Accelerating Polarization via Alphabet Extension

Iwan Duursma[†], Ryan Gabrys[*], Venkatesan Guruswami[‡], Ting-Chun Lin[*§], Hsin-Po Wang[*]
[*]University of California San Diego, CA, USA. Emails: {hsw001, til022, rgabrys} @ucsd.edu
[†]University of Illinois Urbana-Champaign, IL, USA. Email: duursma@illinois.edu
[‡]University of California, Berkeley, CA, USA. Email: venkatg@berkeley.edu
[§]Hon Hai (Foxconn) Research Institute, Taipei, Taiwan.

Abstract

Polarization is an unprecedented coding technique in that it not only achieves channel capacity, but also does so at a faster speed of convergence than any other technique. This speed is measured by the "scaling exponent" and its importance is three-fold. Firstly, estimating the scaling exponent is challenging and demands a deeper understanding of the dynamics of communication channels. Secondly, scaling exponents serve as a benchmark for different variants of polar codes that helps us select the proper variant for real-life applications. Thirdly, the need to optimize for the scaling exponent sheds light on how to reinforce the design of polar code.

In this paper, we generalize the binary erasure channel (BEC), the simplest communication channel and the protagonist of many polar code studies, to the "tetrahedral erasure channel" (TEC). We then invoke Mori–Tanaka's 2×2 matrix over $\mathbb{F}_4$ to construct polar codes over TEC. Our main contribution is showing that the dynamic of TECs converges to an almost–one-parameter family of channels, which then leads to an upper bound of 3.328 on the scaling exponent. This is the first non-binary matrix whose scaling exponent is upper-bounded. It also polarizes BEC faster than all known binary matrices up to 23×23 in size. Our result indicates that expanding the alphabet is a more effective and practical alternative to enlarging the matrix in order to achieve faster polarization.

1 Introduction

A fundamental question at the center of the theory of communication is whether we can fully utilize a noisy channel to transmit information. In modern terminology, can error correcting codes achieve channel capacity? The answer is positive; in fact, multiple code constructions do so. Among them, polar code is a special one as it achieves capacity faster than any other known code.

Polar coding was invented by Arıkan around 2008 [Ari09]. During that time, Arıkan was experimenting with channel combining and splitting. By treating two independent binary channels as a single quaternary channel (combining) and tasking ourselves with guessing certain linear combinations of the inputs (splitting), he synthesized two channels, denoted by $W^{\square}$ and $W^{\bigcirc}$, out of the original channel W. Arıkan realized that, when combining and splitting is applied recursively, the channels undergo an intriguing dynamic that ultimately results in most synthetic channels being either almost noiseless or extremely noisy. This is *channel polarization*, the first ingredient underlying polar codes.

The second ingredient of polar codes, also given by Arıkan in said seminal paper, is the relation between the dynamic of synthetic channels and the construction and performance of the code. Arıkan's insight was that synthetic channels that become almost noiseless can be used to transmit information bits, and synthetic channels that become extremely noisy can be "frozen" to some fixed values. The rate at which we communicate meaningful bits is then the proportion of synthetic channels that are almost noiseless. So, whether we can achieve channel capacity becomes a problem of counting the numbers of good and bad synthetic channels.

It then became apparent, perhaps even appealing, that one can study the dynamic of synthetic channels by means of stochastic processes. Take the *binary erasure channel* (BEC) as an example. Let W be BEC(ε), the BEC with erasure probability ε, where $0 < \varepsilon < 1$. The channels $W^{\square}$ and $W^{\bigcirc}$ are BEC($2\varepsilon - \varepsilon^2$) and BEC($\varepsilon^2$), respectively. A process $\{H_n\}_n$ is thus defined by having $H_0 \coloneqq \varepsilon$ and $H_{n+1} \coloneqq 2H_n - H_n^2$ or H_n^2 with equal probability. It can be shown that if

$$\mathbb{P}\{\, H_n \leqslant f(n) \,\} = 1 - H_0 - g(n),$$

where $f, g > 0$ are functions in n, then there is a polar code of length 2^n, miscommunication probability $2^n f(n)$, and gap to capacity $g(n)$.

It was at this point that the study of polar codes branched. On one branch, called the *error exponent regime*, g is a constant and the asymptotics of f is examined. On the other branch, called the *scaling exponent regime*, f is a constant[1] and the asymptotics of g is examined. On the error exponent branch, it was shown that $f(n)$ is roughly $\exp(-e^{\beta n})$, where $\beta > 0$ is a constant depending on the matrix used in the code construction. The task of determining β for each matrix has been fully resolved; interested readers are referred to [AT09; KŞU10; HMTU13; MT14].

On the scaling exponent branch, making progress is harder and slower. For BECs, [HAU10; KMTU10] managed to estimate that $g(n) \approx 2^{-n/3.527}$. For binary memoryless symmetric (BMS) channels, it was first shown that $g(n) < 2^{-n/\mu}$ for some constant $0 < \mu < \infty$ [GX15]. This makes polar codes the only known code family that converges to capacity at a polynomial rate in the block length. More realistic estimates of μ were given later: $3.553 < \mu$ [GHU12], $3.579 < \mu < 6$ [HAU14], $\mu < 5.702$ [GB14], $\mu < 4.714$ [MHU16], and very recently $\mu < 4.63$ [WLVG22]. Now that we know the μ for polar code and the optimal value being $\mu \approx 2$ for random code [BKB04; Hay09; PPV10], the discrepancy begs the question: Can one modify polar code to reach a smaller scaling exponent?

The answer is positive: Arıkan used the matrix $\left[\begin{smallmatrix}1&0\\1&1\end{smallmatrix}\right]$ (this is called the *kernel* in literature) to combine and split channels. Instead, one can use a larger matrix, for instance

$$\begin{bmatrix} 1&0&0&0&0&0&0&0\\ 1&1&0&0&0&0&0&0\\ 1&0&1&0&0&0&0&0\\ 1&0&0&1&0&0&0&0\\ 1&1&1&0&1&0&0&0\\ 1&1&0&1&0&1&0&0\\ 1&0&1&1&0&0&1&0\\ 1&1&1&1&1&1&1&1 \end{bmatrix},$$

to combine and split channels. In [FV14; YFV19; TT21; Tro21; BBL20; Lin21], binary matrices ranging from 3×3 to 64×64 are deployed and the scaling exponents over BECs are estimated. The best scaling exponent up to every matrix size is plotted in Figure 2. There are also meta-asymptotic results stating that $\mu \approx 2$ can be achieved using larger and larger matrices. This statement was proved over q-ary erasure channels [PU16], binary erasure channels [FHMV21], all BMS channels [GRY22], and finally discrete memoryless channels [WD21].

As much as we want to lower polar code's scaling exponent, there is one caveat that renders large matrices impractical: the smallest matrix whose scaling exponent is strictly better than $\left[\begin{smallmatrix}1&0\\1&1\end{smallmatrix}\right]$ is the 8×8 matrix above. Using this matrix takes twice more time to decode (estimate based on

[1] Not always; sometimes $f \to 0$ but only exponentially fast in n. Note that $2^n f(n)$, the upper bound on the miscommunication probability, is allowed to exceed 1, so the corresponding code can be meaningless. Yet the asymptotics of g capture the behaviors of other meaningful codes.

the method of [BFSTV17]), whereas the benefit we gain is that μ slightly decreases from 3.627 to 3.577. As the matrix gets larger and deviates more from the tensor powers of $\left[\begin{smallmatrix}1&0\\1&1\end{smallmatrix}\right]$, the time complexity grows drastically. For this reason, it is unlikely that we will ever see polar code based on large matrices (unless it is for other concerns [BGLB20]).

