Binary Linear Codes with Optimal Scaling: Polar Codes with Large Kernels

Arman Fazeli
University of California San Diego
9500 Gilman Drive, La Jolla, CA 92093
afazelic@ucsd.edu

Hamed Hassani
University of Pennsylvania
3101 Walnut St, Philadelphia, PA 19104
hassani@seas.upenn.edu

Marco Mondelli
Institute of Science and Technology Austria (IST Austria)
Am Campus 1, 3400 Klosterneuburg, Austria
marco.mondelli@ist.ac.at

Alexander Vardy
University of California San Diego
9500 Gilman Drive, La Jolla, CA 92093
avardy@ucsd.edu

October 15, 2020

Abstract

We prove that, for the binary erasure channel (BEC), the polar-coding paradigm gives rise to codes that not only approach the Shannon limit but do so under the *best possible scaling of their block length* as a function of the gap to capacity. This result exhibits the first known family of binary codes that attain both optimal scaling and quasi-linear complexity of encoding and decoding. Our proof is based on the construction and analysis of binary polar codes *with large kernels*. When communicating reliably at rates within $\varepsilon > 0$ of capacity, the code length n often scales as $O(1/\varepsilon^{\mu})$, where the constant μ is called the *scaling exponent*. It is known that the optimal scaling exponent is $\mu = 2$, and it is achieved by random linear codes. The scaling exponent of conventional polar codes (based on the 2×2 kernel) on the BEC is $\mu = 3.63$. This falls far short of the optimal scaling guaranteed by random codes. Our main contribution is a rigorous proof of the following result: for the BEC, there exist $\ell \times \ell$ binary kernels, such that polar codes constructed from these kernels achieve scaling exponent $\mu(\ell)$ that tends to the optimal value of 2 as ℓ grows. We furthermore characterize precisely how large ℓ needs to be as a function of the gap between $\mu(\ell)$ and 2. The resulting binary codes maintain the recursive structure of conventional polar codes, and thereby achieve construction complexity $O(n)$ and encoding/decoding complexity $O(n \log n)$.

*The research of Arman Fazeli and Alexander Vardy was supported in part by the United States National Science Foundation under Grants CCF-1405119 and CCF-1719139. Marco Mondelli was supported in part by an Early Postdoc Mobility Fellowship from the Swiss National Science Foundation.

1. Introduction

Shannon's coding theorem implies that for every binary-input memoryless symmetric (BMS) channel W, there is a capacity $I(W)$ such that the following holds: for all $\varepsilon > 0$ and $P_e > 0$, there exists a binary code of rate at least $I(W) - \varepsilon$ which enables communication over W with probability of error at most P_e. Ever since the publication of Shannon's famous paper [1], the holy grail of coding theory was to find explicit codes that achieve Shannon capacity with polynomial-time complexity of construction and decoding. Today, several such families of codes are known, and the principal remaining challenge is to characterize *how fast we can approach capacity* as a function of the code block length n. Specifically, we now have explicit binary codes (which can be constructed and decoded in polynomial time) of length n and rate R, such that the gap to capacity $\varepsilon = I(W) - R$ required to achieve any fixed error probability $P_e > 0$ vanishes as a function of n. The fundamental theoretical problem is to characterize how fast this happens. Equivalently, we can fix $\varepsilon = I(W) - R$ and ask how large does the block length n need to be as a function of ε. That is, we are interested in the *scaling between the block length and the gap to capacity*, under the constraint of polynomial-time construction and decoding.

Based on [2], it is known that the optimal scaling is of the form $n = O(1/\varepsilon^{\mu})$, where the constant μ is referred to as the *scaling exponent*. It is furthermore known that the best possible scaling exponent is $\mu = 2$, and it is achieved by random linear codes — although, of course, random codes do not admit efficient decoding. In this paper, we present the first family of binary codes that attain both optimal scaling and quasi-linear complexity on the binary erasure channel (BEC). Specifically, for any fixed $\delta > 0$, we exhibit codes that ensure reliable communication on the BEC at rates within $\varepsilon > 0$ of the Shannon capacity, with block length $n = O(1/\varepsilon^{2+\delta})$, construction complexity $O(n)$, and encoding/decoding complexity $O(n \log n)$.

To establish this result, we use *polar coding*, invented by Arıkan [3] in 2009. However, while Arıkan's polar codes are based upon a specific 2×2 kernel, we use $\ell \times \ell$ binary polarization kernels, where ℓ is a sufficiently large constant. The main technical challenge is to prove that this construction works. To this end, we choose the polarization kernel uniformly at random from the set of all $\ell \times \ell$ nonsingular binary matrices, and show that with probability at least $1 - O(1/\ell)$, the resulting scaling exponent is at most $2 + \delta$. Since ℓ is a constant that depends only on δ, the choice of a polarization kernel can be, in principle, de-randomized using brute-force search whose complexity is independent of the block length.

By way of a disclaimer, the theorems in this paper require the size of the kernel to be *extremely large* to the point that they are not practical at all. Moreover, the decoding complexity of polar codes constructed with large kernels, if done naively and over arbitrary channels, scales *exponentially* with the kernel size, which is another challenge in using these codes. On the positive side, we prove that, for sufficiently large values of ℓ, *almost all* kernels yield fast scaling exponents. This establishes a new family of binary error-correcting codes that are *near optimal* in every aspect at the asymptotic regime; this is of purely theoretical interest. The problem of finding good polarization kernels with reasonable size and decoding complexity remains unsolved. We will review some recent advancements in the field on this problem in the final section.

The rest of this paper is organized as follows. In this section, we provide the necessary background and give an informal statement of our main result (Theorem 1). In Section 2, we present a brief outline of the proof. In Section 3, we formally state our main theorem (Theorem 2), and gradually reduce its proof to a certain statement about $\ell \times \ell$ binary polarization kernels (Theorem 6). We defer the proof of Theorem 6 itself, which is technically quite elaborate, to Section 4. We conclude with a brief discussion in Section 5.

1.1. Background and context

A sequence of papers, starting with [4,5] in 1960s and culminating with [2,6], shows that for any discrete memoryless channel W and *any* code of length n and rate R that achieves error-probability P_e on W, we have

$$I(W) - R \geqslant \frac{\mathrm{const}(P_e, W)}{\sqrt{n}} - O\left(\frac{\log n}{n}\right), \tag{1}$$

where the constant (which is given explicitly in [2]) depends on W and P_e, but not on n. This immediately implies that if $n = O(1/\varepsilon^{\mu})$, where $\varepsilon = I(W) - R$ is the gap to capacity, then $\mu \geqslant 2$. We further note that expressions similar to (1) were derived from the perspective of threshold phenomena in [7] and from the perspective of statistical physics in [8]. The fact that $\mu \geqslant 2$ also follows from a simple heuristic argument. For simplicity, consider the special case of transmission over the BEC with erasure probability p. As $n \to \infty$, the number of erasures will tend to the normal distribution with mean np and standard deviation $\sqrt{np(1-p)}$. Thus, channel randomness yields a variation in the fraction of erasures of order $1/\sqrt{n}$. This indicates that, in order to achieve a fixed error probability, the gap to capacity ε has to scale at least as $1/\sqrt{n}$.

It is well known [2,6] that the lower bound $\mu = 2$ is achieved by random linear codes. For the special case of transmission over the BEC, the proof of this fact reduces to computing the rank of a certain random matrix. Indeed, the generator matrix of a random linear code of length n and rate R is a matrix with Rn rows and n columns whose entries are i.i.d. uniform in $\{0,1\}$. The effect of transmission over the BEC with erasure probability p is equivalent to removing each column of this generator matrix independently with probability p. The probability of error (under maximum-likelihood decoding) is thus equal to the probability that such residual matrix is not full-rank. This probability is easy to compute, and the desired scaling result immediately follows.

Unfortunately, random linear codes cannot be decoded efficiently. On general BMS channels, this task is NP-hard [9]. On the BEC, decoding a general binary linear code takes time $O(n^{\omega})$, where ω is the exponent of matrix multiplication. This leads to the following natural question: what is the lowest possible scaling exponent for binary codes that can be constructed, encoded, and decoded efficiently? For the BEC, we take *efficiently* to mean linear or quasi-linear complexity. Here is a brief survey of the current state of knowledge on this question.

Forney's concatenated codes [10] are a classical example of a capacity-achieving family of codes. However, their construction and decoding complexity are exponential in the inverse gap to capacity $1/\varepsilon$ (see [11,12] for more details), so they are not competitive from an asymptotic perspective. In recent years, three new families of capacity-achieving codes have been discovered; let us review what is known regarding their scaling exponents.

Polar codes: Achieve the capacity of any BMS channel under a successive-cancellation decoding algorithm [3] that runs in time $O(n \log n)$. It was shown in [11,12] that the block length, construction complexity, and decoding complexity are all bounded by a polynomial in $1/\varepsilon$, which implies that the scaling exponent μ is finite. Later, a sequence of papers [14–17] provided rigorous upper and lower bounds on μ. The specific value of μ depends on the channel W. It is known that $\mu = 3.63$ on the BEC. The best-known bounds valid for any BMS channel W are given by $3.579 \leqslant \mu \leqslant 4.714$.

Spatially-coupled LDPC codes: Achieve the capacity of any BMS channel under a belief-propagation decoding algorithm [18] that runs in linear time. A simple heuristic argument yields that the scaling exponent of these codes is roughly 3 (see [19, Section VI-D]). However, a rigorous proof of this statement remains elusive and appears to be technically challenging.

Reed-Muller codes: Achieve capacity of the BEC under maximum-likelihood decoding [20,21] that runs in time $O(n^{\omega})$. While it has been observed empirically that the performance of Reed-Muller codes on the BEC is close to that of random codes [22], no bounds on the scaling exponent of these codes are known.

Let us point out that some papers also define a "scaling exponent" for codes that do not achieve capacity, such as ensembles of LDPC codes, by substituting the specific threshold of the ensemble for channel capacity. In this context, it is known [23] that for a large class of ensembles of LDPC codes and channel models, the scaling exponent is $\mu = 2$. However, the threshold of such LDPC ensembles does not converge to capacity.

1.2. Our main result: Binary linear codes with optimal scaling and quasi-linear complexity

Our main result provides the first family of binary codes for transmission over the BEC that achieves optimal scaling between the gap to capacity ε and the block length n, and that can be constructed, encoded, and decoded in quasi-linear time. In other words, the block length, construction, encoding, and decoding complexity are all bounded by a polynomial in $1/\varepsilon$ and, moreover, the degree of this polynomial approaches the information-theoretic lower bound $\mu \geqslant 2$. Somewhat informally (cf. Theorem 2), this result can be stated as follows.

Theorem 1. *Consider transmission over i.i.d. copies of a binary erasure channel W with capacity $I(W)$. Fix the block error probability $P_{\mathrm{e}} \in (0,1)$ and an arbitrary $\delta \in (0,1]$. Then, there exists a fixed constant $\ell_0(\delta)$ such that for all $R < I(W)$, there exists a binary linear code of rate at least R that guarantees error probability at most P_{e} on the channel W, and whose block length n is at most*

$$n \leqslant \frac{\beta \ell_0(\delta)}{\big(I(W)-R\big)^{2+\delta}}, \tag{2}$$

where $\beta = \big(1+2P_{\mathrm{e}}^{-1}\big)^3$ is a universal constant. Furthermore, as R approaches $I(W)$ and n grows, this code has construction complexity $O(n)$ and encoding/decoding complexity $O(n \log n)$.

A few remarks regarding Theorem 1 are in order. First, in the definition of the constant β, the term P_{e} is raised to the power of -1. We point out that we could have similarly chosen any other negative constant as the exponent of P_{e}. However, picking a smaller exponent for P_{e} requires us to select an upper bound on δ which is stricter than $\delta \leqslant 1$. This, in turn, increases $\ell_0(\delta)$. Second, the error probability in Theorem 1 is upper bounded by a fixed constant P_{e}. However, a somewhat stronger claim is possible. It can be shown that Theorem 1 still holds if the error probability is required to decay *polynomially fast* with the block length n. Lastly, it should be emphasized that $\ell_0(\delta)$ is a constant that only depends on δ. However, its dependence is of an exponential nature, *i.e.* $\ell_0(\delta) = O(\exp(1/\delta^{1.01}))$. This limitation prevents the proposed scaling exponent to be exactly equal to 2 while maintaining the quasi-linear complexity property.

To prove Theorem 1, we will show that there exist $\ell \times \ell$ binary kernels, such that polar codes constructed from these kernels achieve capacity with a scaling exponent $\mu(\ell)$ that tends to the optimal value of 2 as ℓ grows. The claim regarding the construction and encoding/decoding complexities immediately follows from known results on polar codes [3, 24, 25]. Indeed, polar codes constructed from $\ell \times \ell$ binary kernels maintain the recursive structure of conventional polar codes, and thereby inherit construction complexity $O(n)$ and encoding/decoding complexity $O(n \log n)$. We will discuss the decoding complexity in more detail in Section 5.

1.3. A primer on polar codes

Like many fundamental discoveries, polar codes are rooted in a simple and beautiful basic idea. Polarization is induced via a simple linear transformation consisting of many Kronecker products of a binary matrix K, called the *polarization kernel*, with itself. Conventional polar codes, introduced by Arıkan in [3], correspond to

$$K = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}. \tag{3}$$

However, it was shown in [26] that we can construct polar codes from any kernel K that is an $\ell \times \ell$ nonsingular binary matrix, which cannot be transformed into an upper triangular matrix under any column permutations.

Let $W\colon \{0,1\} \to \mathscr{Y}$ be a BMS channel, characterized in terms of its transition probabilities $W(y|x)$, for all $y \in \mathscr{Y}$ and $x \in \{0,1\}$. Further, let $\boldsymbol{U} = (U_1, U_2, \ldots, U_n)$ be a block of $n = \ell^m$ bits chosen uniformly at

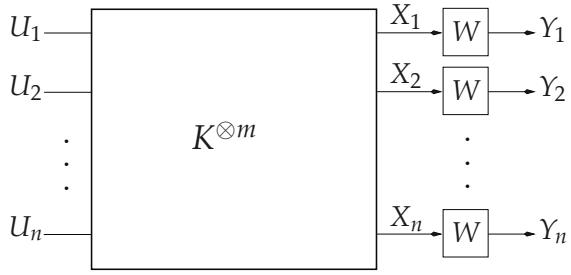

Figure 1: Block diagram of a polar coded communication scheme.

random from $\{0,1\}^n$. We encode $\boldsymbol{U}$ as $\boldsymbol{X} = \boldsymbol{U}K^{\otimes m}$ and transmit $\boldsymbol{X}$ through n independent copies of W, as shown in Figure 1.

