \log_2\left(\tilde{Z}_0^{-1}\right) \cdot \prod_{i=0}^{n-1} \breve{D}_i\right) \leq \tag{33}$$

$$\log_l\left(n \log_2\left(c_2^{-1}\right)\right) + \log_l\left(\log_2\left(\tilde{Z}_0^{-1}\right) \cdot \prod_{i=0}^{n-1} \breve{D}_i\right) =$$

$$n\left(o(1) + \frac{1}{n}\sum_{i=0}^{n-1} \log_\ell\left(\breve{D}_i\right)\right).$$

This results in

$$\Pr\left(\tilde{Z}_n \geq \delta_n\right) \geq$$

$$\Pr\left(o(1) + \frac{1}{n}\sum_{i=0}^{n-1} \log_\ell\left(\breve{D}_i\right) \leq \frac{1}{n}\log_\ell\left(-\log_2\left(\delta_n\right)\right)\right). \tag{34}$$

Now, set $\ell = 4$, $\delta_n = 2^{-4^{\beta n}}$ and $\beta > 1/4 \sum_{i=1}^{4} \log_4\left(D_{2,\max}^{(i)}\right)$ in (34). We have, by Lemma 1, that $\frac{1}{n}\sum_{i=0}^{n-1} \log_\ell\left(\breve{D}_i\right)$ goes almost surely to $1/4 \sum_{i=1}^{4} \log_4\left(D_{2,\max}^{(i)}\right)$, therefore

$$\lim_{n\to\infty} \Pr\left(o(1) + \frac{1}{n}\sum_{i=0}^{n-1} \log_\ell\left(\breve{D}_i\right) \leq \frac{1}{n}\log_\ell \log_2\left(-\delta_n\right)\right) = 1.$$

$\diamondsuit$

## A.2 Proof of Proposition 6

First, a generalization of Proposition 3 can be stated.

**Proposition 7** *There exist positive constants $c_1, c_2$ such that*

$$Z_{max}(W_{n+1}) \leq c_1 \cdot Z_{max}(W_n)^{\hat{D}_n} \tag{35}$$

$$Z_{min}(W_{n+1}) \geq c_2 \cdot Z_{min}(W_n)^{\breve{D}_n} \qquad n \geq 0, \tag{36}$$

*where $Z_{max}(W_n), Z_{min}(W_n)$ are the maximum and the minimum Bhattacharyya parameters of the channel $W_n$. $\left\{\hat{D}_n\right\}_{n\geq 0}, \left\{\breve{D}_n\right\}_{n\geq 0}$ sequences are defined as follows*

$$\hat{D}_n = D_{t,min}^{\tilde{\tau}(B_n)} \quad \breve{D}_n = D_{t,max}^{\tilde{\tau}(B_n)},$$

*where the parameter $t$ indicates the kernel to which the partial distances refer to, i.e. $t \in \mathcal{B} \bigcup \{1\}$. $t$ equals 1, if $N_n = 1$, and otherwise it equals $j$, if $g_j(\cdot)$ is the auxiliary kernel for the alphabet of size $2^{N_n}$. $\tilde{\tau}(\cdot)$ maps between the names of the channels and their ordinal numbers.*

Using Proposition 7, we can prove Proposition 6 in a similar fashion to the proofs of Proposition 4, but instead of using Lemma 1, we use Lemma 2 below. Specifically, for proving (24) we take the same steps we used in the proof of (21), but use the fact that $\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\log_\ell \hat{D}_i \geq \delta_1 \cdot E_1(g)$ for each $\delta_1 \in (0,1)$, almost surely, by Lemma 2. For proving (23) we take same steps we used in the proof of (22), but use the fact that $\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\log_\ell \breve{D}_i \leq (1+\delta_2)\cdot E_2(g)$, for each $\delta_2 \in (0,1)$, almost surely, by Lemma 2.