Large matrix aside, many other techniques emerge with empirical evidence that they improve polar code—concatenation, cyclic redundancy check, and list decoder to name a few. But none of them sees a proof of improvements in the scaling exponent; in fact, quite the opposite was reported [MHU15]. So we are back to the starting point where we want to improve polar codes' scaling exponent while minimizing the complexity penalty.

One approach that seems promising, albeit very little is known due to its innate technical difficulty, is to use a non-binary input alphabet. This line of research started from Şaşoğlu [ŞTA09; Şaş12; Chi14], wherein the goal was to find at least one way to polarize arbitrary finite alphabets regardless of the speed. In particular, the usual matrix $\left[\begin{smallmatrix}1&0\\1&1\end{smallmatrix}\right]$ is known to polarize prime fields. Later, Sahebi–Pradhan [SP11] and Park–Barg [PB13] showed that $\left[\begin{smallmatrix}1&0\\1&1\end{smallmatrix}\right]$ cannot polarize non-prime fields. Then, Mori–Tanaka [MT14] classified all matrices that can polarize finite fields (i.e., the alphabet size must be a prime power). One step forward, Nasser [Nas16] classified all binary operators (i.e., bivariate functions) that can polarize arbitrary finite alphabets. In [BGNRS22; BGS18], the authors showed that, for any polarizing matrix over prime fields, one has $\mu < \infty$. In [WD21], the authors showed that $\mu \approx 2$ is reachable over arbitrary finite alphabets.

Why is a non-binary input alphabet attractive? There are at least three reasons. First, modulation[2]: For quadrature amplitude modulation (QAM) and amplitude and phase-shift keying (APSK), a constellation point is more likely to be confused with constellation points nearer to it. A non-binary channel models this proximity relation more naturally than a series of correlated binary channels do [SSSH13; CKW19]. Second, two-stage polarization: If we weakly-polarize a binary channel with $\left[\begin{smallmatrix}1&0\\1&1\end{smallmatrix}\right]$, treat two binary channels as a quaternary channel, and strongly-polarize the quaternary channel with the 4×4 Reed–Solomon matrix, we can improve the asymptotics of $f(n)$ from $\exp(-2^{0.5n})$ to $\exp(-2^{0.5731n})$ [PSL16] (see also [AMV22; CBM18]). Third, and most importantly, scaling exponent: Several works have observed that non-binary matrices of the form $\left[\begin{smallmatrix}1&0\\\omega&1\end{smallmatrix}\right]$ just polarize faster than $\left[\begin{smallmatrix}1&0\\1&1\end{smallmatrix}\right]$ [YS18; LY21; Sav21]. Could it be that the non-binary scaling exponents are smaller?

Consider [RWLP22]'s technique that uses $\left[\begin{smallmatrix}1&0\\1&1\end{smallmatrix}\right]$ to polarize non-binary channels; their result has an implication that non-binary channels' scaling exponent is at least as good as binary channels'. In this paper, we aim to answer the question of whether the former is strictly better than the latter. By defining a toy model that contains a pair of BECs as a special case and estimating the scaling exponent of $\left[\begin{smallmatrix}1&0\\\omega&1\end{smallmatrix}\right]$, we provide a proof of concept result that an expansion in alphabet size does result in an improvement in scaling exponent. Recall that BECs form a one-parameter family and that this property makes its scaling behavior easy to analyze. This paper's overall strategy is to show that the descendants of a quaternary channel converge to an almost–one-parameter family; we then analyze the scaling behavior of this family and conclude the following.

Theorem 1 (main theorem)**.** *Treating a pair of BECs as a quaternary channel, the* 2×2 *matrix* $\left[\begin{smallmatrix}1&0\\\omega&1\end{smallmatrix}\right]$ *over* $\mathbb{F}_4$ *induces a scaling exponent less than* 3.451*. Here,* $\omega^2 + \omega + 1 = 0$*.*

This paper is organized as follows. Section 2 reviews the essence of polar code. Section 3 defines tetrahedral erasure channels (TECs), defines balanced TECs to be those that possess some

[2]Modulation means translating digital signals to analog signals. A digital signal will be mapped to a point on the complex plane, which represents a sine wave with certain amplitude and phase; such a point is called a constellation point, the union of all points a constellation diagram.

symmetry, and defines edge-heavy TECs to be those that will be polarized faster. Section 4 defines serial combination and parallel combination that will be used to polarize TECs. Section 5 shows that unbalanced TECs tend to become very close to balanced TECs, so it suffices to consider the speed of polarization of the latter. Section 6 shows that edge-light TECs tend to become very close to edge-heavy TECs, so it suffices to consider the speed of polarization of the latter. Section 7 shows the speed of polarization of a generic TEC is faster than the classical BEC.

2 Polar Code

Readers who are familiar with polar code can safely skip this section. This section serves a simplified, high-level summary of classical polar code. More details are found in [Wan21, Chapter 2]. We assume BEC throughout the section.

Let $X \in \mathbb{F}_2$ be a random variable following the uniform distribution. Let $Y \in \mathbb{F}_2 \cup \{?\}$ be a random variable with $\mathbb{P}\{\, Y = X \,\} = 1 - \varepsilon$ and $\mathbb{P}\{\, Y = ? \,\} = \varepsilon$. Here, $\varepsilon \in [0, 1]$ is called the *erasure probability*. The pair $(X|Y)$ is called a *binary erasure channel* (BEC) and denoted by $\mathrm{BEC}(\varepsilon)$. The entropy $H(\mathrm{BEC}(\varepsilon)) = H(X|Y) = \varepsilon$ is defined through Shannon's mean.

Let $(X_1|Y_1)$ and $(X_2|Y_2)$ be two iid copies of $\mathrm{BEC}(\varepsilon)$. Define the serial combination $\mathrm{BEC}(\varepsilon)^{\square}$ to be $(X_1 + X_2|Y_1, Y_2)$. That is, what do we know about $X_1 + X_2$ when given Y_1 and Y_2? One sees that it is information theoretically equivalent to $\mathrm{BEC}(2\varepsilon - \varepsilon^2)$. Define the parallel combination $\mathrm{BEC}(\varepsilon)^{\bigcirc}$ to be $(X_1|Y_1, Y_2, X_1 + X_2)$. That is, what do we know about X_1 when given Y_1, Y_2, and $X_1 + X_2$? One sees that it is information theoretically equivalent to $\mathrm{BEC}(\varepsilon^2)$.