To understand what polarization means in this context, we consider a number of channels associated with this transformation (see also Chapter 5 of [24] and Chapter 2.4 of [27]). Let $W^n\colon \{0,1\}^n \to \mathscr{Y}^n$ be the channel that corresponds to n independent uses of W, and let $W^*\colon \{0,1\}^n \to \mathscr{Y}^n$ be the channel with transition probabilities given by $W^*(\boldsymbol{y}|\boldsymbol{u}) = W^n\big(\boldsymbol{y}\,\big|\,\boldsymbol{u}K^{\otimes m}\big)$. Finally, for all $i \in [n] \triangleq \{1,2,\ldots,n\}$, let $W_i : \{0,1\} \to \mathscr{Y}^n \times \{0,1\}^{i-1}$ be the channel that is "seen" by the bit U_i, defined as

$$W_i\big(\boldsymbol{y},\boldsymbol{v}\big|\,u_i\big) \triangleq \frac{1}{2^{n-1}} \sum_{\boldsymbol{u}'\in\{0,1\}^{n-i}} W^*\Big(\boldsymbol{y}\,\Big|\,(\boldsymbol{v},u_i,\boldsymbol{u}')\Big) = \frac{1}{2^{n-1}} \sum_{\boldsymbol{u}'\in\{0,1\}^{n-i}} W^n\Big(\boldsymbol{y}\,\Big|\,(\boldsymbol{v},u_i,\boldsymbol{u}')K^{\otimes m}\Big), \tag{4}$$

where $(\cdot,\cdot)$ stands for concatenation. We say that W_i is the *i-th bit-channel*. It is easy to see that $W_i\big(\boldsymbol{y},\boldsymbol{v}\big|\,u_i\big)$ is indeed the probability of the event that $(Y_1,Y_2,\ldots,Y_n) = \boldsymbol{y}$ and $(U_1,U_2,\ldots,U_{i-1}) = \boldsymbol{v}$ given that $U_i = u_i$.

The key observation of [3] is that, as n grows, the n bit-channels W_i defined in (4) start *polarizing*: they approach either a *noiseless channel* or a *useless channel*. Formally, given a BMS channel W, its *capacity* $I(W)$, and *Bhattacharyya parameter* $Z(W)$ are defined by

$$\begin{aligned} I(W) &\triangleq \frac{1}{2} \sum_{y\in\mathscr{Y}} \sum_{x\in\{0,1\}} W(y|x) \log_2 \frac{W(y|x)}{\frac{1}{2}W(y|0) + \frac{1}{2}W(y|1)}\,, \\ Z(W) &\triangleq \sum_{y\in\mathscr{Y}} \sqrt{W(y|0)W(y|1)}. \end{aligned} \tag{5}$$

Given $\delta \in (0,1)$, let us say that a bit-channel W_i is *δ-bad* if $Z(W_i) \geqslant 1-\delta$ and *δ-good* if $Z(W_i) \leqslant \delta$. Then the polarization theorem of Arıkan [3, Theorem 1] can be informally stated as follows.

Theorem (Polarization theorem)**.** *For every $\delta \in (0,1)$, almost all bit-channels become either δ-good or δ-bad as $n \to \infty$. In fact, as $n \to \infty$, the fraction of δ-good bit-channels approaches the capacity $I(W)$ of the underlying channel W, while the fraction of δ-bad bit-channels approaches $1 - I(W)$.*

With $\delta = o(1/n)$, this theorem naturally leads to the construction of capacity-achieving *polar codes*. Specifically, an (n,k) polar code is constructed by selecting a set $\mathcal{A}$ of k δ-good bit-channels to carry the information bits, while the input to all the other bit-channels is frozen to zeros. In practice, the code parameters k and δ are usually selected according to the target rate of the code and/or the desired probability of error.

Henceforth, let us focus on the *binary erasure channel* with erasure probability z, which we denote as $\mathrm{BEC}(z)$. It is well known that for $W = \mathrm{BEC}(z)$, we have $Z(W) = z$ and $I(W) = 1 - z$. It is furthermore known (see, for example, [27, Section 3.4], [28], or [29, Section 2.2]) that if $W = \mathrm{BEC}(z)$, then for all $i \in [n]$, the i-th bit-channel W_i is also a binary erasure channel $\mathrm{BEC}\big(p_i(z)\big)$, whose erasure probability $p_i(z)$ is a polynomial of degree at most n in z.

A proof of the polarization theorem for the BEC follows by studying the evolution of these n erasure probabilities $p_i(z)$ as $n = \ell^m$ grows. For a fixed kernel K, this evolution is completely determined by the erasure probabilities of the ℓ bit-channels obtained after a *single step of polarization*. These ℓ erasure probabilities are a central object of study in this paper.

Definition (Polarization behavior)**.** *Let $W = \mathrm{BEC}(z)$ and let K be a fixed $\ell \times \ell$ binary polarization kernel. For each $i \in [\ell]$, we let $f_{K,i}(z)$ denote the erasure probability of the bit-channel W_i given by (4) with $n = \ell$ and $W^*(\boldsymbol{y}|\boldsymbol{u}) = W^{\ell}(\boldsymbol{y}\,|\,\boldsymbol{u}K)$. We refer to the set of ℓ polynomials $\{f_{K,1}(z), f_{K,2}(z), \ldots, f_{K,\ell}(z)\}$ as the polarization behavior of the kernel K.*

Indeed, we shall see later in this paper that $f_{K,i}(z)$ is a polynomial of degree at most ℓ in z, for all i. For example, in the special case of the 2×2 kernel (3), the polarization behavior is given by $f_{K,1}(z) = 2z - z^2$ and $f_{K,2}(z) = z^2$. With this notation, it is advantageous to view the $n = \ell^m$ erasure probabilities $p_i(z)$ as the values taken by a random variable Z_m induced by the uniform distribution on the ℓ^m bit-channels. Given that K is non-singular, one can show that $K^{\otimes m}$ is also non-singular. Furthermore, by applying the chain rule of mutual information, since the matrix $K^{\otimes m}$ is nonsigular, it is easy to see that the polar transform in Figure 1 preserves capacity. We can then study the evolution of this random variable Z_m as m grows. More formally, the recursive construction of $K^{\otimes m}$ makes it possible to introduce the martingale $\{Z_m\}_{m\in\mathbb{N}}$ defined as follows:

$$Z_{m+1} \;=\; f_{K,B_m}(Z_m), \qquad \text{for } B_m \sim \mathsf{Uniform}[\ell], \tag{6}$$

with the initial condition $Z_0 = z$. One can view (6) as a stochastic process on an infinite ℓ-ary tree, where in each step we take one of the ℓ available branches with uniform probability. The polarization theorem then follows from the almost sure convergence given by the martingale convergence theorem, which in this case implies that

$$\mathbb{P}\big(\lim_{m\to\infty} Z_m(1 - Z_m) = 0\big) = 1, \tag{7}$$

where the probability measure is defined with respect to the random selection of the bit-channel indices. This shows that the erasure probabilities $p_i(z)$ of the ℓ^m bit-channels polarize to either 0 or 1 as $m \to \infty$. Hence, the fraction of bit-channels that polarize to 0 approaches $I(W)$. The speed with which this polarization phenomenon takes place is the determining factor in the decay rate of the gap to capacity as a function of the block length $n = \ell^m$. We elaborate on this in the next subsection.

1.4. On the rate of polarization in various regimes

The performance of polar codes has been analyzed in several regimes. In the *error-exponent* regime, the rate $R < I(W)$ is fixed, and we study how the error probability P_{e} scales as a function of the block length n. This is represented by the vertical/blue cut in Figure 2. In [30], it is shown that the error probability under successive-cancellation decoding behaves roughly as $2^{-\sqrt{n}}$. A more refined scaling in this regime is proved in [31].

As common practice for comparison purposes, we consider using a communication channel over a family of channels that can be characterized by a single channel parameter such as the erasure probability in BEC. In the *error-floor* regime, the code is fixed (i.e., the rate R and the block length n are fixed), and we study how the error probability P_{e} scales as a function of the channel parameter. This approach corresponds to taking into account one of the four curves in Figure 2. In [32], it is proved that the stopping distance of polar codes scales as $\sqrt{n}$, which implies good error-floor performance under belief-propagation decoding. The authors of [32] also provide simulation results that show no sign of an error floor for transmission over the BEC and over the binary-input AWGN channel. This problem is completely settled in [17], where it is shown that polar codes do not exhibit error floors under transmission over any BMS channel.

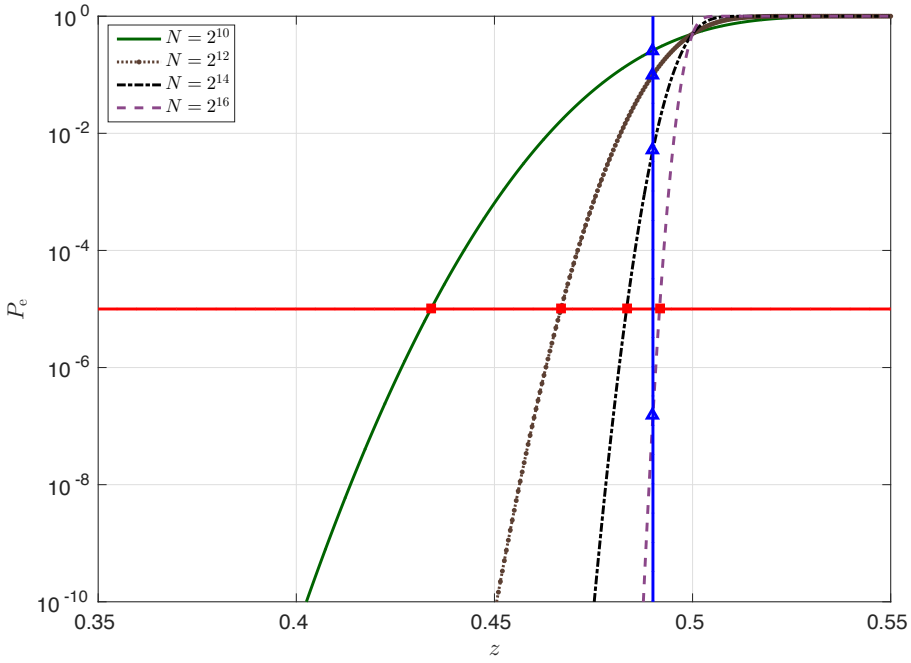

Figure 2: Performance of a family of codes with rate $R = 0.5$ over binary erasure channels. Different curves correspond to different codes of varying block length n. The x-axis is the erasure probability of underlying transmission channel, denoted by z, and the y-axis is the error probability P_e. The error-exponent regime captures the behavior of the blue/vertical cut at a fixed channel parameter z (or, equivalently, at a fixed gap to capacity $I(W) - R$). The error-floor regime captures the behavior of a single curve, of fixed block length n. The scaling-exponent regime captures the behavior of the red/horizontal cut at a fixed error probability P_e. The figure is courtesy of [17].

The focus of this paper is on the *scaling-exponent* regime, where the error probability P_e is fixed, and we study how the gap to capacity $I(W) - R$ scales as a function of the block length n. This approach is represented by the horizontal/red cut in Figure 2. As mentioned earlier, if n is $O\big(1/(I(W) - R)^{\mu}\big)$, we say that the family of codes has *scaling exponent* μ. For polar codes, the value of μ depends on the underlying channel W. In [16], a heuristic method is presented for computing the scaling exponent in the case of transmission over the BEC under successive-cancellation decoding; this method yields $\mu \approx 3.627$. In [11,12], it is shown that the block length, construction, encoding and decoding complexity are all bounded by a polynomial in the inverse of the gap to capacity, for transmission over any BMS channel. This implies that there exists a finite scaling exponent μ. Rigorous bounds on μ are provided in [14,15,17]. In [15], it is proved that $3.579 \leqslant \mu \leqslant 6$, and it is conjectured that the lower bound can be increased to 3.627 (i.e., up to the value heuristically computed for the BEC). In [14], the upper bound is improved to $\mu \leqslant 5.702$. The currently best-known upper bounds on the scaling exponent are established in [17]: for any BMS channel, $\mu \leqslant 4.714$; and for the special case of the BEC, $\mu \leqslant 3.639$, which approaches the value obtained heuristically in [16]. As a side note, let us point out that the heuristic method of [16] is based on a "scaling assumption" which requires the existence of a certain limit. The results of [14,15,17], as well as the results presented in this paper, do not rely on such an assumption.

In a nutshell, the scaling exponent of classical polar codes constructed via Arıkan's 2×2 kernel is around 4. Its exact value depends on the underlying transmission channel and can be bounded as $3.579 \leqslant \mu \leqslant 4.714$. In contrast, random binary linear codes achieve the optimal scaling exponent of 2. This means that, in order

to obtain the same gap to capacity, the block length of polar codes needs to be roughly the square of the block length of random codes. Hence, a natural question is how to improve the scaling exponent of polar codes.

One possible approach is to improve the successive-cancellation decoding algorithm. In particular, the successive cancellation *list decoder* proposed in [33] empirically provides a significant improvement in performance. However, [34] establishes a negative result for list decoders: the introduction of any finite-size list cannot improve the scaling exponent under MAP decoding for transmission over any BMS channel. Furthermore, for the special case of the BEC, it is also proved in [34] that the scaling exponent under successive-cancellation decoding does not change even under a finite number of interventions (that reverse incorrect decisions) from a genie.

Another approach is to consider polarization kernels of size larger than Arıkan's 2×2 matrix (3). Indeed, it is already known that such kernels have the potential to improve the scaling behavior of polar codes. For the error-exponent regime, Korada, Şaşoğlu, and Urbanke proved in [26] that for ℓ sufficiently large, there exist $\ell \times \ell$ binary kernels such that the error probability of the resulting polar codes scales roughly as 2^{-n}, rather than $2^{-\sqrt{n}}$. For the scaling-exponent regime, Fazeli and Vardy [28] observed that the value of μ on the BEC can be reduced from $\mu = 3.627$ for the matrix in (3) to $\mu(K_8) = 3.577$ and $\mu(K_{16}) = 3.356$, where K_8 and K_{16} are specific binary kernels constructed in [28]. Pfister and Urbanke [35] recently proved that, in the case of transmission over the q-ary erasure channel, the optimal scaling-exponent value of $\mu = 2$ can be approached as both the size of the kernel ℓ and the size of the alphabet q grow without bound. Furthermore, Hassani [27] gives evidence supporting the conjecture that, in order to approach $\mu = 2$ on the erasure channel, it suffices to consider large kernels over the *binary alphabet*. Herein, we finally settle this conjecture.

2. Outline of the Proof

The proof of our main result consists of several major steps. The technical part of the proof is, on occasion, quite intricate. To help the reader, we briefly discuss the main ideas behind each of the steps in this section.

Step 1: Characterization of the polarization process. In order to understand the finite-length scaling of polar codes, we need to understand how fast the random process Z_m defined in (6) polarizes. In other words, given a small $\zeta > 0$, how fast does the quantity $\mathbb{P}\{Z_m \in [\zeta, 1-\zeta]\}$ vanish with m? To answer this question, we first relate the decay rate of Z_m with another quantity that can be directly computed from the kernel matrix K.