**Lemma 2** *Let $J$ be a random variable having values from a finite set of numbers $\mathcal{J}$ ($1 \notin \mathcal{J}$). Let $\{Y_n\}_{n\geq 0}$ be a bounded sequence defined by*

$$Y_n = \begin{cases} Y_n^{(1)}, & n \leq T; \\ Y_n^{(J)}, & n > T. \end{cases}$$

*where $\left\{Y_n^{(1)}\right\}_{n\geq 0}$ and $\left\{Y_n^{(j)}\right\}_{n\geq 0}$ $j \in \mathcal{J}$ are i.i.d sequences, $T$ is a r.v. with the property*

$$\lim_{t\to\infty} \Pr(T > t) = 0, \tag{37}$$

*and $T$ and $J$ are statistically independent of the sequences $Y_n^{(j)}$ $j \in \mathcal{J}$. Let $\mu_j = \mathbb{E}\left[Y_n^{(j)}\right]$, and*

$$\mu_{min} = \min_{j\in\mathcal{J}} \mu_j \qquad \mu_{max} = \max_{j\in\mathcal{J}} \mu_j$$

*Then, almost surely, $\frac{1}{n}\sum_{i=1}^{n} Y_i$ converges to a number $\mu$ where $\mu \in \{\mu_j | j \in \mathcal{J}\}$. Specifically, this means that, $\forall \delta_1, \delta_2 \in (0,1)$, as $n \to \infty$, almost surely*

$$\delta_1 \cdot \mu_{min} \leq \frac{1}{n}\sum_{i=1}^{n} Y_i \leq (1+\delta_2)\cdot \mu_{max}.$$

## A.3 Proof of [3, Proposition 15]

We state a slight variation on the first part of [3, Proposition 15].

**Proposition 8** *Let $\{X_n \in (0,1)\}$ be a random process satisfying the following properties. 0) For each $n$, $X_{n+1} = f_{i,n}(X_n)$ w.p. $\frac{1}{\ell}$ $i \in [\ell]$ independently in the value of $X_n$, where $\{f_{i,n}(\cdot)\}_{i\in[\ell],n\geq 0}$ is a sequence of deterministic functions. 1) $X_n$ converges to $X_\infty \in \{0,1\}$ almost surely. 2) There exists a positive constant $q$ such that $f_{i,n}(x) \leq q \cdot x^{d_i}$ $\forall x \in (0,1), n \geq 0, i \in [\ell], d_i \geq 1$.*

*Then,*

$$\lim_{n\to\infty} \Pr\left(X_n < 2^{-\ell^{\beta n}}\right) = \Pr\left(X_\infty = 0\right),$$

*for each $\beta < \Lambda$, where $\Lambda = \frac{1}{\ell}\sum_{i=1}^{\ell}\log_\ell(d_i)$.*

**Proof** We first define a random sequence $\{D_n \geq 1\}$ that takes values from $\{d_i | i \in [\ell]\}$ correspondingly to properties 0 and 2. This means that this sequence is an i.i.d sequence and $X_{n+1} \leq q \cdot X_n^{D_n}$. Furthermore $\Lambda = \mathbb{E}\left[\log_\ell(D_n)\right]$ and $\tilde{\Lambda} = \mathbb{E}\left[D_n\right]$.

We take the path of [7] by first giving the following definitions. For $m < n$ natural numbers define

$$S_{m,n} = \sum_{i=m}^{n-1} \log_\ell(D_i) \tag{38}$$

$$\tilde{S}_{m,n} = \sum_{i=m}^{n-1} D_i \tag{39}$$

**Definition 3** *For a fixed $\gamma \in [0,1)$, let $\mathcal{G}_{m,n}(\gamma)$ and $\tilde{\mathcal{G}}_{m,n}(\gamma)$ be the events defined by*

$$\mathcal{G}_{m,n}(\gamma) = \left\{\frac{1}{n-m} S_{m,n} \geq \gamma \cdot \Lambda\right\}$$

$$\tilde{\mathcal{G}}_{m,n}(\gamma) = \left\{\frac{1}{n-m} \tilde{S}_{m,n} \geq \gamma \cdot \tilde{\Lambda}\right\}.$$