Serial and parallel combinations apply recursively. A polar code of block length 2^n consists of a subset of strings $\mathcal{I} \subseteq \{\square, \bigcirc\}^n$. In this code, a synthetic channel

$$\Big(\cdots\big((\mathrm{BEC}(\varepsilon)^{c_1})^{c_2}\big)\cdots\Big)^{c_n} \tag{1}$$

will be used to transmit useful information iff $(c_1, c_1, \ldots, c_n) \in \mathcal{I}$. The code rate of this polar code is $|\mathcal{I}|/2^n$. The exact miscommunication probability of this polar code is hard to find, but has an upper bound of

$$\sum_{\mathcal{I}} H\bigg(\Big(\cdots\big((\mathrm{BEC}(\varepsilon)^{c_1})^{c_2}\big)\cdots\Big)^{c_n}\bigg).$$

To define a good $\mathcal{I}$, choose a function $f(n)$ and collect all strings $(c_1, c_1, \ldots, c_n) \in \{\square, \bigcirc\}^n$ such that H(formula (1)) is less than $f(n)$. The fact that the erasure probabilities undergo simple evolutions $\varepsilon \mapsto 2\varepsilon - \varepsilon^2$ and $\varepsilon \mapsto \varepsilon^2$ motivates the following stochastic process: define $\{H_n\}_n$ by initial value $H_0 \coloneqq \varepsilon$ and evolution rule $H_{n+1} \coloneqq 2H_n - H_n^2$ or H_n^2 with equal probability. Then the code rate $|\mathcal{I}|/2^n$ coincides with $\mathbb{P}\{\, H_n \leqslant f(n) \,\}$. The gap to capacity $g(n) \coloneqq 1 - H_0 - |\mathcal{I}|/2^n = 1 - H_0 - \mathbb{P}\{\, H_n \leqslant f(n) \,\}$ is thus motivated.

In a way, the study of polar code over BEC is the study of the cdf of H_n, with emphasis put on the hard threshold at $1 - H_0$. Abusing the same logic, this paper is a study of a stochastic process $\{W_n\}_n$ that lives in $[0, 1]^5 \cap \{\, p + q + r + s + t = 1 \,\}$, which happens to have peculiar implications in coding theory.

3 A New Channel Model

We are to define a type of quaternary channels in this section. This should be the smallest possible set of quaternary channels that meet the following: (a) it should model a pair of BECs as a

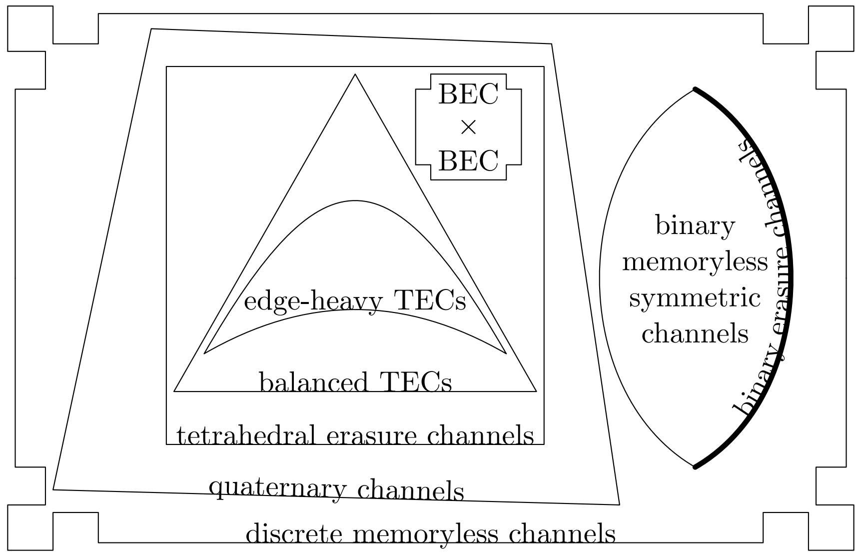

Figure 1: The Euler diagram of channels featured in this paper. The cross is the set of pairs of BECs; it will converge to the set of balanced TECs (Section 5). The balanced TECs will then converge to edge-heavy TECs (Section 6). And then edge-heavy TECs polarize faster than BECs. Note that BECs are a one-parameter family of extreme BMS channels, hence the thick curve.

special case; and (b) it should be closed under pre-processing the input using invertible linear transformations.

3.1 Tetrahedral erasure channel

Let the input alphabet be $\mathbb{F}_2^2$; and we assume the uniform input distribution throughout the paper. For any input $(x_1, x_2) \in \mathbb{F}_2^2$, the output will be in $(\mathbb{F}_2 \cup \{?\})^3$ and assume one of the following five erasure patterns:

- $(x_1, x_1 + x_2, x_2)$ with probability p;
- $(x_1, ?, ?)$ with probability q;
- $(?, x_1 + x_2, ?)$ with probability r;
- $(?, ?, x_2)$ with probability s;
- $(?, ?, ?)$ with probability t.

Here we call p, q, r, s, t the *subspace erasure probabilities* and they sum to 1. Such a channel is denoted by $\mathrm{TEC}(p, q, r, s, t)$. For brevity, we say a TEC outputs (x_1, x_2), outputs x_1, outputs $x_1 + x_2$, outputs x_2, and outputs nothing to represent the five erasure patterns.

A TEC can be related to a tetrahedron whose vertices are $(0,0,0)$, $(1,1,0)$, $(1,0,1)$, and $(0,1,1)$. Outputting (x_1, x_2) corresponds to the vertex $(x_1, x_1 + x_2, x_2)$. Outputting x_1 corresponds to the edge $(x_1, x_1, 0)$—$(x_1, 1 - x_1, 1)$. Outputting nothing corresponds to the tetrahedron per se. That is to say, a TEC takes a vertex as an input and outputs the same vertex with probability p, outputs an edge attached to that vertex with probability $q+r+s$, and output the entire tetrahedron with probability t.

There is another way to interpret a TEC. Consider $\mathbb{F}_4$ and let ω be a primitive element therein. A TEC takes $x \coloneqq x_1\omega + x_2 \in \mathbb{F}_4$ as an input and outputs x, $\operatorname{tr}(x)$, $\operatorname{tr}(\omega x)$, $\operatorname{tr}(x/\omega)$, or nothing, each with probability p, q, r, s, and t. Here, $\operatorname{tr}\colon \mathbb{F}_4 \to \mathbb{F}_2$ is the field trace. It is the matrix trace if we use the matrices $\left[\begin{smallmatrix}0&0\\0&0\end{smallmatrix}\right]$, $\left[\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right]$, $\left[\begin{smallmatrix}1&1\\1&0\end{smallmatrix}\right]$, $\left[\begin{smallmatrix}0&1\\1&1\end{smallmatrix}\right]$ to represent $0, 1, \omega, 1+\omega \in \mathbb{F}_4$.

TEC is not an ad hoc channel that we happen to know how to deal with. It relates to other channels that have been discussed in literature.

Proposition 2. *The "q-ary erasure channel with erasure probability* ε*" [MT10; PU16], when* $q = 4$*, is a TEC of the form* $\operatorname{TEC}(\,1-\varepsilon\,,\,0\,,\,0\,,\,0\,,\,\varepsilon\,)$.

Proposition 3. *When transmitting two bits* x_1 *and* x_2 *through* $\operatorname{BEC}(\delta)$ *and* $\operatorname{BEC}(\varepsilon)$*, respectively, the outputs can be simulated by* $\operatorname{TEC}(\,(1-\delta)(1-\varepsilon),\,(1-\delta)\varepsilon,\,0, \delta(1-\varepsilon),\,\delta\varepsilon\,)$.

The proofs are trivial. The propositions imply that any scaling exponent estimate for TEC immediately generalizes to 4-ary erasure channels and BECs.