As the first step along these lines, we consider the behavior of another random process $Y_m = g_\alpha(Z_m)$, where $g_\alpha(z) = z^\alpha(1-z)^\alpha$, and $\alpha > 0$ is a parameter to be determined later. Note that $Z_m \in [\zeta, 1-\zeta]$ if and only if Y_m is lower-bounded by $\zeta^\alpha(1-\zeta)^\alpha$. Therefore, by Markov's inequality, we have

$$\mathbb{P}\{Z_m \in [\zeta, 1-\zeta]\} \leqslant \frac{\mathbb{E}[g_\alpha(Z_m)]}{\zeta^\alpha(1-\zeta)^\alpha} \tag{8}$$

In order to derive an upper bound on $\mathbb{E}[g_\alpha(Z_m)]$, we write:

$$\begin{aligned} g_\alpha(Z_m) &= \Big(f_{K,B_m}(Z_{m-1})\big(1 - f_{K,B_m}(Z_{m-1})\big)\Big)^\alpha \\ &= Z_{m-1}^\alpha (1-Z_{m-1})^\alpha \left(\frac{f_{K,B_m}(Z_{m-1})\big(1 - f_{K,B_m}(Z_{m-1})\big)}{Z_{m-1}(1-Z_{m-1})}\right)^\alpha \\ &= g_\alpha(Z_{m-1}) \left(\frac{f_{K,B_m}(Z_{m-1})\big(1 - f_{K,B_m}(Z_{m-1})\big)}{Z_{m-1}(1-Z_{m-1})}\right)^\alpha. \end{aligned} \tag{9}$$

Proceeding along these lines, we eventually conclude that

$$\mathbb{E}[g_\alpha(Z_m)] \leqslant \big(\lambda^*_{\alpha,K}\big)^m, \tag{10}$$

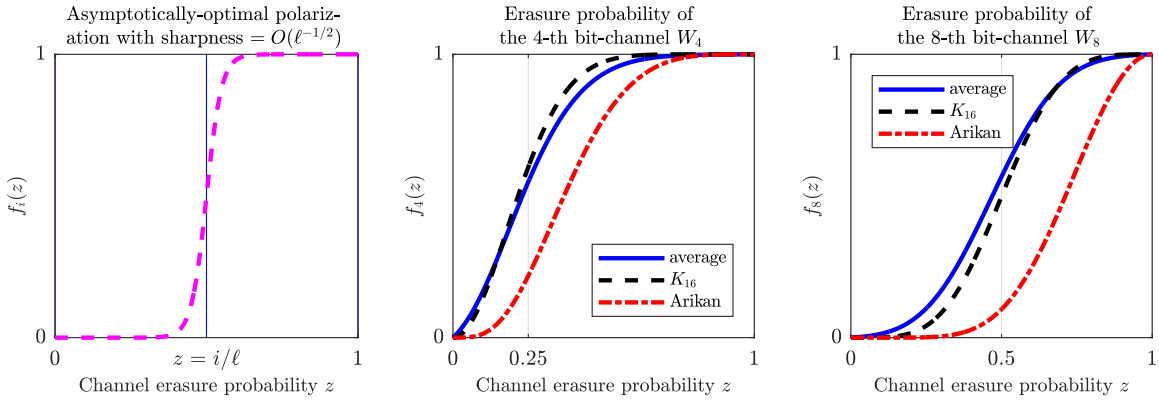

Figure 3: The figure on the left illustrates the fact that $f_{K,i}(z)$ has a sharp transition of order roughly $O\big(\ell^{-1/2}\big)$ when the kernel K is chosen at random. The two figures on the right compare three different choices of the kernel: the red curve corresponds to Arıkan's kernel; the black curve to the kernel K_{16} from [28], and the blue curve is obtained by taking the average of the functions $f_{K,i}(z)$ for all nonsingular 16×16 kernels.

where

$$\lambda^*_{\alpha,K} \triangleq \sup_{z\in(0,1)} \frac{1}{\ell} \frac{\sum_{i=1}^{\ell} \Big(f_{K,i}(z)\big(1-f_{K,i}(z)\big)\Big)^{\alpha}}{\big(z(1-z)\big)^{\alpha}}. \tag{11}$$

The discussion above is presented formally in Lemma 5.

Step 2: Sharp transitions in the polarization behavior. We fix $\alpha = 1/\log \ell$ and show that as ℓ grows, with probability at least $1 - o_\ell(1)$ over the random choice of a non-singular $\ell \times \ell$ binary kernel K, we have

$$\lambda^*_{\alpha,K} \;=\; O(\ell^{-1/2}\log \ell)\,. \tag{12}$$

To do so, we prove that, as ℓ grows, the polarization-behavior polynomials $f_{K,i}(z)$ will "look like" step functions for most nonsingular kernels. First note that $f_{K,i}(z)$ is an increasing polynomial with $f_{K,i}(0) = 0$ and $f_{K,i}(1) = 1$, for any i and any K. As ℓ increases, we show that $f_{K,i}(z)$ is likely to have a sharp transition threshold around the point $z = i/\ell$. More precisely, we prove that

$$\begin{aligned} f_{K,i}(z) &\leqslant \ell^{-(2+\log \ell)}, \qquad && \text{for } z \leqslant \frac{i}{\ell} - c_5 \ell^{-1/2}\log \ell, \\ f_{K,i}(z) &\geqslant 1 - \ell^{-(2+\log \ell)}, && \text{for } z \geqslant \frac{i}{\ell} + c_5 \ell^{-1/2}\log \ell, \end{aligned} \tag{13}$$

with probability at least $1 - O(1/\ell)$ over the random choice of K, where c_5 is a universal constant. This threshold behavior is illustrated (both schematically and for certain specific kernels of size $\ell = 16$) in Figure 3.

Step 3: Finite-length scaling law. For any fixed $\delta > 0$, we can derive the finite-length scaling law for polar codes using the results of the previous two steps. From (8), (10), and (12), we conclude that

$$\mathbb{P}\big\{Z_m \in [\zeta, 1-\zeta]\big\} \;=\; O\Big(\zeta^{-\alpha}\big(\ell^{-1/(2+\delta)}\big)^m\Big). \tag{14}$$

Denote the desired block error probability by P_e, and set $\varepsilon = P_e/n = P_e\ell^{-m}$ in (14), as is common in the polarization theory literature due the union upper bound on the block error rate. Then we have

$$\mathbb{P}\big\{Z_m \in [P_e\ell^{-m}, 1-P_e\ell^{-m}]\big\} \;=\; O\big(\ell^{-m/(2+\delta)}\big) \tag{15}$$

The foregoing is an upper bound on the fraction of bit-channels that are not yet sufficiently polarized after m polarization steps. Later, we will also provide a simple bound on the fraction $\mathbb{P}\{Z_m \geqslant 1 - P_e \ell^{-m}\}$ of bit-channels that are polarized to the useless state. Note that if we transmit information only on those bit-channels whose erasure probability is at most P_e/n, then a straightforward union-bound argument shows that the overall probability of error under successive-cancellation decoding is at most P_e. In essence, the bound in (15) implies that the fraction of such "good" bit-channels is at least $I(W) - O\big(\ell^{-m/(2+\delta)}\big)$. Since the block length n is ℓ^m, this means that the gap to capacity scales roughly as $n^{-1/(2+\delta)}$, which is the desired scaling law. Lemma 4 captures the above discussed argument.

3. Main Result

We begin by specializing Theorem 1 to polar codes and stating this result more precisely in Theorem 2. We then gradually reduce the proof of Theorem 2 to more and more specialized statements about large binary kernels. To do so, we start from the following definitions.

Definition 1 (Polar codes with large kernels)**.** *Consider transmission over a binary erasure channel W with capacity $I(W)$. Let $\mathrm{GL}(\ell, \mathbb{F}_2)$ denote the general linear group of all $\ell \times \ell$ non-singular matrices over $\mathbb{F}_2$. Let K be a specific polarization kernel chosen from $\mathrm{GL}(\ell, \mathbb{F}_2)$. We define $\mathcal{C}_K(n, R, P_e)$ to be a code of rate $R < I(W)$ obtained by polarizing K whose block length $n = \ell^m$ is the smallest such that the error probability under successive cancellation decoding is at most P_e.*

We observe that the code $\mathcal{C}_K(n, R, P_e)$ defined above always exists by the results of [26] and [24, Theorem 5.4], as long as the matrix K is not equivalent under column permutations to an upper triangular matrix.

Definition 2 (Upper bound on the scaling exponent)**.** *Consider transmission over a binary erasure channel W with capacity $I(W)$. For a fixed polarization kernel $K \in \mathrm{GL}(\ell, \mathbb{F}_2)$, we define $\overline{\mu}(K)$ to be an upper bound on the scaling exponent of K, if for any probability of error $P_e \in (0,1)$ and any rate $R < I(W)$, there exists a $\mathcal{C}_K(n, R, P_e)$ polar code whose block length n is upper bounded by*

$$n \leqslant \frac{\beta}{(I(W) - R)^{\overline{\mu}(K)}}, \tag{16}$$

where β is a constant that depends only on K and P_e.

Note that these definitions are consistent with the way the scaling exponent is defined in the literature, see e.g., [14, Theorem 1], [17, Theorem 1]. We are now ready to present our main results.

Theorem 2 (Binary polar codes with near-optimal scaling)**.** *Consider transmission over a binary erasure channel W with capacity $I(W)$. Let $K \in \mathrm{GL}(\ell, \mathbb{F}_2)$ be a kernel selected uniformly at random. Fix $\delta \in (0, 1]$. Then, there exists $\ell_0(\delta)$ such that for any $\ell > \ell_0(\delta)$, with high probability over the choice of K, the scaling exponent of K is upper-bounded by*

$$\overline{\mu}(K) \leqslant 2 + \delta. \tag{17}$$

Furthermore, the constant β in (16) is given by $\ell(1 + 2P_e^{-1})^3$.

In fact, what we prove is slightly stronger. Let $\{\mathcal{C}_K(n = \ell^m, R)\}_{m=1}^{\infty}$ be the family of rate-R large-kernel polar codes obtained by polarizing K for m many steps, where $R < I(W)$. We will show that as ℓ grows, with high probability over the choice of K, the scaling exponent can be upper-bounded by

$$\mu = 2 + O\left(\frac{\log(\log \ell)}{\log \ell}\right). \tag{18}$$

In order to treat (18) as $\mu \leqslant 2+\delta$, it suffices to pick a fixed ℓ that is at least in the order of $\exp\left(\delta^{-1.01}\right)$. We denote this minimum value of ℓ by $\ell_0(\delta)$. Once again, we emphasize that $\ell_0(\delta)$ scales exponentially with $1/\delta$. Note that a fixed value of ℓ, although being extremely large, does not change the asymptotic code-length, n, nor the decoding complexity, $O(n \log n)$, which is the main focus of this work. However, given how large ℓ should be and the fact that the decoding complexity of polar codes with arbitrary $\ell \times \ell$ kernels is also multiplied by 2^ℓ, it becomes clear that the large-kernel polar codes, whose kernels are chosen at random, are not suitable for the practical purposes. We address the recent advancements on the decoding problem of large kernels at the end of the paper.

We also point out that, as the rate R approaches the channel capacity $I(W)$, which consequently makes the block length n grow, these codes have construction complexity $O(n)$ and encoding/decoding complexity $O(n \log n)$. The claim on the construction complexity follows from the fact that the erasure probabilities of the bit-channels can be computed exactly according to the recursion (6). The claim on the encoding/decoding complexity follows from [26, Section VII].

The foregoing theorem follows from the following result that characterizes the behavior of the polarization process defined in (6).

Theorem 3 (Near-optimal scaling of the polarization process)**.** *Let $K \in \mathrm{GL}(\ell, \mathbb{F}_2)$ be a kernel selected uniformly at random from all $\ell \times \ell$ nonsingular binary matrices. Let Z_m be the random process defined in (6) with initial condition $Z_0 = z$. Fix $P_{\mathrm{e}} \in (0,1)$ and a small constant $\delta > 0$. Then, there exists $\ell_0(\delta)$ such that for all $\ell > \ell_0(\delta)$ and for all $m \geqslant 1$, and almost all K, we have*

$$\mathbb{P}\{Z_m \leqslant P_{\mathrm{e}} \ell^{-m}\} \;\geqslant\; 1 - z \;-\; \left(1 + 2\,P_{\mathrm{e}}^{-1}\right) \ell^{-\frac{m}{2+\delta}}. \tag{19}$$

For the sake of clarity, note that *(i)* in (19) the kernel K is fixed and the probability space is defined with respect to the random process Z_m, and *(ii)* the result (19) holds with high probability over the choice of the kernel K. We are now ready to present the proof of Theorem 2.

Proof of Theorem 2. Fix any rate R with $R < I(W)$. Assuming Theorem 3 holds, consider transmission over $\mathrm{BEC}(z)$ of a polar code with block length $n = \ell^m$ and rate R obtained by polarizing the $\ell \times \ell$ kernel K, where $\ell > \ell_0(\delta)$. By Theorem 3, with high probability over the choice of K, at least a $1 - z - (1 + 2\,P_{\mathrm{e}}^{-1})\ell^{-m/2+\delta}$ fraction of the bit-channels have erasure probability at most $P_{\mathrm{e}}\ell^{-m}$. Given that $R < 1 - z$, one can always find a positive integer m such that

$$R \;\leqslant\; 1 - z - (1 + 2\,P_{\mathrm{e}}^{-1})\ell^{-\frac{m}{2+\delta}}. \tag{20}$$

A simple union bound yields that the error probability under successive cancellation decoding is at most P_{e}. Given that $I(W) = 1 - z$, we can re-arrange (20) to obtain that

$$(1 + 2P_{\mathrm{e}}^{-1})\ell^{-\frac{m}{2+\delta}} \leqslant I(W) - R, \tag{21}$$

which is equivalent to

$$\frac{1 + 2P_{\mathrm{e}}^{-1}}{\left(I(W) - R\right)} \;\leqslant\; n^{\frac{1}{2+\delta}} \;\Leftrightarrow\; \frac{(1 + 2P_{\mathrm{e}}^{-1})^{2+\delta}}{\left(I(W) - R\right)^{2+\delta}} \leqslant n. \tag{22}$$

W.l.o.g., we can assume that $\delta < 1$ and, hence, we can take β as prescribed in Theorem 2. Considering that n is a power of ℓ, for (20) to hold, it suffices to have

$$\left\lceil \log_\ell \left(\frac{\beta}{\left(I(W) - R\right)^{2+\delta}} \right) \right\rceil \leqslant \log_\ell n\,, \tag{23}$$

where ℓ can be an arbitrary integer such that $\ell \geqslant \ell_0$. Therefore, the smallest value of n for which the desired code exists is the first integer power of ℓ that is not smaller than $\beta/(I(W)-R)^{2+\delta}$. Thus, there exists a code with

$$n \leqslant \frac{\beta \ell}{(I(W)-R)^{2+\delta}} \tag{24}$$

such that (23) holds. □

The rest of the section is devoted to the proof of Theorem 3. The basic idea is to bound the number of unpolarized bit-channels. To this end, let us introduce the polarization measure function $g_\alpha(z)$, defined as follows:

$$g_\alpha(z) \triangleq z^\alpha (1-z)^\alpha, \tag{25}$$

where $\alpha \in (0,1)$ is a constant parameter to be determined later. The first step is to show that an upper bound on $\mathbb{E}[g_\alpha(Z_m)]$ yields a lower bound on $\mathbb{P}\{Z_m \leqslant P_{\mathrm{e}} \ell^{-m}\}$. This is accomplished in the following lemma.