**Definition 4** *Let $\rho, \beta \in (0,1)$. The events $\mathcal{C}_n(\rho)$ and $\mathcal{D}_n(\beta)$ are defined as*

$$\mathcal{C}_n(\rho) = \{X_n \leq \rho^n\} \tag{40}$$

$$\mathcal{D}_n(\beta) = \left\{X_n \leq 2^{-\ell^{n\beta\Lambda}}\right\} \tag{41}$$

We prove that the event $\mathcal{C}_n(\rho)$ has probability arbitrarily close to $\Pr(X_\infty = 0)$ as $n \to \infty$. This is used to prove that the event $\mathcal{D}_n(\beta)$ has a probability arbitrarily close to $\Pr(X_\infty = 0)$ as $n \to \infty$.

**Proposition 9** *For an arbitrary fixed $\rho \in (0,1)$*

$$\lim_{n\to\infty} \Pr(\mathcal{C}_n(\rho)) = \Pr(X_\infty = 0).$$

**Proof** As mentioned in [7], the proof is similar to [1, Theorem 2]. We now elaborate on it. By the law of large numbers we have

$$\lim_{n-m\to\infty} \Pr\left(\tilde{\mathcal{G}}_{m,n}(\gamma)\right) = 1 \quad \gamma \in [0,1) \tag{42}$$

Now define for $\alpha \in (0,1)$,

$$\mathcal{T}_m(\alpha) = \{\omega \in \Omega | X_i \leq \alpha \quad , \forall i \geq m\},$$

obviously,

$$\lim_{m\to\infty} \Pr\left(\mathcal{T}_m(\alpha)\right) \geq \Pr\left(X_\infty = 0\right). \tag{43}$$

Now, for $\omega \in \mathcal{T}_m(\alpha)$

$$\frac{X_{i+1}(\omega)}{X_i(\omega)} \leq \alpha^{D_i - 1},$$

therefore

$$X_n(\omega) = X_m(\omega) \cdot \prod_{i=m}^{n-1} \frac{X_{i+1}(\omega)}{X_i(\omega)} \leq \alpha^{(n-m)\left(\frac{1}{n-m}\tilde{S}_{n,m}-1\right)} =$$

$$= \left(\alpha^{\frac{1}{n-m}\tilde{S}_{n,m}-1}\right)^{n-m}.$$

By the law of large numbers (see Remark 1), for each $\gamma, \delta \in (0,1)$ there exists $n_0$ such that for each $n > n_0$

$$\Pr(\tilde{\mathcal{G}}_{m,n}(\gamma)) = \Pr\left(\alpha^{\frac{1}{n-m}\tilde{S}_{n,m}-1} \leq \alpha^{\gamma\tilde{\Lambda}-1}\right) \geq 1 - \delta/2 \tag{44}$$

Also, because of (43), for each $\delta, \alpha$ there exists $m_0$, such that, for each $m > m_0$

$$\Pr\left(\mathcal{T}_m(\alpha)\right) \geq \Pr\left(X_\infty = 0\right) - \delta/2 \tag{45}$$

Now, if we take $\alpha$ such that $\alpha^{\gamma\tilde{\Lambda}-1} < \rho - \epsilon$, then there exists an $n_1$ such that for each $n > n_1$

$$\left(\alpha^{\gamma\tilde{\Lambda}-1}\right)^n \cdot \left(\alpha^{\gamma\tilde{\Lambda}-1}\right)^{-m} < \rho^n$$

This means that for $n > \max\{n_0, n_1\}$

$$\Pr\left(\mathcal{T}_{m_0}(\alpha) \bigcap X_n \leq \rho^n\right) \geq \Pr\left(X_\infty = 0\right) - \delta.$$

Therefore,

$$\Pr\left(X_n < \rho^n\right) \geq \Pr\left(X_\infty = 0\right).$$

Because $0 < \rho < 1$ it means that (we assume that $X_\infty \in \{0,1\}$)

$$\Pr\left(X_n < \rho^n\right) \leq \Pr\left(X_n < 1\right) \to \Pr\left(X_\infty = 0\right),$$

as $n \to \infty$ almost surely.