3.2 Channel functionals

The *conditional entropy* (hereafter *entropy*) of a TEC is defined by the following; it is meant to be compatible with Shannon's definition:

$$H(\operatorname{TEC}(p,q,r,s,t)) \coloneqq \frac{q+r+s}{2} + t.$$

The *edge mass* of a TEC is defined by the following; it measures the "polarizability" of a TEC:

$$E(\operatorname{TEC}(p,q,r,s,t)) \coloneqq q+r+s.$$

The *Quetelet index* of a TEC W is defined by

$$Q(W) \coloneqq \frac{E(W)}{H(W)(1-H(W))}.$$

Clearly, $0 \leqslant E(W) \leqslant 2\min(H(W), 1-H(W))$ and $0 \leqslant Q(W) \leqslant 4$. We call a TEC W *edge-heavy* if $Q(W) \geqslant 2\sqrt{7}-4$. Adolphe Quetelet invented the body mass index that determines if a person is overweight or underweight. Here, we use Quetelet index to determine if a TEC possesses too much edge mass (easy to polarize) or too little (hard to polarize).

A TEC is *balanced* if $q = r = s$. Put it another way, the edges of the tetrahedron weigh the same. It is not hard to see that H and E uniquely determine a balanced TEC by

$$p = 1 - H(W) - \frac{E(W)}{2}, \qquad q = r = s = \frac{E(W)}{3}, \qquad t = H(W) - \frac{E(W)}{2}.$$

The *moment of inertia* of a TEC is defined by

$$A(\operatorname{TEC}(p,q,r,s,t)) \coloneqq (q-r)^2 + (r-s)^2 + (s-q)^2.$$

A TEC is balanced iff its moment of inertia vanishes. See also the "symmetric over the product" condition in [CLMS21] and the "equidistance" condition in [ŞTA09].

4 Channel Synthesis

TECs can be serially combined or parallelly combined as in the theory of density evolution [LH06].

4.1 Serial combination

Let $U \coloneqq \mathrm{TEC}(p,q,r,s,t)$ and $V \coloneqq \mathrm{TEC}(p',q',r',s',t')$ be two TECs. The *serial combination* of U and V is defined to be the task of guessing (u_1+v_1, u_2+v_2) given the output of inputting (u_1,u_2) into U and the output of inputting (v_1,v_2) into V. Let us go over all 25 erasure patterns that are classified into five scenarios.

Scenario one—U outputs (u_1,u_2) and V outputs (v_1,v_2): Now we know (u_1+v_1,u_2+v_2) in its entirety. This scenario happens with probability pp'.

Scenario two—U outputs u_1 with or without u_2, and V outputs v_1 with or without v_2, but either u_2 or v_2 is missing: In this case, we can infer u_1+v_1, but we cannot infer u_2+v_2. So this case feels like $(x_1,x_2) \coloneqq (u_1+v_1,u_2+v_2)$ underwent a TEC and only x_1 went through. The probability that only x_1 went through is $pq'+qq'+qp'$.

Scenario three—U outputs (u_1,u_2) or u_1+u_2, and V outputs (v_1,v_2) or v_1+v_2, but scenario one does not happen: For this case, we know neither u_1+v_1 nor u_2+v_2. But we can infer $(u_1+v_1)+(u_2+v_2)$. So this case feels like $(x_1,x_2) \coloneqq (u_1+v_1,u_2+v_2)$ underwent a TEC and only x_1+x_2 went through. The probability that only x_1+x_2 went through is $pr'+rr'+rp'$.

Scenario four—U outputs u_2 with or without u_1, and V outputs v_2 with or without v_1, but either u_1 or v_1 is missing: In this case, we can infer $x_2 \coloneqq u_2+v_2$ but not $x_1 \coloneqq u_1+v_1$. So this case feels like x_1 is erased. This scenario happens with probability $ps'+ss'+sp'$.

Scenario five—U outputs one bit (u_1 or u_1+u_2 or u_2) and V outputs one bit (v_1 or v_1+v_2 or v_2) but the erasure patterns do not match; or at least one of U and V outputs nothing: We cannot infer u_1+v_1 because either u_1 or v_1 is missing. We cannot infer u_2+v_2 because either u_2 or v_2 is missing. We cannot infer $(u_1+v_1)+(u_2+v_2)$, either. So this case feels like both $x_1 \coloneqq u_1+v_1$ and $x_1 \coloneqq u_2+v_2$ are erased, so is x_1+x_2. The probability that we learn nothing about (x_1,x_2) is $(q+r+s)(q'+r'+s')-qq'-rr'-ss'+t+t'-tt'$.

Note that these five scenarios correspond to the five erasure patterns in the definition of TEC. Denote by $U \boxast V$ the serial combination of U and V; it is a TEC with subspace erasure probabilities

$$U \boxast V \coloneqq \mathrm{TEC}(\quad pp', \quad pq'+qq'+qp', \quad pr'+rr'+rp', \quad ps'+ss'+sp', \quad 1-\text{the four to the left} \quad).$$

4.2 Parallel combination

The *parallel combination* of $U \coloneqq \mathrm{TEC}(p,q,r,s,t)$ and $V \coloneqq \mathrm{TEC}(p',q',r',s',t')$ is defined to be the task of guessing (u_1,u_2) given (u_1+v_1,u_2+v_2) (the perfect output of $U \boxast V$), the result of feeding (u_1,u_2) into U, and the result of feeding (v_1,v_2) into V.

Denote by $U \circledast V$ the parallel combination of U and V. One can go over its erasure scenarios like the previous subsection does. For instance, if U outputs u_1 and V outputs v_1+v_2, then we can infer v_1 (using u_1 and u_1+v_1), followed by v_2 (using v_1 and v_1+v_2), and finally u_2 (using v_2 and u_2+v_2); and hence we can completely recover u_1 and u_2. Details omitted, it can be shown that $U \circledast V$ is a TEC with subspace erasure probabilities

$$U \circledast V \coloneqq \mathrm{TEC}(\quad 1-\text{the four to the right}, \quad tq'+qq'+qt', \quad tr'+rr'+rt', \quad ts'+ss'+st', \quad tt' \quad).$$

Note that there is a duality between $\mathrm{TEC}(p,q,r,s,t)$ and $\mathrm{TEC}(t,s,r,q,p)$ that keeps E as is, maps H to $1-H$, and swaps parallel and serial combinations. The duality grants us the convenience of proving half of a theorem and the other half follows by symmetry.

4.3 Mori–Tanaka's twisting kernel

A 2×2 polarization *kernel* K over $\mathbb{F}_4$ is defined with a "twist" as follows: For a pair of inputs $u, v \in \mathbb{F}_4$, let K be the linear transformation that reads $(u, v) \longmapsto (u + \omega v, v)$ or, equivalently,

$$\begin{bmatrix} u & v \end{bmatrix} \longmapsto \begin{bmatrix} u & v \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \omega & 1 \end{bmatrix}.$$

This kernel was studied by Mori–Tanaka [MT14] and is shown to be polarizing. If we treat $\mathbb{F}_4$ as $\mathbb{F}_2^2$, then K reads $\bigl((u_1, u_2), (v_1, v_2) \bigr) \longmapsto \bigl((u_1 + v_1 + v_2, u_2 + v_1), (v_1, v_2) \bigr)$ or, equivalently,

$$\begin{bmatrix} u_1 & u_2 & v_1 & v_2 \end{bmatrix} \longmapsto \begin{bmatrix} u_1 & u_2 & v_1 & v_2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix},$$

where $u_1, u_2, v_1, v_2 \in \mathbb{F}_2$. The kernel K combines two TECs U and V to synthesize $U \boxast (V\omega)$ and $U \circledast (V\omega)$, where $V\omega$ is the channel that multiplies the input by ω before feeding it into V. For brevity, $W \boxast (W\omega)$ and $W \circledast (W\omega)$ are denoted by $W^\square$ and $W^\bigcirc$, respectively.