Lemma 4. *Let $K \in \mathrm{GL}(\ell, \mathbb{F}_2)$ be an $\ell \times \ell$ nonsingular binary kernel such that none of its column permutations is upper triangular. Let Z_m be the random process defined in (6) with initial condition $Z_0 = z$. Fix a constant $\alpha \in (0,1)$ and define $g_\alpha(z)$ as in (25). Further fix $\rho > 0, P_{\mathrm{e}} \in (0,1)$ and assume that*

$$\mathbb{E}[g_\alpha(Z_m)] \leqslant \ell^{-m\rho}. \tag{26}$$

for all $m \geqslant 1$. Then, for any $m \geqslant 1$, we have

$$\mathbb{P}\{Z_m \leqslant P_{\mathrm{e}} \ell^{-m}\} \;\geqslant\; 1 - z - \left(2P_{\mathrm{e}}^{-\alpha} + P_{\mathrm{e}}\right) \ell^{-m(\rho-\alpha)}. \tag{27}$$

Proof. First of all, we upper bound $\mathbb{P}\{Z_m \in [P_{\mathrm{e}} \ell^{-m}, 1 - P_{\mathrm{e}} \ell^{-m}]\}$ as follows:

$$\begin{aligned}
\mathbb{P}\left\{Z_m \in \left[P_{\mathrm{e}} \ell^{-m}, 1 - P_{\mathrm{e}} \ell^{-m}\right]\right\} &\overset{\text{(a)}}{=} \mathbb{P}\left\{g_\alpha(Z_m) \geqslant g_\alpha(P_{\mathrm{e}} \ell^{-m})\right\} \\
&\overset{\text{(b)}}{\leqslant} \frac{\mathbb{E}[g_\alpha(Z_m)]}{g_\alpha(P_{\mathrm{e}} \ell^{-m})} \\
&\overset{\text{(c)}}{\leqslant} \frac{\ell^{-m\rho}}{g_\alpha(P_{\mathrm{e}} \ell^{-m})} \\
&\overset{\text{(d)}}{\leqslant} 2P_{\mathrm{e}}^{-\alpha} \ell^{-m(\rho-\alpha)},
\end{aligned} \tag{28}$$

where equality (a) uses the concavity of $g_\alpha(\cdot)$ together with its symmetry around $1/2$; inequality (b) follows from Markov's inequality; inequality (c) uses the hypothesis $\mathbb{E}[g_\alpha(Z_m)] \leqslant \ell^{-m\rho}$; and inequality (d) uses the fact that $1 - P_{\mathrm{e}} \ell^{-m} \geqslant 1/2$ for all $m \geqslant 1$. Now, let us fix m from this point forward and define

$$\begin{aligned}
A &\triangleq \mathbb{P}\left\{Z_m \in \left(0, P_{\mathrm{e}} \ell^{-m}\right)\right\}, \\
B &\triangleq \mathbb{P}\left\{Z_m \in \left[P_{\mathrm{e}} \ell^{-m}, 1 - P_{\mathrm{e}} \ell^{-m}\right]\right\}, \\
C &\triangleq \mathbb{P}\left\{Z_m \in \left(1 - P_{\mathrm{e}} \ell^{-m}, 1\right)\right\},
\end{aligned} \tag{29}$$

and let A', B', and C' be the fraction of bit-channels in A, B, and C, respectively, that will have a vanishing erasure probability as $n \to \infty$. More formally, we define

$$\begin{aligned}
A' &\triangleq \lim_{m' \to \infty} \mathbb{P}\left\{Z_m \in \left(0, P_{\mathrm{e}} \ell^{-m}\right), Z_{m+m'} \leqslant \ell^{-m'}\right\}, \\
B' &\triangleq \lim_{m' \to \infty} \mathbb{P}\left\{Z_m \in \left[P_{\mathrm{e}} \ell^{-m}, 1 - P_{\mathrm{e}} \ell^{-m}\right], Z_{m+m'} \leqslant \ell^{-m'}\right\}, \\
C' &\triangleq \lim_{m' \to \infty} \mathbb{P}\left\{Z_m \in \left(1 - P_{\mathrm{e}} \ell^{-m}, 1\right), Z_{m+m'} \leqslant \ell^{-m'}\right\}.
\end{aligned} \tag{30}$$

We now show that the limits in (30) exists, and consequently the quantities A', B' and C' are well defined. To do so, it suffices to prove that, for any $z \in (0,1)$, the following limit exists

$$\lim_{m' \to \infty} \mathbb{P}\left\{ Z_{m+m'} \leqslant \ell^{-m'} \mid Z_m = z \right\}, \tag{31}$$

which is equivalent to proving the existence of

$$\lim_{m' \to \infty} \mathbb{P}\left\{ Z_{m'} \leqslant \ell^{-m'+m} \mid Z_0 = z \right\}. \tag{32}$$

Note that, for any $\beta > 0$,

$$\liminf_{m' \to \infty} \mathbb{P}\left\{ Z_{m'} \leqslant \ell^{-m'+m} \mid Z_0 = z \right\} \geqslant \liminf_{m' \to \infty} \mathbb{P}\left\{ Z_{m'} \leqslant 2^{-\ell^{m'\beta}} \mid Z_0 = z \right\}. \tag{33}$$

From the proof of [26, Theorem 2], we have that the random variable $Z_{m'}$ converges almost surely to a $\{0,1\}$ random variable Z_∞. Furthermore, an application of [24, Lemma 5.9] gives that, for $\beta < E(K)$,

$$\liminf_{m' \to \infty} \mathbb{P}\left\{ Z_{m'} \leqslant 2^{-\ell^{m'\beta}} \mid Z_0 = z \right\} = \mathbb{P}\left\{ Z_\infty = 0 \right\}, \tag{34}$$

where $E(K) > 0$ for all $\ell \times \ell$ nonsingular binary matrices none of whose column permutations is upper triangular. Suppose now that

$$\limsup_{m' \to \infty} \mathbb{P}\left\{ Z_{m'} \leqslant \ell^{-m'+m} \mid Z_0 = z \right\} > \mathbb{P}\left\{ Z_\infty = 0 \right\}. \tag{35}$$

Then, $Z_{m'}$ cannot converge in probability to Z_∞ [41, Page 70, Eq. (5)]. However, this cannot be possible since almost sure convergence implies convergence in probability [41, Theorem 4.1.2]. Thus,

$$\limsup_{m' \to \infty} \mathbb{P}\left\{ Z_{m'} \leqslant \ell^{-m'+m} \mid Z_0 = z \right\} = \liminf_{m' \to \infty} \mathbb{P}\left\{ Z_{m'} \leqslant \ell^{-m'+m} \mid Z_0 = z \right\}, \tag{36}$$

and the limits in (31) and (32) exist. Note that

$$A' + B' + C' \;=\; \lim_{m' \to \infty} \mathbb{P}\left\{ Z_{m+m'} \leqslant \ell^{-m'} \right\} \;=\; 1 - z. \tag{37}$$

In addition, from (28) we have that

$$B' \leqslant B \leqslant 2\, P_{\mathrm{e}}^{-\alpha}\, \ell^{-m(\rho-\alpha)}. \tag{38}$$

In order to upper bound C', we proceed as follows:

$$\begin{aligned} C' &= \lim_{m' \to \infty} \mathbb{P}\left\{ Z_{m+m'} \leqslant \ell^{-m'} \mid Z_m \in \left(1 - P_{\mathrm{e}}\, \ell^{-m}, 1\right] \right\} \cdot \mathbb{P}\left\{ Z_m \in \left(1 - P_{\mathrm{e}}\, \ell^{-m}, 1\right] \right\} \\ &\leqslant \lim_{m' \to \infty} \mathbb{P}\left\{ Z_{m+m'} \leqslant \ell^{-m'} \mid Z_m \in \left(1 - P_{\mathrm{e}}\, \ell^{-m}, 1\right] \right\}. \end{aligned} \tag{39}$$

By using again the fact that the kernel K is polarizing, we obtain that the last term equals the capacity of a BEC with erasure probability at least $1 - P_{\mathrm{e}}\, \ell^{-m}$. Consequently,

$$C' \leqslant P_{\mathrm{e}}\, \ell^{-m}. \tag{40}$$

As a result, we conclude that $\mathbb{P}\{Z_m \in [0, P_{\mathrm{e}}\, \ell^{-m})\} = A$ is bounded as follows

$$A \;\geqslant\; A' \overset{\text{(a)}}{=} 1 - z - B' - C' \overset{\text{(b)}}{\geqslant} 1 - z - 2\, P_{\mathrm{e}}^{-\alpha}\, \ell^{-m(\rho-\alpha)} - P_{\mathrm{e}}\, \ell^{-m} \overset{\text{(c)}}{\geqslant} 1 - z - \left(2\, P_{\mathrm{e}}^{-\alpha} + P_{\mathrm{e}}\right) \ell^{-m(\rho-\alpha)},$$

where equality (a) uses (37); inequality (b) uses (38) and (40); and inequality (c) uses the fact that since $\alpha, \rho \in (0,1)$ then $\rho - \alpha < 1$. This chain of inequalities implies the desired result. □

The second step is to derive an upper bound on $\mathbb{E}[g_\alpha(Z_m)]$ of the form $(\lambda^*_{\alpha,K})^m$, where $\lambda^*_{\alpha,K}$ depends on the particular kernel K. This is accomplished in Lemma 5, whose statement and proof follow.

Lemma 5. *Let $K \in \mathrm{GL}(\ell, \mathbb{F}_2)$ be a fixed $\ell \times \ell$ binary kernel. Let Z_m be the random process defined in (6) with initial condition $Z_0 = z$. Fix $\alpha \in (0,1)$ and define $g_\alpha(z)$ as in (25). For $z \in (0,1)$, define $\lambda_{\alpha,K}(z)$ as*

$$\lambda_{\alpha,K}(z) \triangleq \frac{\frac{1}{\ell}\sum_{i=1}^{\ell} g_\alpha(f_{K,i}(z))}{g_\alpha(z)}, \tag{41}$$

and let $\lambda^*_{\alpha,K}$ be its supremum, i.e.,

$$\lambda^*_{\alpha,K} \triangleq \sup_{z\in(0,1)} \lambda_{\alpha,K}(z). \tag{42}$$

Then, for any $m \geqslant 0$, we have that

$$\mathbb{E}[g_\alpha(Z_m)] \leqslant (\lambda^*_{\alpha,K})^m g_\alpha(z). \tag{43}$$

Proof. We prove the claim by induction. The base step $m = 0$ follows immediately from the fact that $Z_0 = z$. To prove the inductive step, we write

$$\mathbb{E}\big[g_\alpha(Z_{m+1})\big] = \mathbb{E}\big[\mathbb{E}[g_\alpha(f_{K,B_m}(Z_m)) \mid Z_m]\big], \tag{44}$$

where the first (outer) expectation on the RHS is with respect to Z_m and the second (inner) expectation is with respect to B_m. Then, we have that

$$\begin{aligned}\mathbb{E}\big[\mathbb{E}[g_\alpha(f_{K,B_m}(Z_m)) \mid Z_m]\big] &= \mathbb{E}\left[g_\alpha(Z_m)\frac{\frac{1}{\ell}\sum_{i=1}^{\ell} g_\alpha(f_{K,i}(Z_m))}{g_\alpha(Z_m)}\right] \\ &\leqslant \mathbb{E}\big[g_\alpha(Z_m)\big] \underbrace{\sup_{z\in\{0,1\}} \frac{\frac{1}{\ell}\sum_{i=1}^{\ell} g_\alpha(f_{K,i}(z))}{g_\alpha(z)}}_{\lambda^*_{\alpha,K}}.\end{aligned} \tag{45}$$

□

The third and final step is to prove that $\lambda^*_{\alpha,K}$ concentrates around $1/\sqrt{\ell}$, when K is selected uniformly at random among all $\ell \times \ell$ nonsingular binary matrices. This is done in Theorem 6, which is stated below.

Theorem 6 (Concentration of $\lambda^*_{\alpha,K}$)**.** *Let $K \in \mathrm{GL}(\ell, \mathbb{F}_2)$ be a kernel selected uniformly at random among all $\ell \times \ell$ non-singular binary matrices. Set $\alpha = 1/\log \ell$ and define $\lambda^*_{\alpha,K}$ as in (42). Then, there exists a universal constant c such that as ℓ grows, we have*

$$\mathbb{P}\left\{\lambda^*_{\alpha,K} \leqslant c\ell^{-\frac{1}{2}}\log \ell\right\} \geqslant 1 - o(1) \quad , \tag{46}$$

where the probability space is defined over the choice of the kernel K.

At this point, we are ready to put everything together and present a proof of Theorem 3, assuming that Theorem 6 holds. The proof of Theorem 6 is deferred to the next section.

Proof of Theorem 3. A simple counting over the binary subspaces of dimension ℓ shows that

$$\frac{\text{number of non-singular but non-polarizing } \ell \times \ell \text{ binary kernels}}{\text{number of non-singular } \ell \times \ell \text{ binary matrices}} = \frac{\ell! 2^{\frac{\ell(\ell-1)}{2}}}{\prod_{i=0}^{\ell-1}(2^\ell - 2^i)} \leqslant \frac{1}{2^\ell} \quad \text{for all} \quad \ell \geqslant 5. \tag{47}$$

Therefore, as ℓ grows, with probability at least $1 - 2^{-\ell} = 1 - o(1)$ over the choice of the kernel K, K is such that none of its column permutations is upper triangular. By Theorem 6, as ℓ grows, with probability at least $1 - o(1)$ over the choice of the kernel K, we also have that $\lambda^*_{\alpha,K} \leqslant c\ell^{-1/2}\log\ell$. Given that the intersection of these two sets also has probability at least $1 - o(1)$, most choices of K satisfy both conditions for sufficiently large ℓ. Fix any such kernel. Consequently, as $g_\alpha(z) \leqslant 1$ for all $z \in (0,1)$, by Lemma 5 we have that

$$\mathbb{E}[g_\alpha(Z_m)] \leqslant \left(c\ell^{-1/2}\log\ell\right)^m, \tag{48}$$

where the expectation is over the uniform selection of the polar bit-channel index, or in other words, the random process Z_m. Taking into the account that $\alpha = 1/\log\ell$, we can apply Lemma 4 to deduce that

$$\mathbb{P}\{Z_m \leqslant P_e\ell^{-m}\} \;\geqslant\; 1 - z - c_1(2c\ell^{-1/2}\log\ell)^m, \tag{49}$$

where $c_1 = 2P_e^{-\alpha} + P_e$ and the probability space is defined with respect to the random selection of the index. Note that, as $\alpha \leqslant 1$ and $P_e \leqslant 1$, we have that $c_1 \leqslant 1 + 2P_e^{-1}$. The theorem immediately follows by picking $\ell_0(\delta)$ to be large enough such that

$$2c\ell_0(\delta)^{-1/2}\log\ell_0(\delta) \leqslant \ell_0(\delta)^{-\frac{1}{2+\delta}}, \tag{50}$$

which is of the order $O(\exp(\delta^{-1.01}))$. □

4. Proof of Theorem 6: Concentration of $\lambda^*_{\alpha,K}$

Recall that our goal is to show that for most non-singular binary kernels $K \in \mathrm{GL}(\ell, \mathbb{F}_2)$,

$$\lambda_{\alpha,K}(z) \leqslant c\ell^{-\frac{1}{2}}\log\ell \quad \forall z \in (0,1), \tag{51}$$

where $\alpha = 1/\log\ell$ is fixed and c is some universal constant. Our strategy is to split the interval $(0,1)$ into the three sub-intervals $(0, 1/\ell^2)$, $[1/\ell^2, 1 - 1/\ell^2]$, and $(1 - 1/\ell^2, 1)$. Then, we will show that (51) holds for each of these sub-intervals. In fact, as we shall see, polarization is much faster at the tail intervals. Proposition 1 captures this approach.