so, finally

$$\lim_{n\to\infty} \Pr\left(X_n < \rho^n\right) = \Pr\left(X_\infty = 0\right).$$

$\diamondsuit$

**Remark 1** *Given the event $T_m(\alpha)$ the sequence $\{D_i\}_{i\geq m}$ is not necessarily i.i.d anymore.*

However in (44) we still want to use the law of large numbers. To do so, take $\alpha$ such that $\alpha < q^{-1}$. This means that if $X_n \leq \alpha$, then $X_{n+1} \leq q \cdot X_n^{D_n}$, which means that $X_{n+1} \leq \alpha^{D_n - 1}$, so in case $D_n \geq 2$, we have that $X_{n+1} \leq \alpha$. For $D_n = 1$ it may happen that $X_{n+1} > \alpha$. This means that for the sequence $D_n$,

$$\Pr\left(D_n \geq 2 \middle| T_m(\alpha), X_n = x\right) \geq \Pr\left(D_n \geq 2\right)$$

$$\Pr\left(D_n = 1 \middle| T_m(\alpha), X_n = x\right) \leq \Pr\left(D_n = 1\right)$$

Construct a sequence $\tilde{D}_n$ in the following way. Given $X_n = x$ and $T_m(\alpha)$, we have

$$\tilde{D}_n = \begin{cases} D_n, & \text{w.p. } 1 - \pi(n,x); \\ 1, & \text{w.p. } \pi(n,x). \end{cases}$$

where $\pi(n,x)$ is chosen such that

$$\Pr\left(\tilde{D}_n = 1 \middle| T_m(\alpha), X_n = x\right) = \Pr\left(D_n = 1\right).$$

Note, that in this case

$$\Pr\left(\tilde{D}_n = i \middle| T_m(\alpha), X_n = x\right) = \Pr\left(D_n = i\right) \quad \forall i, x$$

Which means that $\tilde{D}_n$ is an i.i.d sequence that distributes like the sequence $D_n$. We also have that $\tilde{D}_n \leq D_n$, so

$$X_n(\omega) = X_m(\omega) \prod_{i=m}^{n-1} \frac{X_{i+1}(\omega)}{X_i(\omega)} \leq \tag{46}$$

$$\alpha^{n-m\left(\frac{1}{n-m}\tilde{S}_{n,m}-1\right)} \underbrace{\leq}_{\alpha<1} \left(\alpha^{\frac{1}{n-m}\sum_{i=m}^{n-1}\tilde{D}_i - 1}\right)^{n-m}.$$

Now, we can use the law of large numbers on the right side of (46).

Now, we follow the bootstrapping idea that was presented in [8] and used again in [7]. The idea is that for some $m << n$, once a realization of $X_m$ becomes sufficiently small, one can assure with probability close to 1, that samples conditionally generated on the realization of $X_m$ will converge to 0, exponentially fast. We follow the steps of [7, Section 4.E]. Define a process $\{L_i\}$ using the process $\{X_i\}$ as follows for fixed $m$.

$$L_i = \log_2 X_i \quad i = 0, 1, 2..., m \tag{47}$$

$$L_{i+1} = D_i \cdot L_i + \underbrace{\log_2(q)}_{\zeta} \quad i \geq m. \tag{48}$$

The inequality $X_i \leq 2^{L_i}$ holds for this sample basis for all $i \geq 0$. We have that