Multiplying a TEC by ω behaves like a rotation of order 3 (after all, $\omega^3 = 1$ and it is rotating the tetrahedron). It maps $\mathrm{TEC}(p, q, r, s, t)$ to $\mathrm{TEC}(p, s, q, r, t)$. If W is balanced, rotation does not alter it: $W = W\omega$. If it is not balanced, then the rotation helps mis-match q, r, s so that a large probability is paired with a small probability. More precisely,

$$\mathrm{TEC}(p,q,r,s,t)^\square \coloneqq \mathrm{TEC}(\quad p^2, \quad ps+sq+qp, \quad pq+qr+rp, \quad pr+rs+sp, \quad 1-\text{the other four} \quad),$$

$$\mathrm{TEC}(p,q,r,s,t)^\bigcirc \coloneqq \mathrm{TEC}(\quad 1-\text{the other four}, \quad ts+sq+qt, \quad tq+qr+rt, \quad tr+rs+st, \quad t^2 \quad).$$

Twisting makes it easier to reduce q, r, and s and redistribute the mass to p and t.

4.4 Channel process

For a TEC W, we call $W^\square$ the *serial-child* of W and $W^\bigcirc$ the *parallel-child* of W. Together, they are the *children* of W. The *descendants* of W are the children of W together with the descendants of the children of W. The *nth-generation* descendants of W are the $(n-1)$th-generation descendants of the children of W; the 0th is W itself.

When W is understood from the context, let W_0 be W. For n a positive integer, let W_n be a random child of W_{n-1} with equal probability.

The common strategy used to estimate the scaling exponent concerns a concave function $\psi\colon [0,1] \to \mathbb{R}$ such that $\psi(0) = \psi(1) = 0$ and is positive elsewhere. With ψ, one finds a $0 < \mu < \infty$ such that

$$\frac{\psi(H(W^\square)) + \psi(H(W^\bigcirc))}{2\psi(H(W))} \leqslant 2^{-1/\mu}$$

With this "eigenvalue," a routine argument [Wan21, Sections 5.8–5.10] will show that

$$\mathbb{P}\bigl\{ H(W_n) < \exp\bigl(-e^{n^{1/3}}\bigr) \bigr\} > 1 - H(W_0) - 2^{-n/\mu}.$$

5 Unbalanced TEC Becomes Balanced

In this section, we argue that TECs undergoing the polarization process tend to become more balanced than before. We do so by showing that the moments of inertia are decreasing.

Theorem 4 (uniform loss of inertia)**.** *$A(W^\square), A(W^\bigcirc) \leqslant A(W)(1 - A(W)/3)$ for any TEC W.*

A proof of the theorem is in Appendix A.1. Now the recurrence relation $A(W_{n+1}) \leqslant A(W_n) \times (1 - A(W_n)/3)$ is equivalent to $A(W_{n+1}) - A(W_n) \leqslant -A(W_n)^2/3$ and analogous to the ordinary differential equation $f'(n) \leqslant -f(n)^2/3$. Solving it, we get $f(n) = O(1/n)$; analogously, $A(W_n) = O(1/n)$.

Corollary 5 (ultimate loss of inertia)**.** *Fix a TEC W, then $A(W_n) = O(1/n)$ as $n \to \infty$.*

Another way to look at it is the average decay of $A(W_n)$.

Proposition 6 (average loss of inertia)**.** *$A(W^\square) + A(W^\bigcirc) \leqslant A(W)$ for any TEC W.*

A proof of the proposition is in Appendix A.2. By the proposition, $A(W) \geqslant A(W^\square)+A(W^\bigcirc) \geqslant A(W^{\square\square}) + A(W^{\square\bigcirc}) + A(W^{\bigcirc\square}) + A(W^{\bigcirc\bigcirc}) \geqslant A(W^{\square\square\square}) \cdots$. Hence the expectation of $A(W_n)$ over all W_n is at most $A(W)/2^n$.

Corollary 5 and Proposition 6 imply that any unbalanced TEC will promptly become very similar to a balanced one.[3] The speed of polarization of unbalanced TECs is thus dominated by that of balanced TECs. We now turn to the analysis of the polarization speed of balanced TECs.

6 Balanced TECs Hoard Edge Mass

In this section, we argue that the Quetelet index $Q(W_n) := E(W_n)/H(W_n)(1 - H(W_n))$ of a sufficiently deep descendant is about 1.6. Put another way, there is a "trap" that constrains the relation between $E(W_n)$ and $H(W_n)$.

Recall that a balanced TEC W is edge-heavy if $Q(W) \geqslant 2\sqrt{7} - 4$. Let $\alpha := 2\sqrt{7} - 4 \approx 1.3$.

Theorem 7 (trapping region)**.** *If W is balanced and edge-heavy, then its children are edge-heavy.*

A proof of the theorem is in Appendix B.1. The theorem implies that all descendants of an edge-heavy are edge-heavy. For a TEC that is not edge-heavy, its descendants will still become "edge-heavier" by the following theorem.

Theorem 8 (attraction toward the trap)**.** *Fix any $\varepsilon > 0$; choose $\delta := 3\varepsilon/8$. Let W be any balanced TEC. We have that $Q(W) \leqslant \alpha - \varepsilon$ implies $Q(W^\square) \geqslant Q(W)\bigl(1 + H(W)\delta\bigr)$ and $Q(W^\bigcirc) \geqslant Q(W)\bigl(1 + (1 - H(W))\delta\bigr)$.*

A proof of the theorem is in Appendix B.2. It is clear that the factors $H(W)$ and $1-H(W)$ before δ slow down the rate at which $Q(W_n)$ approaches $2\sqrt{7} - 4$, especially when $H(W)$ is close to 0 or 1, respectively. These factors cannot be optimized away. To see why, suppose that $H(W) = x \approx 1$ and $E(W) = y \approx 0$. Then $H(W^\bigcirc)$ is about $x^2 + O(y^2)$ and $E(W^\bigcirc)$ is about $2xy + O(y^2)$. Hence $Q(W^\bigcirc)$ is about $2xy/x^2(1 - x^2) \approx y/x(1 - x) = Q(W)$. That being the case, we would like to add that TECs whose Quetelet index can hardly be improved are already polarized, so we shall not worry about them. Besides, we can prove uniform attraction using Theorem 8.

Theorem 9 (uniform attraction)**.** *Fix any $\varepsilon > 0$. For any balanced TEC W such that $Q(W) \leqslant \alpha - \varepsilon$, there exists an integer $m > 0$ such that $Q(W_n) \geqslant Q(W)(1 + \varepsilon/8)$ for all $n \geqslant m$.*

[3] Note that Corollary 5 is a weak statement about every single descendant of W, while Proposition 6 implies a strong statement about $A(W_n)$ averaged over all nth-generation descendants. Only Corollary 5 will be used later.

A proof of the theorem is in Appendix B.3. Uniform attraction means that every child is at least making some positive progress toward the trap. Small steps of the descendants accumulate to a giant leap of the family.