Proposition 1. *Let $K \in \mathrm{GL}(\ell, \mathbb{F}_2)$ be a kernel selected uniformly at random from all $\ell \times \ell$ nonsingular binary matrices. Set $\alpha = 1/\log\ell$ and define $\lambda_{\alpha,K}(z)$ as in (41). Then, as ℓ grows, the following results hold.*

1. *Near optimal polarization in the middle:*

$$\mathbb{P}\left\{\lambda_{\alpha,K}(z) < c\ell^{-\frac{1}{2}}\log\ell, \;\; \forall z \in \left[\frac{1}{\ell^2}, 1 - \frac{1}{\ell^2}\right]\right\} > 1 - o(1), \tag{52}$$

2. *Faster polarization at the tails:*

$$\mathbb{P}\left\{\lambda_{\alpha,K}(z) < c\ell^{-1}\log\ell, \;\; \forall z \in \left(0, \frac{1}{\ell^2}\right) \cup \left(1 - \frac{1}{\ell^2}, 1\right)\right\} > 1 - o(1), \tag{53}$$

where the probability spaces are defined over the choice of the kernel K and c is a universal constant.

Proof of Theorem 6. Let A and B be the sets of kernels such that the events in (52) and (53) respectively hold. Then, by Proposition 1, we have that

$$\begin{aligned} \mathbb{P}(A) &> 1 - o(1), \\ \mathbb{P}(B) &> 1 - o(1). \end{aligned} \tag{54}$$

Furthermore, we have that

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cup B) \geqslant \mathbb{P}(A) + \mathbb{P}(B) - 1. \tag{55}$$

By combining (54) and (55), we conclude that

$$\mathbb{P}\left\{ \lambda_{\alpha,K}(z) < c\ell^{-\frac{1}{2}} \log \ell, \ \ \forall z \in (0,1) \right\} \ > \ 1 - o(1). \tag{56}$$

□

In what follows, we first analyze the probability of error under successive-cancellation decoding for the special case of transmission over the BEC. We formulate the erasure probability of the i-th polar bit-channel (at the kernel level) as a polynomial in the erasure probability of the underlying channel. Then, we utilize this formulation to introduce and compute the average polarization behavior and provide several auxiliary lemmas/propositions to establish the *sharp* transitions of $f_{K,i}(z)$ *on average* that was depicted earlier in Figure 3. Eventually, we put these propositions together and prove a concentration theorem, which, in turn, completes the proof for Proposition 1.

4.1. Successive cancellation decoding on binary erasure channels

Let $K \in \mathrm{GL}(\ell, \mathbb{F}_2)$ be a nonsingular binary kernel, and let s denote the number of erasures that occurred during the transmission over $\mathrm{BEC}(z)$. There are a total of $\binom{\ell}{s}$ distinct and equally-likely erasure patterns, and each of them occurs with probability $z^s(1-z)^{\ell-s}$. Let $\eta_s^{(i)}$ denote the number of erasure patterns with s erasures, which make u_i undecodable. Thus the erasure probability of the i-th bit-channel is given by

$$f_{K,i}(z) = \sum_{s=0}^{\ell} z^s (1-z)^{\ell-s} \eta_s^{(i)}. \tag{57}$$

Fix an erasure pattern with s erasures. To simplify notation in what follows, let us assume that the s erasures are in the last s positions. As in (4), let us write $\boldsymbol{u} = (\boldsymbol{v}, u_i, \boldsymbol{u}')$ for the vector encoded by the polar transformation in (1), and let $\boldsymbol{y} = (y_1, y_2, \ldots, y_{\ell-s})$ denote the vector observed at the channel output in which the erasure locations are removed. Then $\boldsymbol{y} = \boldsymbol{u}K|_{\ell-s}$ where $K|_{\ell-s}$ denotes the submatrix of K consisting of its first $\ell - s$ columns. Notice, however, that in addition to $\boldsymbol{y}$, the successive-cancellation decoder knows the vector $\boldsymbol{v}$ consisting of the first $i-1$ bits of $\boldsymbol{u}$. Thus let us write

$$\boldsymbol{y} \ = \ \boldsymbol{u}K\big|_{\ell-s} = \ (\boldsymbol{v}, u_i, \boldsymbol{u}')K\big|_{\ell-s} = \ (\boldsymbol{v}, 0, \boldsymbol{0})K\big|_{\ell-s} \ + \ (\boldsymbol{0}, u_i, \boldsymbol{u}')K\big|_{\ell-s} \tag{58}$$

and define $\boldsymbol{x} = \boldsymbol{y} - (\boldsymbol{v}, 0, \boldsymbol{0})K|_{\ell-s}$. Since this vector $\boldsymbol{x}$ can be computed by the decoder, it follows that the decoding task is to determine u_i given

$$\boldsymbol{x} \ = \ (u_i, \boldsymbol{u}')K'\big|_{\ell-s} \tag{59}$$

where $K'|_{\ell-s}$ denotes the submatrix of $K|_{\ell-s}$ consisting of its last $\ell-(i-1)$ columns. It is easy to see that u_i can be determined uniquely from $\boldsymbol{x}$ if and only if the vector $(1,0,0,\ldots,0)^t$ is in the column space of the matrix $K'|_{\ell-s}$. Thus we arrive at the following decodability condition:

$$u_i \text{ is decodable} \iff (1,\underbrace{0,0,\ldots,0}_{\ell-i})^t \in \text{ column space of } K'\big|_{\ell-s} \tag{60}$$

As we shall see, it is advantageous to rephrase this condition in terms of the column space of the $\ell\times(\ell-s)$ matrix $K|_{\ell-s}$, since we know that all the columns of this matrix are linearly independent. Clearly, $(1,0,0,\ldots,0)^t$ is in the column space of $K'|_{\ell-s}$ if and only if

$$\exists\,\psi_1,\psi_2,\ldots,\psi_{i-1}\in\mathbb{F}_2: \quad (\psi_1,\psi_2\cdots,\psi_{i-1},1,\underbrace{0,0,\cdots,0}_{\ell-i})^t \in \text{ column space of } K\big|_{\ell-s}. \tag{61}$$

Now let $\mathbf{e}_j$ denote the j-th element of the canonical basis for $\mathbb{F}_2^\ell$ and define the linear subspace E_j of $\mathbb{F}_2^\ell$ as $E_j \triangleq \langle \mathbf{e}_1,\mathbf{e}_2,\cdots,\mathbf{e}_j\rangle$, where $\langle\cdot\rangle$ denotes the linear span over $\mathbb{F}_2$. With this, in view of (60) and (61), the decodability condition can be rephrased as follows:

$$u_i \text{ is decodable} \iff (E_i\setminus E_{i-1})\cap\big(\text{column space of } K\big|_{\ell-s}\big) \neq \emptyset. \tag{62}$$

In what follows, we use (62) to derive an explicit formula for the probability that u_i is decodable — that is, for $\mathbb{P}\big\{(E_i\setminus E_{i-1})\cap\big(\text{column space of } K|_{\ell-s}\big)\neq\emptyset\big\}$ when K is selected uniformly at random from $\mathrm{GL}(\ell,\mathbb{F}_2)$.

4.2. Average polarization behavior

In this subsection, we study the erasure probability of the i-th bit-channel W_i given that (i) the kernel is selected uniformly at random from $\mathrm{GL}(\ell,\mathbb{F}_2)$, and ($ii$) the transmission channel is BEC(z). Explicitly, for all $i\in[\ell]$, we define the *average erasure probability* $\mathcal{F}_i(z)$ as follows:

$$\mathcal{F}_i(z) \triangleq \mathbb{E}_K\big[f_{K,i}(z)\big] = \frac{\sum\limits_{K\in\mathrm{GL}(\ell,\mathbb{F}_2)} f_{K,i}(z)}{\big|\mathrm{GL}(\ell,\mathbb{F}_2)\big|} = \frac{\sum\limits_{K\in\mathrm{GL}(\ell,\mathbb{F}_2)} f_{K,i}(z)}{\prod\limits_{j=0}^{\ell-1}(2^\ell-2^j)}. \tag{63}$$

In what follows, we analyze the asymptotic behavior of $\mathcal{F}_i(z)$ and show that, as ℓ grows, $\mathcal{F}_i(z)$ becomes close to a step function with a jump at $z\sim{}^i\!/_\ell$. Later on, we prove concentration results that show that, with high probability over the choice of the kernel, $f_{K,i}(z)$ is also close to a sharp step function centered around $z\sim{}^i\!/_\ell$. This is captured in Propositions 2 and 3.

Proposition 2 (Lower bound on the average erasure probability)**.** *Let $\mathcal{F}_i(z)$ be the average erasure probability of the i-th bit-channel as defined in (63). Fix $\beta,\sigma\in\mathbb{R}^+\triangleq\{x: x\in\mathbb{R}, x>0\}$ and assume that*

$$z > \frac{i}{\ell}+\frac{\lceil\sigma\log\ell\rceil}{\ell}+\left(\frac{\beta\ln\ell}{2\ell}\right)^{1/2}, \tag{64}$$

where log *and* ln *denote the logarithm in base* 2 *and e, respectively. Then, we have that*

$$\mathcal{F}_i(z) > \big(1-\ell^{-\beta}\big)\big(1-\ell^{-\sigma}\big). \tag{65}$$

Proposition 3 (Upper bound on the average erasure probability)**.** *Let $\mathcal{F}_i(z)$ be the average erasure probability of the i-th bit-channel as defined in (63). Fix $\beta, \sigma \in \mathbb{R}^+$ and assume that*

$$z \;<\; \frac{i}{\ell} - \frac{h(\sigma)}{\ell} \;-\; \left(\frac{\beta \ln \ell}{2\ell}\right)^{1/2}, \tag{66}$$

where log *and* ln *denote the logarithms in base* 2 *and e, respectively, and*

$$h(\sigma) \;=\; \left\lfloor \frac{\sigma \log \ell + \log 6}{\log 3 - 1} \right\rfloor \;=\; O(\sigma \log \ell). \tag{67}$$

Then, we have that

$$\mathcal{F}_i(z) < \ell^{-\beta} + \ell^{-\sigma}. \tag{68}$$

Recall that $\mathcal{F}_i(z)$ is the probability of observing an erasure at the i-th bit-channel, when there are two sources of randomness: (i) the selection of the kernel, and (ii) the number and location of the erased bits. Let the random variable S denote the number of erased bits at the receiver. As z is the erasure probability of the underlying transmission channel, we have that

$$\mathbb{P}\{S = s\} \;=\; \binom{\ell}{s} z^s (1-z)^{\ell - s}. \tag{69}$$

Since we also average over all $\ell \times \ell$ nonsingular kernels, the location of these s erasures does not affect the average erasure probability. Hence, without loss of generality, we can assume that the erasures are in the last s positions. Let $\mathcal{R}_{\ell-s} \subset \mathbb{F}_2^\ell$ denote the linear span of the first $\ell - s$ columns of the kernel. Since the kernel is selected uniformly at random from $\mathrm{GL}(\ell, \mathbb{F}_2)$, it is easy to see that $\mathcal{R}_{\ell-s}$ is also chosen uniformly at random from all subspaces of dimension $\ell - s$ in $\mathbb{F}_2^\ell$. Recalling the decodability condition (62), we have that

$$\mathbb{P}\{u_i = \text{erasure} \,|\, S = s\} \;=\; \mathbb{P}\big\{\mathcal{R}_{\ell-s} \cap (E_i \setminus E_{i-1}) = \emptyset\big\}, \tag{70}$$

where $\mathcal{R}_{\ell-s}$ is a subspace of dimension $\ell - s$ in $\mathbb{F}_2^\ell$ that is chosen uniformly at random. Note that the *event* on the Left Hand Side (LHS) is reliant on a specific number of erasures, s, and is computed over all possible locations of erasures and selections of K. However, the *event* on the RHS is independent of the location and number of erasures, and thus is computed over all selections of random subspace $\mathcal{R}_{\ell-s}$. Therefore, the probability that the i-th bit is erased given s erasures is a claim solely on the structure of the kernel. Now, we can rewrite $\mathcal{F}_i(z)$ as

$$\mathcal{F}_i(z) \;=\; \sum_{s=0}^{\ell} \mathbb{P}\{S = s\} \mathbb{P}\{u_i = \text{erasure} \,|\, S = s\} \;=\; \sum_{s=0}^{\ell} \binom{\ell}{s} z^s (1-z)^{\ell-s} p_{i|s} \;, \tag{71}$$

where we define the *average conditional erasure probability* $p_{i|s}$ as follows:

$$p_{i|s} \;\triangleq\; \mathbb{P}\{u_i = \text{erasure} \,|\, S = s\} \;=\; \mathbb{P}\{\mathcal{R}_{\ell-s} \cap (E_i \setminus E_{i-1}) = \emptyset\}, \tag{72}$$

where the right-most equality is derived from (70).