$$\begin{aligned} L_n &= D_{n-1} \cdot L_{n-1} + \zeta = \\ &= D_{n-1} \cdot (D_{n-2} \cdot L_{n-2} + \zeta) + \zeta = \dots \\ \dots &= L_m \prod_{i=m}^{n-1} D_i + \zeta \cdot \sum_{j=m+1}^{n-1} \prod_{r=j}^{n-1} D_r = \\ &= \prod_{i=m}^{n-1} D_i \cdot \left(L_m + \frac{\zeta \cdot \sum_{j=m+1}^{n-1} \prod_{r=j}^{n-1} D_r}{\prod_{i=m}^{n-1} D_i}\right) = \\ &= \prod_{i=m}^{n-1} D_i \cdot \left(L_m + \zeta \cdot \sum_{j=m+1}^{n-1} \left(\prod_{r=m}^{j-1} D_r\right)^{-1}\right) \leq \\ &\prod_{i=m}^{n-1} D_i \cdot (L_m + \zeta \cdot (n-m)) = \ell^{S_{m,n}} \left(L_m + \zeta \cdot (n-m)\right). \end{aligned} \tag{49}$$

**Lemma 3** *Fix $\gamma \in (0,1)$ and $\epsilon > 0$ and let $\rho$ be such that $\log_2(\rho) = -\left(\epsilon + \zeta \frac{n-m}{m}\right)$ holds. Then, conditional on $\mathcal{C}_m(\rho) \bigcap \mathcal{G}_{m,n}(\gamma)$ one has from (49)*

$$L_n \leq -\ell^{\gamma \cdot \Lambda \cdot (n-m)} \cdot \epsilon \cdot m \tag{50}$$

**Proposition 10** *For an arbitrary $\beta \in (0, \Lambda)$, we have*

$$\lim_{n\to\infty} \Pr\left(\mathcal{D}_n(\beta)\right) = \Pr\left(X_\infty = 0\right)$$

**Proof** Given $\beta \in (0,\Lambda)$, choose $\gamma, \alpha \in (0,1)$ such that $\frac{\beta}{\Lambda} = \gamma \cdot (1-\alpha)$. Take $m = \alpha n$, where $\alpha \in (0,1)$, and let $\{L_i\}$ denote the process defined in (47) and (48) with respect to $m = \alpha n$. Then for any $\epsilon > 0$ using Lemma 3, conditional on the event $\mathcal{C}_{\alpha n}(\rho) \bigcap \mathcal{G}_{\alpha n,n}(\gamma)$ (using $\rho$ as defined in the Lemma) we have the inequality

$$L_n \leq -\ell^{\gamma \cdot \Lambda \cdot (1-\alpha) n} \cdot \epsilon \cdot \alpha n = -\ell^{\beta n} \cdot \epsilon \cdot \alpha n$$

This means that

$$\left\{X_n \leq 2^{-\ell^{\beta n} \cdot \epsilon \cdot \alpha n}\right\} \supseteq \mathcal{C}_{\alpha n}(\rho) \bigcap \mathcal{G}_{\alpha n,n}(\gamma). \tag{51}$$

For any $n \geq (\epsilon\alpha)^{-1}$, $\beta \cdot n \leq \overbrace{\gamma(1-\alpha)\Lambda}^{\beta} n + \log_\ell(\epsilon\alpha n)$. Therefore,

$$\mathcal{D}_n(\beta) \supseteq \left\{X_n \leq 2^{-\ell^{\gamma \cdot \Lambda \cdot (1-\alpha) n} \cdot \epsilon \cdot \alpha n}\right\} \tag{52}$$

So using (52) and the independence of $\mathcal{C}_{\alpha n}(\rho)$ and $\mathcal{G}_{\alpha n,n}(\gamma)$ we have that

$$\Pr\left(D_n(\beta)\right) \geq \Pr\left(\mathcal{G}_{\alpha n,n}(\gamma)\right) \cdot \Pr\left(\mathcal{C}_{\alpha n}(\rho)\right). \tag{53}$$

Hence, using Proposition 9 we have

$$\lim_{n\to\infty} \Pr\left(\mathcal{D}_n(\beta)\right) \geq$$
$$\lim_{n\to\infty} \underbrace{\Pr\left(\mathcal{G}_{\alpha n,n}(\gamma)\right)}_{\to \Pr(X_\infty = 0)} \cdot \underbrace{\Pr\left(\mathcal{C}_{\alpha n}(\rho)\right)}_{\to 1} = \Pr(X_\infty = 0)$$

♢