Corollary 10 (ultimate attraction)**.** *For any $\varepsilon > 0$ and any balanced TEC W such that $Q(W) > 0$, there exists an integer $m > 0$ such that $Q(W_n) \geqslant \alpha - \varepsilon$ for all $n \geqslant m$.*

Proof. Apply the uniform attraction theorem repeatedly. Every application improves the Quetelet index by a factor of $1 + (\alpha - Q(W_n))/8$. So after a finite number of applications the Quetelet index can be made $\geqslant \alpha - \varepsilon$. □

To summarize this and the previous section, we have two trends: unbalanced TECs tend to become balanced; and "edge-light" TECs tend to become edge-heavy.

The following proposition is a bound on Quetelet index in the opposite direction.

Proposition 11 (attraction on the other side)**.** *Let W be a balanced TEC with $Q(W) \leqslant 2$. Then $Q(W^{\square}) \leqslant 2$ and $Q(W^{\bigcirc}) \leqslant 2$.*

Some comments on how to prove this proposition is in Appendix B.4.

The following proposition gives a tighter trapping region than Theorem 8 and Proposition 11 do. A proof is omitted but similar to those of Theorem 8 and Proposition 11. For the optimal trapping region, see the discussion in Appendix D.

Proposition 12. *Let $f(x) := x(1-x)(1.66 - 0.38x(1-x))$. Then $E(W) \leqslant f(H(W))$ implies $E(W^{\square}) \leqslant f(H(W^{\square}))$ and $E(W^{\bigcirc}) \leqslant f(H(W^{\bigcirc}))$. Let $g(x) := x(1-x)(2 - 2x(1-x)/3)$. Then $E(W) \geqslant g(H(W))$ implies $E(W^{\square}) \geqslant g(H(W^{\square}))$ and $E(W^{\bigcirc}) \geqslant g(H(W^{\bigcirc}))$.*

7 Edge-heavy TECs Polarize Faster

Let W be any balanced TEC with a fixed $H(W) = x$ and a variable $E(W) = y$. Then $H(W^{\square}) = 2x - x^2 + y^2/12$ is increasing in y and $H(W^{\bigcirc}) = x^2 - y^2/12$ is decreasing in y.

The monotonicity has two applications. Application one: If we know too little to lower bound $Q(W)$, we will upper bound $H(W^{\bigcirc})$ using x^2. In this case, the speed of polarization is at least $\mu \approx 3.627$, the number induced by the standard polar code. Application two: If we know $Q(W) \geqslant \alpha$, we will upper bound $H(W^{\bigcirc})$ using $x^2 - (\alpha x(1-x))^2/12$. This time, $H(W^{\bigcirc})$ and $H(W^{\square})$ are more separated so the speed of polarization is strictly better than $\mu \approx 3.627$. Any positive α, not necessarily $2\sqrt{7} - 4$, can improve the scaling. This is demonstrated by the following lemma that uses $9/7$ in place of α.

Lemma 13 (eigenfunction and eigenvalue)**.** *Let $\psi(x) := (x(1-x))^{0.697}(5 - \sqrt{x(1-x)})$. For balanced TECs with $Q(W) \geqslant 9/7$,*

$$\frac{\psi(H(W^{\square})) + \psi(H(W^{\bigcirc}))}{2\psi(H(W))} < 0.818.$$

Comments on how to verify the lemma is in Appendix C.

Theorem 14 (main theorem)**.** *Consider a pair of BECs treated as a TEC, or consider any TEC where $pqrst > 0$. The 2×2 matrix $\left[\begin{smallmatrix}1 & 0\\ \omega & 1\end{smallmatrix}\right]$ over $\mathbb{F}_4$ induces a scaling exponent less than 3.451.*

Proof. Two iid copies of BEC(ε) can be seen as $W \coloneqq \text{TEC}(\,(1-\varepsilon)^2,\,(1-\varepsilon)\varepsilon,\,0,\varepsilon(1-\varepsilon),\,\varepsilon^2\,)$. If ε is 0 or 1, there is nothing to prove. Suppose $0 < \varepsilon < 1$, then both $W^\square$ and $W^\bigcirc$ have five positive subspace erasure probabilities. (That is, their "p, q, r, s, t" are all positive). The descendants of a TEC with five positive subspace erasure probabilities satisfy the same property. In particular, all descendants have positive Quetelet index.

Let W be a TEC whose descendants all have positive Q. By Corollary 5, it takes W a finite number of generations to become very similar to a balanced TEC. That is, for any $\delta > 0$ there exists an $m > 0$ such that $A(W_m) < \delta$. Although W_n is never balanced, what we proved about balanced TECs still hold for "almost-balanced" TECs up to a diminishing error term. So we may proceed as if W_n is balanced for $n \geqslant m$.

By Corollary 10, it takes another finite number of generations to become "almost edge-heavy." In particular, there exists an m' such that $Q(W_{m'}) \geqslant 9/7$ (note that $9/7 \approx 1.286$ and $2\sqrt{7} - 4 \approx 1.291$).

Before the m'th generation, the eigenvalues of the form

$$\frac{\psi(H(W_n^\square)) + \psi(H(W_n^\bigcirc))}{2\psi(H(W_n))}$$

was less than 1. After the m'th generation, the eigenvalues of said form will be less than $0.818 < 2^{-1/3.451}$, by Lemma 13. As n goes to infinity, 0.818 dominates the overall scaling behavior. Hence W, and hence any BEC, enjoys scaling exponent less than 3.451. $\square$

In the abstract, we claim that the scaling exponent of $\left[\begin{smallmatrix}1 & 0\\ \omega & 1\end{smallmatrix}\right]$ over TECs (and hence BECs) is < 3.328. This number will be derived in Appendix D with more intense numerical calculations. In particular, there is a new trapping region that is bounded by two linear splines and is significantly smaller than the region bounded by $ax(1-x)$ for $a = 2\sqrt{7} - 4$ and 2; the attraction toward the new trap is witnessed by sampling TECs with low edge-mass. In Appendix E, we also examine the actual values of $H(W_n)$ and its asymptotic behavior aligns with the estimate 3.328.

8 Conclusions

In this paper, we argue that $\left[\begin{smallmatrix}1 & 0\\ \omega & 1\end{smallmatrix}\right]$ polarizes BECs faster than $\left[\begin{smallmatrix}1 & 0\\ 1 & 1\end{smallmatrix}\right]$ does. We first show that a pair of BECs will be transformed into balanced TECs. We then show that balanced TECs will be transformed into edge-heavy TECs. Finally, we show that edge-heavy TECs assume a better scaling exponent.

Our rigorous overestimate of the scaling exponent is 3.451; there is another overestimate of 3.328 with strong numerical evidence. Compared to Arıkan's 2×2 matrix with $\mu \approx 3.627$, Fazeli–Vardy's 8×8 matrix with $\mu \approx 3.577$ [FV14], Trofimiuk–Trifonov's 16×16 matrix with $\mu \approx 3.346$ [TT21], and Yao–Fazeli–Vardy's 32×32 matrix with $\mu \approx 3.122$ [YFV19], our result suggests that one should consider expanding the alphabet size prior to enlarging the matrix size. More precisely, the rigorous estimate is analogous to a 15×15 binary matrix; the more accurate estimate is analogous to a 20×20 binary matrix (see Figure 2).