Lemma 7 (Closed-form for the average conditional erasure probability)**.** *Let $p_{i|s}$ be the average conditional erasure probability defined in (72). Then, for any i and s, we have*

$$p_{i|s} \;=\; \begin{bmatrix} \ell \\ \ell - s \end{bmatrix}^{-1} \sum_{t=\max\{i-s,0\}}^{\min\{\ell-s, i-1\}} \begin{bmatrix} i-1 \\ t \end{bmatrix} \prod_{j=0}^{\ell-s-t-1} \frac{2^\ell - 2^{i+j}}{2^{\ell-s} - 2^{t+j}} \;, \tag{73}$$

where $\begin{bmatrix} a \\ b \end{bmatrix}$ *is the binary Gaussian binomial coefficient that denotes the total number of subspaces with dimension b in $\mathbb{F}_2^a$.*

Proof. Let $\Delta_{\ell-s}$ denote the number of subspaces of dimension $\ell-s$ in $\mathbb{F}_2^\ell$. That is,

$$\Delta_{\ell-s} \triangleq \begin{bmatrix} \ell \\ \ell-s \end{bmatrix} = \prod_{j=0}^{\ell-s-1} \frac{2^\ell - 2^j}{2^{\ell-s} - 2^j}. \tag{74}$$

Define $\Gamma(t;\ell,s,i)$ as the number of subspaces A of dimension $\ell-s$ in $\mathbb{F}_2^\ell$ such that $A \cap (E_i \setminus E_{i-1}) = \emptyset$ and $\dim(A \cap E_{i-1}) = t$. Recall that E_i and E_{i-1} are linear subspaces of $\mathbb{F}_2^\ell$ with respective dimensions of i and $i-1$, and $E_{i-1} \subset E_i$. Therefore, $\Gamma(t;\ell,s,i)$ is equal to the number of subspaces A of dimension $\ell-s$ in $\mathbb{F}_2^\ell$ such that $\dim(A \cap E_{i-1}) = \dim(A \cap E_i) = t$. Consequently, the integer t in the definition of $\Gamma(t;\ell,s,i)$ satisfies

$$\max\{i-s,0\} \leqslant t \leqslant \min\{\ell-s, i-1\}. \tag{75}$$

A simple basis counting argument (see, for example, [36, Section II.C]) yields that

$$\Gamma(t;\ell,s,i) = \begin{bmatrix} i-1 \\ t \end{bmatrix} \prod_{j=0}^{\ell-s-t-1} \frac{2^\ell - 2^{i+j}}{2^{\ell-s} - 2^{t+j}} \tag{76}$$

where the first term in (76) counts the number of subspace of dimension t in E_{i-1} whereas the second term counts the (normalized) number of basis extensions from dimension t to dimension $\ell-s$. Enumerating over all possible values of t given by (75), the desired conditional erasure probability can be written as

$$p_{i|s} = \frac{\sum_{t=\max\{i-s,0\}}^{\min\{\ell-s,i-1\}} \Gamma(t;\ell,s,i)}{\Delta_{\ell-s}} = \begin{bmatrix} \ell \\ \ell-s \end{bmatrix}^{-1} \sum_{t=\max\{i-s,0\}}^{\min\{\ell-s,i-1\}} \begin{bmatrix} i-1 \\ t \end{bmatrix} \prod_{j=0}^{\ell-s-t-1} \frac{2^\ell - 2^{i+j}}{2^{\ell-s} - 2^{t+j}}. \tag{77}$$

□

Next, we use this closed-form expression to provide upper and lower bounds on the average conditional erasure probability $p_{i|s}$ and on the average erasure probability $\mathcal{F}_i(z)$.

Lemma 8 (Lower bound on the average conditional erasure probability)**.** *Let $p_{i|s}$ be the average conditional erasure probability defined in (72). Then, for any i and s, we have*

$$p_{i|s} \geqslant 1 - 2^{-(s-i)}. \tag{78}$$

Proof. If $i \geqslant s$, then the lemma holds vacuously. Henceforth, let us assume that $i < s$. We drop all but the first term from (77) to write

$$p_{i|s} = \frac{\sum_{t=0}^{\min\{\ell-s,i-1\}} \Gamma(t,\ell,s,i)}{\Delta_{\ell-s}} \geqslant \frac{\Gamma(0;\ell,s,i)}{\Delta_{\ell-s}} = \frac{\prod_{j=0}^{\ell-s-1} \frac{2^\ell - 2^{i+j}}{2^{\ell-s} - 2^j}}{\prod_{j=0}^{\ell-s-1} \frac{2^\ell - 2^j}{2^{\ell-s} - 2^j}} = \prod_{j=0}^{\ell-s-1} \frac{2^\ell - 2^{i+j}}{2^\ell - 2^j}. \tag{79}$$

The proof now reduces to the following calculation:

$$\prod_{j=0}^{\ell-s-1} \frac{2^\ell - 2^{i+j}}{2^\ell - 2^j} > \prod_{j=0}^{\ell-s-1} \frac{2^\ell - 2^{i+j}}{2^\ell} = \prod_{j=0}^{\ell-s-1} \left(1 - 2^{-(\ell-i)+j}\right) \overset{(a)}{\geqslant} 1 - \sum_{j=0}^{\ell-s-1} 2^{-(\ell-i)+j} \overset{(b)}{>} 1 - 2^{-(s-i)}, \tag{80}$$

where (a) is because of

$$\prod_{i=1}^{n}(1-x_i) \geqslant 1-\sum_{i=1}^{n} x_i \quad \text{for all} \quad 0 \leqslant x_i \leqslant 1, \tag{81}$$

which can be shown by induction, and (b) is due to the fact that for all $n \in \mathbb{N}$, we have $2^{-n} = \sum_{i=n+1}^{\infty} 2^{-i}$. □

Proof of Proposition 2. We begin by dropping the first $i + \lceil \sigma \log(\ell) \rceil + 1$ terms in (71) and applying Lemma 8 to obtain

$$\begin{aligned} \mathcal{F}_i(z) \;=\; \sum_{s=0}^{\ell} \binom{\ell}{s} z^s (1-z)^{\ell-s} p_{i|s} \quad &> \sum_{s=i+\lceil \sigma \log \ell \rceil + 1}^{\ell} \binom{\ell}{s} z^s (1-z)^{\ell-s} (1-2^{-(s-i)}) \\ &\geqslant (1-\ell^{-\sigma}) \sum_{s=i+\sigma\lceil \log \ell \rceil + 1}^{\ell} \binom{\ell}{s} z^s (1-z)^{\ell-s}. \end{aligned} \tag{82}$$

Now, we point out that the sum on the RHS of (82) is the tail probability of a binomial distribution with ℓ trials and a success rate of z. More formally, $X \sim B(\ell, z)$ is a binomial random variable with ℓ trials and success probability z. Then, from (82) we immediately obtain that

$$\mathcal{F}_i(z) \;>\; \left(1-\ell^{-\sigma}\right) \mathbb{P}\{X > i + \lceil \sigma \log \ell \rceil\}. \tag{83}$$

Now, we invoke Hoeffding's inequality [37] for a sequence of of i.i.d. Bernoulli random variables with success probability $1-z$ and ℓ trials, which states that for any $\nu > 0$,

$$\mathbb{P}\big(X \geqslant (1-z+\nu)\ell\big) \leqslant \exp(-2\nu^2\ell). \tag{84}$$

By replacing the value of ν with the expressions from (2), we have

$$\begin{aligned} \mathbb{P}\{X > i + \lceil \sigma \log \ell \rceil\} &= 1 - \mathbb{P}\{X \leqslant i + \lceil \sigma \log \ell \rceil\} \\ &\overset{\text{(a)}}{\geqslant} 1 - \exp\left(-2\frac{\big(z\ell - (i + \lceil \sigma \log \ell \rceil)\big)^2}{\ell}\right) \\ &\overset{\text{(b)}}{\geqslant} 1 - \ell^{-\beta}, \end{aligned} \tag{85}$$

where in (a) we have used Hoeffding's inequality and in (b) we have used (64). The lemma now readily follows by combining (82) and (85). □

Next, we use the closed-form expression in Lemma 7 in order to derive a lower bound on the average conditional erasure probability and on the average erasure probability.

Lemma 9 (Upper bound on the average conditional erasure probability)**.** *Let $p_{i|s}$ be the average conditional erasure probability defined in (72). Then, for any i and s,*

$$p_{i|s} \;\leqslant\; 2\left(\frac{2}{3}\right)^{i-s-1}. \tag{86}$$

Proof. If $s \geqslant i-1$, the bound holds vacuously. Henceforth, let us assume that $s < i-1$. We start by proving that the term with $t = i-s$ is the dominant one in the expression (77) for $p_{i|s}$. For all $t > i-s$, we have that

$$\frac{\Gamma(t;\ell,s,i)}{\Gamma(t-1;\ell,s,i)} = \frac{\begin{bmatrix} i-1 \\ t \end{bmatrix}}{\begin{bmatrix} i-1 \\ t-1 \end{bmatrix}} \times \left(\prod_{j=0}^{\ell-s-t-1} \frac{2^\ell - 2^{i+j}}{2^{\ell-s} - 2^{t+j}} \right) \Big/ \left(\prod_{j=0}^{\ell-s-t} \frac{2^\ell - 2^{i+j}}{2^{\ell-s} - 2^{t+j-1}} \right), \tag{87}$$

which using a straightforward manipulation can be simplified as

$$\frac{(2^{i-1}-2^{t-1})(2^{\ell-s}-2^{t-1})}{2^{t-1}(2^t-1)(2^\ell - 2^{i+\ell-s-t})} \leqslant \frac{1}{2^{t-1}} \cdot \frac{2^{i-1}\cdot 2^{\ell-s}}{2^{t-1}\cdot 2^{\ell-1}} = \frac{2^{i-s-t+1}}{2^{t-1}} \leqslant 2^{-t+1} \leqslant \frac{1}{2}. \tag{88}$$

Therefore, for any $t > i-s$, we have that

$$\Gamma(t;\ell,s,i) \;\leqslant\; 2^{-(t-(i-s))}\,\Gamma(i-s;\ell,s,i), \tag{89}$$

which implies that

$$p_{i|s} \;\leqslant\; \frac{\Gamma(i-s;\ell,s,i)}{\Delta_{\ell-s}}\Big(1+2^{-1}+2^{-2}+\cdots\Big) \;\leqslant\; \frac{2\Gamma(i-s;\ell,s,i)}{\Delta_{\ell-s}}. \tag{90}$$

In a similar fashion, we fix ℓ and i, and study the exponential decay of the dominant term in $p_{i|s}$, denoted by $\xi_s \triangleq \Gamma(i-s;\ell,s,i)/\Delta_{\ell-s}$, as s decreases. We again use straightforward manipulation to obtain

$$\begin{aligned} \frac{\xi_s}{\xi_{s+1}} &= \frac{\Delta_{\ell-s-1}}{\Delta_{\ell-s}} \times \frac{\begin{bmatrix} i-1 \\ i-s \end{bmatrix}}{\begin{bmatrix} i-1 \\ i-s-1 \end{bmatrix}} \times \left(\prod_{j=0}^{\ell-i-1} \frac{2^\ell - 2^{i+j}}{2^{\ell-s} - 2^{i-s+j}} \right) \Big/ \left(\prod_{j=0}^{\ell-i-1} \frac{2^\ell - 2^{i+j}}{2^{\ell-s-1} - 2^{i-s-1+j}} \right) \\ &= \frac{(2^{i-1}-2^{i-s-1})(2^{\ell-s}-1)}{(2^{i-s}-1)(2^\ell - 2^{\ell-s-1})} = \left(\frac{2^s-1}{2^{s+1}-1}\right)\frac{1-2^{-(\ell-s)}}{1-2^{-(i-s)}} \\ &\leqslant \frac{1}{2} \times \frac{1}{1-2^{-(i-s)}} \leqslant \frac{1}{2} \times \frac{1}{1-1/4} = \frac{2}{3}. \end{aligned} \tag{91}$$

As a result, we conclude that, for any $s < i-1$,

$$p_{i|s} \overset{(90)}{\leqslant} \frac{2\Gamma(i-s;\ell,s,i)}{\Delta_{\ell-s}} \overset{(91)}{\leqslant} 2\xi_{i-1}\left(\frac{2}{3}\right)^{i-s-1} \leqslant 2\left(\frac{2}{3}\right)^{i-s-1}, \tag{92}$$

where the last inequality follows from the fact that $\Gamma(i-s;\ell,s,i)$ is the number of subspaces of dimension $\ell-s$ in $\mathbb{F}_2^\ell$ with some additional properties, while $\Delta_{\ell-s}$ denotes the total nummber of such subspaces, and thus, $\xi_s \leqslant 1$ for all s that ξ_s is well defined. □

Proof of Proposition 3. Let us recall the formulation of $\mathcal{F}_i(z)$ from (71) and split the summation into two parts, where a trivial upper bound is applied to each part: we drop $\binom{\ell}{s} z^s (1-z)^{\ell-s}$ for all terms in the summation with $s \leqslant i - h(\sigma) - 1$, and we drop $p_{i|s}$ from the remaining terms that correspond to $s \geqslant i - h(\sigma)$. More formally, we have

$$\begin{aligned} \mathcal{F}_i(z) = &\sum_{s=0}^{i-h(\sigma)-1} \binom{\ell}{s} z^s (1-z)^{\ell-s} p_{i|s} \;+\; \sum_{s=i-h(\sigma)}^{\ell} \binom{\ell}{s} z^s (1-z)^{\ell-s} p_{i|s} \\ < &\sum_{s=0}^{i-h(\sigma)-1} p_{i|s} \qquad\qquad\qquad + \sum_{s=i-h(\sigma)}^{\ell} \binom{\ell}{s} z^s (1-z)^{\ell-s}. \end{aligned} \tag{93}$$

We apply the upper bound in (86) to the first summation, and obtain that

$$\sum_{s=0}^{i-h(\sigma)-1} p_{i|s} \leqslant \sum_{s=0}^{i-h(\sigma)-1} 2\left(\frac{2}{3}\right)^{i-s-1} \leqslant \sum_{s=h(\sigma)}^{\infty} 2\left(\frac{2}{3}\right)^{s} = 6\left(\frac{2}{3}\right)^{h(\sigma)} = \ell^{-\sigma}. \tag{94}$$

Utilizing the assumption in (66), the second summation is again upper bounded by applying Hoeffding's inequality on the tail probability of the binomial distribution $X \sim B(\ell, z)$ with ℓ trials and a success rate of z as follows:

$$\sum_{s=i-h(\sigma)}^{\ell} \binom{\ell}{s} z^s (1-z)^{\ell-s} = \mathbb{P}\{X \geqslant i - h(\sigma)\} \leqslant \mathbb{P}\left\{X \geqslant z\ell + \left(\frac{\beta \ell \ln \ell}{2}\right)^{1/2}\right\} \leqslant \ell^{-\beta}. \tag{95}$$

□

4.3. Proof of Proposition 1

At this point, we have gathered all the required tools to prove Proposition 1. Our proof consists of two steps. First, we show that the polarization behavior of a random non-singular $\ell \times \ell$ kernel is given, with high probability, by the function $\mathcal{F}_i(z)$ analyzed in the previous subsection. Then, we explain how to relate this fact to an upper bound on $\lambda_{\alpha,K}(z)$. Note that throughout this section, all probabilities are defined with respect to the random selection of non-singular kernels and there is no randomness in i. In fact, the polarization behavior of a desired kernel should be similar to $\mathcal{F}_i(z)$ for all i. As the theorem suggests, we split the proof into two parts: the first part takes care of the middle interval and proves (52), while the second part takes care of the tail intervals and proves (53).