9 Acknowledgment

This work was supported by NSF grants CCF-1764104, CCF-2107346, and CCF-2210823 and a Simons Investigator Award.

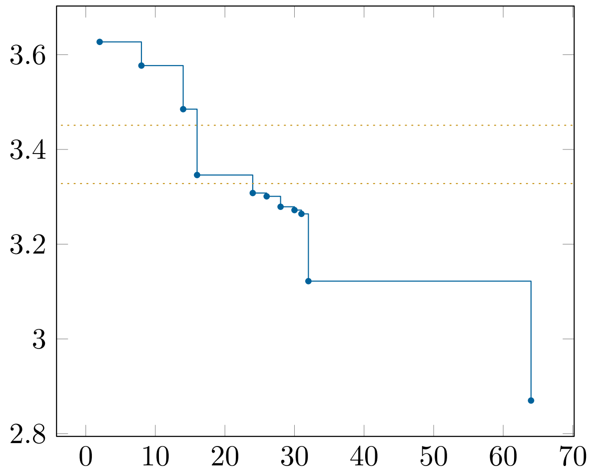

Figure 2: Horizontal axis: matrix size; vertical axis: scaling exponent of the best known matrix [FV14; YFV19; TT21; Tro21; BBL20]. A matrix size will be skipped if no known matrix outruns all smaller matrices. Underlying channel is BEC. Our estimates 3.451 and 3.328 are marked as dotted lines.

A Proofs for Balancing Channels

The following fact is useful for this appendix and follows from the definition. For any TEC $W = \text{TEC}(p, q, r, s, t)$,

$$\begin{aligned} A(W) &= (q-r)^2 + (r-s)^2 + (s-q)^2, \\ A(W^{\square}) &= (q-r)^2(s+p)^2 + (r-s)^2(q+p)^2 + (s-q)^2(r+p)^2, \\ A(W^{\bigcirc}) &= (q-r)^2(s+t)^2 + (r-s)^2(q+t)^2 + (s-q)^2(r+t)^2. \end{aligned}$$

A.1 Uniform loss of inertia (Theorem 4)

Now we want to prove $A(W^{\bigcirc}) \leqslant A(W)(1 - A(W)/3)$ for any TEC W, as $A(W^{\square}) \leqslant A(W)(1 - A(W)/3)$ will follow by duality.

Proof. We have $(s+t)^2 = (1-p-q-r)^2 \leqslant (1-q-r)^2 \leqslant (1-q+r)(1+q-r) = 1-(q-r)^2$. So we see that the three terms that sum to $A(W^{\bigcirc})$ can be bounded by

$$\begin{aligned} (q-r)^2(s+t)^2 &\leqslant (q-r)^2(1-(q-r)^2), \\ (r-s)^2(q+t)^2 &\leqslant (r-s)^2(1-(r-s)^2), \\ (s-q)^2(r+t)^2 &\leqslant (s-q)^2(1-(s-q)^2). \end{aligned}$$

The average of these three terms is at most $(A(W)/3)(1-A(W)/3)$ because $A(W)/3$ is the average of $(q-r)^2$, $(r-s)^2$, and $(s-q)^2$ and because $x(1-x)$ is concave in x. We then conclude that $A(W^{\bigcirc}) \leqslant A(W)(1 - A(W)/3)$. The upper bound on $A(W^{\square})$ follows by duality. □

A.2 Average loss of inertia (Proposition 6)

We want to prove $A(W^{\square}) + A(W^{\bigcirc}) \leqslant A(W)$.

Proof. First, note that $(q-r)^2(r-s)(s-q)+(r-s)^2(s-q)(q-r)+(s-q)^2(q-r)(r-s)$ is zero by factoring out $(q-r)(r-s)(s-q)$. Next, we want to prove that the following is nonnegative:

$$\begin{aligned} A(W)-A(W^{\square})-A(W^{\bigcirc}) = &\ (q-r)^2\big(1-(s+p)^2-(s+t)^2\big) \\ &+(r-s)^2\big(1-(q+p)^2-(q+t)^2\big) \\ &+(s-q)^2\big(1-(r+p)^2-(r+t)^2\big) \\ = &\ (q-r)^2\big(1-(s+p)^2-(s+t)^2-(r-s)(s-q)\big) \\ &+(r-s)^2\big(1-(q+p)^2-(q+t)^2-(s-q)(q-r)\big) \\ &+(s-q)^2\big(1-(r+p)^2-(r+t)^2-(q-r)(r-s)\big) \end{aligned}$$

It remains to show that $1-(s+p)^2-(s+t)^2-(r-s)(s-q)$ is nonnegative, as the rest will follow by symmetry. To show so, replace 1 with $(p+q+r+s+t)^2$. One sees that $(p+q+r+s+t)^2-(s+p)^2-(s+t)^2-(r-s)(s-q)=(2p+q+r+s+2t)(q+r)+qr+2pt$ is nonnegative. $\square$

B Proofs for Trapping Channels

The following fact is useful for this appendix and follows from the definition. Let W be a balanced TEC with $H(W)=x$ and $E(W)=y$. Then

$$H(W^{\bigcirc})=x^2-\frac{y^2}{12},\quad E(W^{\bigcirc})=2xy-\frac{2y^2}{3},\quad H(W^{\square})=2x-x^2+\frac{y^2}{12},\quad E(W^{\square})=2y-2xy-\frac{2y^2}{3}.$$

B.1 Trapping region (Theorem 7)

Given $Q(W)\geqslant\alpha\coloneqq 2\sqrt{7}-4$, now we want to prove $Q(W^{\bigcirc})\geqslant\alpha$ as $Q(W^{\square})\geqslant\alpha$ will follow by duality.

Proof. Consider a balanced TEC with entropy $H(W)=x$ and edge mass $E(W)=y$. We first prove that $Q(W^{\bigcirc})\geqslant\alpha$ whenever $y=\alpha x(1-x)$. To do so, we want $E(W^{\bigcirc})\geqslant\alpha H(W^{\bigcirc})(1-H(W^{\bigcirc}))$; so we want the following quantity to be nonnegative:

$$\begin{aligned} E(W^{\bigcirc})-\alpha H(W^{\bigcirc})(1-H(W^{\bigcirc})) &= 2xy-\frac{2y^2}{3}-\alpha\Big(x^2-\frac{y^2}{12}\Big)\Big(1-x^2+\frac{y^2}{12}\Big) \\ &= \alpha x^2(1-x)^2\cdot\frac{\alpha^2x(2-x)(24-\alpha^2x^2)+\alpha^4x^2-12\alpha^2-96\alpha+144}{144} \\ &= \alpha x^2(1-x)^2\cdot\frac{\alpha^2x(2-x)(24-\alpha^2x^2)+\alpha^4x^2}{144}. \end{aligned}$$

The subformula $-12\alpha^2-96\alpha+144$ is removed from the numerator because α is a root of it. The new numerator is clearly nonnegative because $0\leqslant x\leqslant 1$ and $\alpha<2$. This affirms that $E(W^{\bigcirc})\geqslant\alpha H(W^{\bigcirc})(1-H(W^{\bigcirc}))$ given that $y=\alpha x(1-x)$.