Proof of (52). First, we combine the results of Proposition 2 and Proposition 3 to show that $\mathcal{F}_i(z)$ roughly behaves as a step function. In the previous subsection, we have shown that

$$\begin{cases} \mathcal{F}_i(z) > (1-\ell^{-\beta})(1-\ell^{-\sigma}), & \text{if } z > \frac{i}{\ell} + \frac{\lceil \sigma \log \ell \rceil}{\ell} + \left(\frac{\beta \ln \ell}{2\ell}\right)^{1/2} \\ \mathcal{F}_i(z) < \ell^{-\beta} + \ell^{-\sigma}, & \text{if } z < \frac{i}{\ell} - \frac{\left\lfloor \frac{\sigma \log \ell + \log 6}{\log 3 - 1} \right\rfloor}{\ell} - \left(\frac{\beta \ln \ell}{2\ell}\right)^{1/2} \end{cases}. \tag{96}$$

Our strategy is to show that, with high probability over the choice of the kernel, $f_{K,i}(z)$ is sharp for each fixed value of i. Then, we will use a union-bound-like argument to show that $f_{K,i}(z)$ is sharp for all $i \in [\ell]$. To this end, we first set $\beta = \sigma = 4.5 + \log \ell$ in (96). Given that the nature of our results is asymptotic, we assume that $\ell \geqslant 32$. Now, it is easy to derive the following from (96).

$$\begin{cases} \mathcal{F}_i(z) > 1 - 2\ell^{-4.5-\log \ell} > 1 - (2\ell^{4+\log \ell})^{-1}, & \text{if } z \geqslant \frac{i}{\ell} + c_3 \ell^{-1/2} \log \ell \\ \mathcal{F}_i(z) < 2\ell^{-4.5-\log \ell} < (2\ell^{4+\log \ell})^{-1}, & \text{if } z \leqslant \frac{i}{\ell} - c_4 \ell^{-1/2} \log \ell \end{cases}, \tag{97}$$

where

$$\begin{aligned} c_3 &= \max_{\ell \geqslant 32} \frac{\frac{\lceil (4.5+\log \ell) \log \ell \rceil}{\ell} + \left(\frac{(4.5+\log \ell) \ln \ell)}{2\ell}\right)^{1/2}}{\ell^{-1/2} \log \ell}, \\ c_4 &= \max_{\ell \geqslant 32} \frac{\frac{\left\lceil \frac{(4.5+\log \ell) \log \ell + \log 6}{\log 3 - 1} \right\rceil}{\ell} + \left(\frac{(4.5+\log \ell) \ln \ell)}{2\ell}\right)^{1/2}}{\ell^{-1/2} \log \ell}. \end{aligned} \tag{98}$$

Note that both c_3 and c_4 are finite numbers since the numerators in the RHSs of (98) decay faster than their respective denominators as ℓ grows. It is also possible to remove the ceilings and show that both expressions are decreasing functions of ℓ if $\ell \geqslant 32$, which means that they attain their maximum at $\ell = 32$. Thus, $c_3 < 2.51$ and $c_4 < 3.86$.

Our goal is to prove the simultaneous concentration of $f_{K,i}(z)$'s around their means, $\mathcal{F}_i(z)$, with regards to where and how fast they transition from $f_{K,i}(z) \sim 0$ to $f_{K,i}(z) \sim 1$. To do so, we first show that for any fixed value of i, $f_{K,i}(z)$ behaves similar to the average behavior, with high probability over the choice of K. Next, we provide a union-bound-like argument to prove that, with high probability over the choice of K, $f_{K,i}(z)$ is close to the average for all values of i. For the first step, we recall the erasure probability of the i-th bit-channel from (57) and expand it as

$$f_{K,i}(z) = \sum_{\text{all erasure patterns } \mathbf{e} \in \mathbb{F}_2^\ell} \mathbf{1}[\mathbf{e} \text{ makes } i\text{-th bit-channel undecodable}] z^{wt(\mathbf{e})} (1-z)^{\ell - wt(\mathbf{e})}. \tag{99}$$

Considering that each $z^{wt(\mathbf{e})}(1-z)^{\ell-wt(\mathbf{e})}$ term in the function above is a continuous and increasing function of z, we deduce that $f_{K,i}(z)$ is also a continuous and increasing function of z. Therefore, to show the sharp transition of $f_{K,i}(z)$ around $z = i/\ell$, it suffices to consider only two points in $(0,1)$, one slightly larger than $z = i/\ell$ and one slightly smaller. Let us do so by defining $c_5 \triangleq \max\{c_3, c_4\}$ and

$$a_i \triangleq \frac{i}{\ell} + c_5 \ell^{-1/2} \log \ell. \tag{100}$$

From (97) and (63), we have that

$$\mathbb{E}_K\big[1 - f_{K,i}(a_i)\big] = 1 - \mathcal{F}_i(a_i) < (2\ell^{4+\log \ell})^{-1}. \tag{101}$$

From Markov's inequality, we deduce that

$$\mathbb{P}\big\{f_{K,i}(a_i) \leqslant 1 - \frac{1}{\ell^{2+\log \ell}}\big\} = \mathbb{P}\big\{1 - f_{K,i}(a_i) \geqslant \frac{1}{\ell^{2+\log \ell}}\big\} \leqslant \frac{\mathbb{E}_K\big[1 - f_{K,i}(a_i)\big]}{1/\ell^{2+\log \ell}} \leqslant \frac{1}{2\ell^2} \ . \tag{102}$$

Define

$$\mathcal{A}_i \triangleq \big\{K \in \mathbb{F}_2^{\ell \times \ell} \big| K \text{ is nonsingular and } f_{K,i}(a_i) \geqslant 1 - \frac{1}{\ell^{2+\log \ell}}\big\} \ . \tag{103}$$

Therefore, (102) can be re-written as

$$\mathbb{P}\{K \in \mathcal{A}_i\} \geqslant 1 - \frac{1}{2\ell^2} \ . \tag{104}$$

Similarly, set

$$b_i \triangleq \frac{i}{\ell} - c_5 \ell^{-1/2} \log \ell, \tag{105}$$

and define

$$\mathcal{B}_i \triangleq \big\{K \in \mathbb{F}_2^{\ell \times \ell} \big| K \text{ is nonsingular and } f_{K,i}(b_i) \leqslant \frac{1}{\ell^{2+\log \ell}}\big\}. \tag{106}$$

A very similar use of Markov's inequality shows that

$$\mathbb{P}\{K \in \mathcal{B}_i\} \geqslant 1 - \frac{1}{2\ell^2} \ . \tag{107}$$

Then, define

$$\mathcal{D} = \left(\cap_{j=1}^{\ell} \mathcal{A}_j\right) \cap \left(\cap_{j=1}^{\ell} \mathcal{B}_j\right). \tag{108}$$

By union bound, we obtain that

$$\mathbb{P}\{K \in \mathcal{D}\} \geqslant 1 - \sum_{i=1}^{\ell} \mathbb{P}\{K \notin \mathcal{A}_i\} - \sum_{i=1}^{\ell} \mathbb{P}\{K \notin \mathcal{B}_i\} \geqslant 1 - \frac{2\ell}{2\ell^2} = 1 - \frac{1}{\ell}\ . \tag{109}$$

We assume that $K \in \mathcal{D}$ throughout the remainder of proof. This implies that, for $i \in [\ell]$,

$$\begin{cases} f_{K,i}(z) \geqslant 1 - \frac{1}{\ell^{2+\log \ell}}, & \text{for } z = \frac{i}{\ell} + c_5 \ell^{-1/2} \log \ell \\ f_{K,i}(z) \leqslant \frac{1}{\ell^{2+\log \ell}}, & \text{for } z = \frac{i}{\ell} - c_5 \ell^{-1/2} \log \ell \end{cases}. \tag{110}$$

As $f_{K,i}(z)$ is an increasing function of z, (110) is equivalent to

$$\begin{cases} f_{K,i}(z) \geqslant 1 - \frac{1}{\ell^{2+\log \ell}}, & \text{for } z \geqslant \frac{i}{\ell} + c_5 \ell^{-1/2} \log \ell \\ f_{K,i}(z) \leqslant \frac{1}{\ell^{2+\log \ell}}, & \text{for } z \leqslant \frac{i}{\ell} - c_5 \ell^{-1/2} \log \ell \end{cases}. \tag{111}$$

Given these concentration results, we can proceed to the second step of the proof. Let us define

$$\begin{aligned} T_0(z,\ell) &\triangleq z\ell - c_5 \ell^{1/2} \log \ell, \\ T_1(z,\ell) &\triangleq z\ell + c_5 \ell^{1/2} \log \ell. \end{aligned} \tag{112}$$

Note that

$$\frac{i}{\ell} - c_5 \ell^{-1/2} \log \ell < z < \frac{i}{\ell} + c_5 \ell^{-1/2} \log \ell \iff z\ell - c_5 \ell^{1/2} \log \ell < i < z\ell + c_5 \ell^{1/2} \log \ell. \tag{113}$$

Therefore, for any $z \in (0,1)$, the number of indices i such that $f_{K,i}(z)$ does not satisfy (111) is upper bounded by

$$T_1(z,\ell) - T_0(z,\ell) = 2c_5 \ell^{1/2} \log \ell. \tag{114}$$

We can re-write $\lambda_{\alpha,K}(z)$ which was defined earlier in (41) as

$$\lambda_{\alpha,K}(z) = \frac{\frac{1}{\ell} \sum\limits_{i \in (T_0(z,\ell), T_1(z,\ell))} g_\alpha(f_{K,i}(z))}{g_\alpha(z)} + \frac{\frac{1}{\ell} \sum\limits_{i \notin (T_0(z,\ell), T_1(z,\ell))} g_\alpha(f_{K,i}(z))}{g_\alpha(z)}. \tag{115}$$

By using (111), we have that, for any $i \notin \left(T_0(z,\ell), T_1(z,\ell)\right)$,

$$g_\alpha(f_{K,i}(z)) \leqslant g_\alpha\left(\frac{1}{\ell^{2+\log \ell}}\right) < \left(\ell^{-2-\log \ell}\right)^\alpha. \tag{116}$$

By combining (116) with the trivial upper bound of $g_\alpha(f_{K,i}(z)) \leqslant 1$ for the left summation, and upper bounding the number of indices not in (T_0, T_1) by ℓ for the other summation, we obtain that

$$\lambda_{\alpha,K}(z) \leqslant \frac{\frac{1}{\ell} 2c_5 \ell^{1/2} \log \ell}{g_\alpha(z)} + \frac{\left(\ell^{-2-\log \ell}\right)^\alpha}{g_\alpha(z)}. \tag{117}$$

Furthermore, note that, for any $z \in (1/\ell^2, 1-1/\ell^2)$,

$$g_\alpha(z) \geqslant \left(\ell^{-2}(1-\ell^{-2})\right)^\alpha. \tag{118}$$

By combining (117) and (118), we have that

$$\lambda_{\alpha,K}(z) \leqslant \frac{1}{(1-\ell^{-2})^\alpha}\left(2c_5\ell^{-1/2+2\alpha}\log\ell + \ell^{-\alpha\log\ell}\right). \tag{119}$$

Given that $\alpha = 1/\log\ell$, we can simplify (117) according to the following.

$$\ell^{2\alpha} = (\ell^\alpha)^2 = 4, \text{ and } \ell^{-\alpha\log\ell} = \ell^{-1}. \tag{120}$$

Also, $(1-\ell^{-2})^{-\alpha}$ is a decreasing function with ℓ for $\ell \geqslant 2$ and an increasing function with α for $\alpha \leqslant 1$, which attains its maximum at $(\alpha, \ell) = (1, 2)$. That is

$$(1-\ell^{-2})^{-\alpha} \leqslant 4/3 < 2, \tag{121}$$

By applying the inequalities in (120) and (121) to (119), we finally obtain that

$$\lambda_{\alpha,K}(z) \leqslant 4c_5\ell^{-1/2+2\alpha}\log\ell + 2\ell^{-\alpha\log\ell} \leqslant 16c_5\ell^{-1/2}\log\ell + \ell^{-1} = O(\ell^{-1/2}\log\ell), \tag{122}$$

which establishes the existence of the universal constant c in (52) and concludes the proof.

□

Proof of (53). The proof of the tail intervals also follows from analyzing the average erasure probabilities. We present the proof mainly for the lower tail, where $z \in (0, 1/\ell^2)$. Similar arguments yield the proof for the upper tail.

Let $K \in GL(\ell, \mathbb{F}_2)$ denote an $\ell \times \ell$ random non-singular kernel. Define an indicator random variable $X_{i,s}$ as

$$X_{i,s} = X_{i,s}(K) = \mathbf{1}(\mathcal{R}_{\ell-s} \cap (E_i \setminus E_{i-1}) = \emptyset), \tag{123}$$

where $\mathcal{R}_{\ell-s}$ is the linear span of the first $\ell - s$ columns in K and $1 \leqslant i \leqslant \ell$, and $0 \leqslant s \leqslant \ell$. Given that K is non-singular, $\mathcal{R}_{\ell-s}$ represents a random subspace of dimension $\ell - s$ in $\mathbb{F}_2^{\ell-s}$. By recalling the inequality in (86), we have

$$\mathbb{E}[X_{i,s}] = p_{i|s} \leqslant 3\left(\frac{2}{3}\right)^{i-s}. \tag{124}$$

To establish a concentration result for $X_{i,s}$, we use Markov's inequality to get

$$\mathbb{P}\left(X_{i,s} \geqslant (1/3)\ell^2(\log\ell)(2/3)^{i-s}\right) \leqslant \frac{9}{\ell^2(\log\ell)}. \tag{125}$$

Next, we apply the union bound over all values of i and s to get

$$\begin{aligned}\mathbb{P}\left(X_{i,s} < (1/3)\ell^2(\log\ell)(2/3)^{i-s},\ \forall i,s\right) &= 1 - \mathbb{P}\left(\cup_{i,s} X_{i,s} \geqslant (1/3)\ell^2(\log\ell)(2/3)^{i-s}\right) \\ &\geqslant 1 - \ell(\ell+1).\frac{9}{\ell^2(\log\ell)} = 1 - o(1).\end{aligned} \tag{126}$$

Let us define the random variable $f_{K,i}(z)$ as

$$f_{K,i}(z) = \sum_{s=1}^{\ell} X_{i,s}(K) \binom{\ell}{s} z^s (1-z)^{\ell-s}. \tag{127}$$

We can invoke the inequality in (126) to deduce that with probability at least $1 - o(1)$ over the choice of K, we have

$$\begin{aligned} f_{K,i}(z) &\leqslant \sum_{s=1}^{\ell} (1/3)\ell^2 (\log \ell) \left(\frac{2}{3}\right)^{i-s} \binom{\ell}{s} z^s (1-z)^{\ell-s} \\ &= (1/3)\ell^2 (\log \ell) \left(\frac{2}{3}\right)^{i} \sum_{s=1}^{\ell} \binom{\ell}{s} \left(\frac{3z}{2}\right)^{s} (1-z)^{\ell-s}. \end{aligned} \tag{128}$$

Note that

$$\begin{aligned} \sum_{s=1}^{\ell} \binom{\ell}{s} \left(\frac{3z}{2}\right)^{s} (1-z)^{\ell-s} &= \left(1+\frac{z}{2}\right)^{\ell} - (1-z)^{\ell} \\ &< (1+z)^{\ell} - (1-z)^{\ell} \\ &\leqslant 2\ell z (1+z)^{\ell-1}, \end{aligned} \tag{129}$$

where the last inequality in (129) comes from the mean-value theorem for the function $f(x) = (1+x)^{\ell}$, which states that $f(x) - f(-x) = (x-(-x))f'(x_0)$ for some $x_0 \in [-x, x]$. Thus, for any $x \in (0,1)$, there exists x_0 in $(-x, x)$ such that $f(x) = 2\ell x(1+x_0)^{\ell-1} \leqslant 2\ell x(1+x)^{\ell-1}$, since $(1+x)^{\ell-1}$ is increasing with x for $x \geqslant 0$.

Next, we point out that, for any $z < \ell^{-2}$ and any $\ell \geqslant 32$, we have

$$(1+z)^{\ell-1} \leqslant \left(1+\frac{1}{\ell^2}\right)^{\ell-1} \leqslant \left(1+\frac{1}{\ell^2}\right)^{\frac{\ell^2}{\ell-1}} < \exp(1/(\ell-1)) < 9/8. \tag{130}$$

Now, we replace (129) and (130) in (128) to deduce that with probability at least $1 - o(1)$ over the choice of K, we have

$$f_{K,i}(z) < (3/4)\ell^3 (\log \ell) \left(\frac{2}{3}\right)^{i} z. \tag{131}$$

For such kernels, we can use (131) to derive the following upper bound on $\lambda_{\alpha,K}(z)$ for any $z \in (0, 1/\ell^2)$:

$$\begin{aligned} \lambda_{\alpha,K}(z) &= \frac{\frac{1}{\ell}\sum_{i=1}^{\ell} g_\alpha(f_{K,i}(z))}{g_\alpha(z)} = \frac{1}{\ell}\sum_{i}^{\ell} \frac{\left(f_{K,i}(z)(1-f_{K,i}(z))\right)^{\alpha}}{\left(z(1-z)\right)^{\alpha}} \\ &< \frac{1}{\ell}\sum_{i=1}^{\ell} \left(f_{K,i}(z)\right)^{\alpha} z^{-\alpha}(1-z)^{-\alpha} < \ell^{3\alpha-1}(\log \ell)^{\alpha} \left(\frac{3/4}{1-z}\right)^{\alpha} \left(\sum_{i=1}^{\ell} \left(\frac{2}{3}\right)^{i\alpha}\right). \end{aligned} \tag{132}$$

Given that $\alpha > 0$ and $z < 1/\ell^2 < 1/4$, we have

$$\left(\frac{3/4}{1-z}\right)^{\alpha} < 1. \tag{133}$$

Furthermore,

$$\sum_{i=1}^{\ell}\left(\frac{2}{3}\right)^{i\alpha} < \sum_{i=1}^{\infty}\left(\frac{2}{3}\right)^{i\alpha} \overset{(a)}{=} \left.\frac{x}{1-x}\right|_{x=(2/3)^{\alpha}} = \left.\frac{1}{x-1}\right|_{x=(3/2)^{\alpha}} \overset{(b)}{\leqslant} \frac{1}{\ln(3/2)^{\alpha}} = \frac{1}{\alpha\ln(3/2)}, \tag{134}$$

where (a) comes from the power series expansion, and (b) is because of $\ln(x) \leqslant x-1$ for all $x > 0$. Moreover, given that $\alpha = 1/\log\ell$, we obtain that

$$\ell^{3\alpha} = 8 \text{ and } (\log\ell)^{\alpha} = (\log\ell)^{1/\log\ell} < 2. \tag{135}$$

By combining (132), (133), (134), and (135), we conclude that, for any $z \in (0, 1/\ell^2)$,

$$\lambda_{\alpha,K}(z) < \frac{16}{\ln(3/2)}\ell^{-1}\log\ell = O(\ell^{-1}\log\ell), \tag{136}$$

which yields the desired bound on the lower tail. By following steps similar to (125)-(136), we can also show that, for any $z \in (1-1/\ell^2, 1)$,

$$\lambda_{\alpha,K}(z) \leqslant c\ell^{-1}\log\ell, \tag{137}$$

for some universal constant c with probability at least $1-o(1)$ over the choice of the kernel. By combining (136) and (137) and using one last union bound, we conclude that as ℓ grows, we have

$$\mathbb{P}\left\{\lambda_{\alpha,K}(z) < c\ell^{-1}\log\ell, \ \ \forall z \in \left(0, \frac{1}{\ell^2}\right) \cup \left(1-\frac{1}{\ell^2}, 1\right)\right\} > 1-o(1). \tag{138}$$

□

5. Discussion and Open Problems

This paper concerns the case of transmission over the binary erasure channel (BEC). One natural question is whether our results can be extended to the transmission over any binary memoryless symmetric channel (BMSC). After a preliminary version of this manuscript has appeared in [42,43], the question above has been resolved in [13]. In the rest of this section, we go over some unsolved challenges in the context of large-kernel polar codes. Most of these problems are initiated by the requirements on the size of the kernel. It was already mentioned that ℓ scales exponentially with the inverse of gap to the optimal scaling exponent, $\mu = 2$. This forces ℓ to be extremely large even for moderately good scaling exponents. In the following we address multiple issues with large ℓs:

- *Computation of the scaling exponent.* The computation of the scaling exponent even for the binary erasure channel is NP-hard [28]. While there are methods to improve the efficiency of these calculations for small values of ℓ, we are not aware of any algorithm that can do it for arbitrary 64×64 kernels.
- *Explicit construction of fast polarizing kernels.* In this paper, we showed that, given sufficiently large ℓ, almost all binary non-singular $\ell \times \ell$ matrices are suitable polarization kernel candidates. However, the problem of finding one, or a family, of such kernels remains unsolved. Note that exhaustive search only works up to $\ell \sim 8$, while there are a few heuristic construction algorithms for $\ell = 16, 32$, and 64.
- *Construction of polar codes.* In the polar coding terminology, the construction problem refers to the problem of finding the best bit-channels for which we transmit information over. There are multiple known algorithms for classical polar codes including the Tal-Vardy method in [25] and the Gaussian Approximation

in [38]. Unfortunately, there is yet another computation complexity blow-up if one replaces the 2×2 kernels with arbitrarily large $\ell \times \ell$ matrices, which leaves us with the Monte-Carlo method for finding less noisy bit-channels. However, this method is known to perform poorly in the precision/complexity trade-off.

• *Decoding complexity.* The recursive implementation of successive-cancellation decoding for polar codes is based on the butterfly-like graph, where each node represents a polarization kernel. These kernel-nodes perform successive-cancellation decoding of the kernel itself, and then communicate with each other on a specific schedule to reveal the uncoded information bits sequentially and efficiently. It is well known that the overall decoding complexity for conventional polar codes is $O(n \log n)$. However, the internal successive-cancellation computations within the kernels become more complicated when the 2×2 conventional kernel is replaced with an $\ell \times \ell$ kernel. This effectively changes the asymptotic decoding complexity to $O(2^{\ell} n \log n)$. This is probably the most controversial problem with using polar codes constructed from large kernels if the underlying channel is not a BEC. Moreover, in the case of BEC, decoding can be accomplished by using Gaussian Elimination. This raises the main question about practicality of polar codes with large kernels. Recently, there have been multiple attempts at finding/constructing fast-polarizing kernels with enough structure that would allow us to design a decoding algorithm with reasonable decoding complexity. See for example [39,40]. However, a general approach to reducing the decoding complexity of large kernels is still lacking from the literature.

Despite all these problems, we view the fact that polar codes constructed from random large kernels perform nearly as good as the random codes to be of theoretical interest. That is, we have shown that polarization kernels with optimal scaling for the BEC exist. The problem of finding practical such kernels is a topic of further research.

References

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423 and pp. 623–656, Apr./Oct. 1948.

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

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

[4] R. L. Dobrushin, "Mathematical problems in the Shannon theory of optimal coding of information," in *Proc. 4th Berkeley Symp. Mathematics, Statistics, and Probability*, vol. 1, 1961, pp. 211–252.

[5] V. Strassen, Asymptotische abschätzungen in Shannon's informationstheorie, in *Trans. 3rd Prague Conf. Inf. Theory*, pp. 689–723, 1962.

[6] M. Hayashi, "Information spectrum approach to second-order coding rate in channel coding," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 4947–4966, Nov. 2009.

[7] J. P. Tillich, and G. Zémor, "Discrete isoperimetric inequalities and the probability of a decoding error," *Combinatorics, Probability and Computing*, vol. 9, pp. 465–479, Sept. 2000.

[8] A. Montanari, "Finite size scaling and metastable states of good codes," in *Proc. Allerton Conf. on Commun., Control, and Comput.*, Monticello, IL, USA, Oct. 2001.

[9] E. R. Berlekamp, R. J. McEliece, and H. C. A. van Tilborg, "On the inherent intractability of certain coding problems," *IEEE Trans. Inf. Theory*, vol. 24, no. 3, pp. 384–386, May 1978.

[10] G. D. Forney, Jr., *Concatenated Codes*, PhD thesis, MIT, 1966.

[11] V. Guruswami and P. Xia, "Polar codes: Speed of polarization and polynomial gap to capacity," in *Proc. 54-th Annual IEEE Symp. Foundations of Computer Science* (FOCS), Berkeley, CA, Oct. 2013.

[12] V. Guruswami and P. Xia, "Polar codes: Speed of polarization and polynomial gap to capacity," *IEEE Trans. Inf. Theory*, vol. 61, no. 1, pp. 3–16, Jan. 2015.

[13] V. Guruswami, A. Riazanov and M. Ye, "Arikan meets Shannon: polar codes with near-optimal convergence to channel capacity," in *Proc. 52-nd Annual ACM SIGACT Symp. on Theory of Computing* (STOC), 2020, pp. 552–564.

[14] D. Goldin and D. Burshtein, "Improved bounds on finite length scaling of polar codes," *IEEE Trans. Inf. Theory*, vol. 60, no. 11, pp. 6966–6978, Nov. 2014.

[15] S.H. Hassani, K. Alishahi, and R.L. Urbanke, "Finite-length scaling for polar codes," *IEEE Trans. Inf. Theory*, vol. 60, no. 10, pp. 5875–5898, Oct. 2014.

[16] S. B. Korada, A. Montanari, E. Telatar, and R.L. Urbanke, "An empirical scaling law for polar codes," in *Proc. IEEE Intern. Symp. Inf. Theory*, Austin, TX, June 2010, pp. 884–888.

[17] M. Mondelli, S. H. Hassani, and R. L. Urbanke, "Unified scaling of polar codes: error exponent, scaling exponent, moderate deviations, and error floors," *IEEE Trans. Inf. Theory*, vol. 62, no. 12, pp. 6698–6712, Dec. 2016.

[18] S. Kudekar, T. J. Richardson, and R. L. Urbanke, "Spatially coupled ensembles universally achieve capacity under belief propagation," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 7761–7813, Dec. 2013.

[19] M. Mondelli, S. H. Hassani, and R. L. Urbanke, "How to achieve the capacity of asymmetric channels," *IEEE Trans. Inf. Theory*, vol. 64, no. 5, pp. 3371–3393, May 2018.

[20] S. Kudekar, S. Kumar, M. Mondelli, H. D. Pfister, E. Şaşoğlu, and R. L. Urbanke, "Reed-Muller codes achieve capacity on erasure channels," in *Proc. of the Annual ACM Symposium on Theory of Computing (STOC)*, Boston, MA, USA, June 2016, pp. 658–669.

[21] S. Kudekar, S. Kumar, M. Mondelli, H. D. Pfister, E. Şaşoğlu, and R. L. Urbanke, "Reed-Muller codes achieve capacity on erasure channels," *IEEE Trans. Inf. Theory*, vol. 63, no. 7, pp. 4298–4316, July 2017.

[22] M. Mondelli, S. H. Hassani, and R. L. Urbanke, "From polar to Reed-Muller codes: A technique to improve the finite-length performance," *IEEE Trans. Commun.*, vol. 62, no. 9, pp. 3084–3091, Sept. 2014.

[23] A. Amraoui, A. Montanari, T. Richardson, and R. L. Urbanke, "Finite-length scaling for iteratively decoded LDPC ensembles," *IEEE Trans. Inf. Theory*, vol. 55, no. 2, pp. 473–498, Feb. 2009.

[24] E. Şaşoğlu, "Polarization and polar codes, " *Foundations and Trends in Communications and Information Theory*, vol. 8, no. 4, pp. 259–381, Oct. 2012.

[25] I. Tal and A. Vardy, How to construct polar codes, *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6562–6582, Oct. 2013.

[26] S. B. Korada, E. Şaşoğlu, and R. L. Urbanke, "polar codes: Characterization of exponent, bounds, and constructions," *IEEE Trans. Inf. Theory*, vol. 56, no. 12, pp. 6253-6264, Dec. 2010.

[27] S. H. Hassani, *Polarization and Spatial Coupling: Two Techniques to Boost Performance,* Ph.D. dissertation, EPFL, Lausanne, Switzerland, Mar. 2013.

[28] A. Fazeli and A. Vardy, "On the scaling exponent of binary polarization kernels," in *Proc. of the Allerton Conf. on Commun., Control, and Computing*, Monticello, IL, USA, Oct. 2014, pp. 797–804.

[29] A. Fazeli, *New Frontiers in Polar Coding: Large Kernels, Convolutional Decoding, and Deletion Channels,* Ph.D. dissertation, University of California-San Diego, June 2018.

[30] E. Arıkan and I. E. Telatar, "On the rate of channel polarization," in *Proc. of the IEEE Int. Symposium on Inf. Theory*, Seoul, South Korea, July 2009, pp. 1493–1495.

[31] S. H. Hassani, R. Mori, T. Tanaka, and R. L. Urbanke, "Rate-dependent analysis of the asymptotic behavior of channel polarization," *IEEE Trans. Inf. Theory*, vol. 59, no. 4, pp. 2267–2276, Apr. 2013.

[32] A. Eslami and H Pishro-Nik, "On finite-length performance of polar codes: stopping sets, error floor, and concatenated design," *IEEE Trans. Commun.*, vol. 61, no. 3, pp. 919–929, Mar. 2013.

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

[34] M. Mondelli, S. H. Hassani, and R. L. Urbanke, "Scaling exponent of list decoders with applications to polar codes," *IEEE Trans. Inf. Theory*, vol. 61, no. 9, pp. 4838–4851, Sept. 2015.

[35] H. D. Pfister, and R. L. Urbanke, "Near-optimal finite-length scaling for polar codes over large alphabets," *IEEE Trans. Inf. Theory*, vol. 65, no. 9, pp. 5643–5655, Sept. 2019.

[36] A. Vardy and Y. Be'ery, "Maximum-likelihood soft decision decoding of BCH codes," *IEEE Trans. Inf. Theory*, vol. 40, no. 2, pp. 546–554, Mar. 1994.

[37] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58 , pp. 13–30, Mar. 1963.

[38] P. Trifonov, "Efficient design and decoding of polar codes," *IEEE Trans. Commun.*, vol. 60, no. 11, pp. 3221–7, Aug. 2012.

[39] S. Buzaglo, A. Fazeli, P. Siegel, V. Taranalli, and A. Vardy, "Permuted successive cancellation decoding for polar codes," in *Proc. 2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 2618–2622.

[40] G. Trofimiuk and P. Trifonov "Construction of binary polarization kernels for low complexity window processing," in *Proc. 2019 IEEE Information Theory Workshop (ITW)*, Aug. 2019, pp. 1–5.

[41] K. L. Chung, *A Course in Probability Theory*. 3rd ed. San Diego: Academic Press, 2001.

[42] A. Fazeli, H. Hassani, M. Mondelli, and A. Vardy, "Binary Linear Codes with Optimal Scaling and Quasi-Linear Complexity," arxiv.org/abs/1711.01339v1, preprint of Nov. 3, 2017.

[43] A. Fazeli, H. Hassani, M. Mondelli, and A. Vardy, "Binary linear codes with optimal scaling: Polar codes with large kernels," in *Proc. 018 IEEE Information Theory Workshop (ITW)*, Nov. 2018.