Now that we finished proving $Q(W^{\bigcirc})\geqslant\alpha$ when $y=\alpha x(1-x)$, it remains to consider the case when $y>\alpha x(1-x)$. For that, consider the map

$$\pi(x,y)\coloneqq\Big(x^2-\frac{y^2}{12},\,2xy-\frac{2y^2}{3}\Big).$$

We want to study the image of the hyperbola $x^2-y^2/12=c$, where c is a constant. Firstly, notice that the hyperbola intersects the parabola $y=\alpha x(1-x)$ at exactly one point. Secondly, notice that the x-coordinate of the image of the hyperbola is fixed (and is c). Thirdly, notice that the total derivative of $x^2 - y^2/12$ should be zero, i.e., $2x\partial x - y\partial y/6 = 0$ and consequently $\partial x = y\partial y/12x$.

We next focus on the y-coordinate of the image. It is $2xy - 2y^2/3$ and its total derivative is

$$2x\partial y + 2y\partial x - \frac{4y\partial y}{3} = 2x\partial y + \frac{2y^2\partial y}{12x} - \frac{4y\partial y}{3} = \frac{(6x-y)(2x-y)}{6x} \cdot \partial y.$$

This implies that when the point (x, y) is moving upward along the hyperbola $x^2 - y^2/12 = c$, the image $\pi(x, y)$ is also moving upward. Therefore, when (x, y) is moving above and away from the parabola $y = \alpha x(1-x)$, the image $\pi(x, y)$ will move above and away from the same parabola. This finishes the proof of that when $y \geqslant \alpha x(1-x)$, the image $\pi(x, y)$ lies above the concerned parabola. □

B.2 Attraction toward the trap (Theorem 8)

We want to prove that $Q(W) \leqslant \alpha - \varepsilon$ implies $Q(W^{\bigcirc}) \geqslant Q(W)\big(1 + (1 - H(W))\delta\big)$. And then $Q(W^{\square}) \geqslant Q(W)\big(1 + H(W)\delta\big)$ will follow by duality.

Proof. Let W be a balanced TEC with entropy $H(W) = x$, edge mass $E(W) = y$, and Quetelet index $Q(W) = y/x(1-x) = b$. Suppose $b < \alpha - \varepsilon$, then

$$\begin{aligned} E(W^{\bigcirc}) - bH(W^{\bigcirc})(1 - H(W^{\bigcirc})) &= bx^2(1-x)^2 \cdot \frac{b^2x(2-x)(24-b^2x^2) + b^4x^2 - 12b^2 - 96b + 144}{144} \\ &\geqslant bx^2(1-x)^2 \cdot \frac{b^2x(2-x)(24-b^2x^2) + b^4x^2 + 108\varepsilon}{144} \geqslant bx^2(1-x)^2 \cdot \frac{3\varepsilon}{4}. \end{aligned}$$

Here, $-12b^2 - 96b + 144 = -12(b-\alpha)(b - (-2\sqrt{7} - 4)) > 12\varepsilon(2\sqrt{7} + 4) > 108\varepsilon$. Therefore

$$Q(W^{\bigcirc}) = \frac{E(W^{\bigcirc})}{H(W^{\bigcirc})(1-H(W^{\bigcirc}))} \geqslant b + \frac{3bx^2(1-x)^2\varepsilon}{4H(W^{\bigcirc})(1-H(W^{\bigcirc}))} \geqslant b + \frac{3bx^2(1-x)^2\varepsilon}{4x^2(1-x^2)} \geqslant b + \frac{3b(1-x)\varepsilon}{8}.$$

So $3\varepsilon/8$, our choice of δ, is valid. □

B.3 Uniform attraction (Theorem 9)

We want to prove that if $Q(W) \leqslant \alpha - \varepsilon$, there exists an integer $m > 0$ such that $Q(W_n) \geqslant Q(W)(1 + \varepsilon/8)$ for all $n \geqslant m$.

Proof. In this proof, we call a descendant W' of W *good* if $Q(W') \geqslant Q(W)(1 + \varepsilon/8)$. Being good is hereditary: if a balanced TEC is good, its descendants are all good because their Q's are non-decreasing (Theorem 8). Now imagine the family tree consisting of the root W and all the descendants that are not good. The goal of this theorem is to show that this tree is finite.

Fix a balanced TEC W, we know either $H(W) \geqslant 1/3$ or $H(W) \leqslant 2/3$. For the former case,

$$Q(W^{\square}) \geqslant Q(W)\big(1 + H(W)\delta\big) \geqslant Q(W)\Big(1 + \frac{1}{3} \cdot \frac{3\varepsilon}{8}\Big) \geqslant Q(W)\Big(1 + \frac{\varepsilon}{8}\Big).$$

For the latter case,

$$Q(W^{\bigcirc}) \geqslant Q(W)\big(1 + (1 - H(W))\delta\big) \geqslant Q(W)\Big(1 + \frac{1}{3} \cdot \frac{3\varepsilon}{8}\Big) \geqslant Q(W)\Big(1 + \frac{\varepsilon}{8}\Big).$$

We infer that $H(W) \geqslant 1/3$ implies $W^{\square}$ good and $H(W) \leqslant 2/3$ implies $W^{\bigcirc}$ good. We see that the family tree of the bad descendants is uniparous—every node in this tree has at most one child.

At this point, the only concern is whether there exists an infinite path of TECs $W, W_1, W_2, W_3, \ldots$ such that each is a child of the previous TEC, and none of them has entropy lying in $[1/3, 2/3]$. To see why this cannot happen, suppose we begin with $x \coloneqq H(W) > 2/3$. We know $W^{\square}$ is definitely good so W_1 must be $W^{\bigcirc}$. Given that $H(W^{\bigcirc}) = x^2 - y^2/12$ and $0 \leqslant y \leqslant 2 - 2x$, we see

$$H(W_1) = x^2 - \frac{y^2}{12} \geqslant x^2 - \frac{(1-x)^2}{3} = \frac{11}{27} \geqslant \frac{1}{3}.$$

This implies that the "gap" $[1/3, 2/3]$ is too large and that the path of TECs $W, W_1, W_2, \ldots$ cannot cross this gap—it must stay within $(2/3, 1)$ if $H(W)$ began there.

According to the last few paragraphs, what will contradict the theorem is a path of TECs $W, W_1, \ldots$ such that each is the parallel-child of the previous and all of them have entropy $> 2/3$. But this cannot happen because the H of the parallel-child is at most the square of the previous H, and squaring a number in $(2/3, 1)$ will eventually make it less than $2/3$. By duality, there cannot be a path of TECs such that each is the serial-child of the previous and all of them have entropy $< 1/3$. This finishes the proof. □

B.4 Attraction on the other side (Proposition 11)

To prove that $Q(W) \leqslant 2$ implies $Q(W^{\bigcirc}) \leqslant 2$ and $Q(W^{\square}) \leqslant 2$, apply the same strategy as Theorem 7 (Appendix B.1): expand $E(W^{\bigcirc}) - 2H(W^{\bigcirc})(1 - H(W^{\bigcirc}))$ (using a computer algebra system) and show that it is a negative polynomial in x. To show a polynomial negative, one may apply ultra-high–precision numerical approximations to "get a feeling." One then applies Sturm's theorem to do the final check.

C Eigenvalue and Eigenvector (Lemma 13)

Lemma 13 has $\psi(x) \coloneqq (x(1-x))^{0.697}(5 - \sqrt{x(1-x)})$ and W balanced and $Q(W) \geqslant 9/7$. We want to verify that

$$\frac{\psi(H(W^{\square})) + \psi(H(W^{\bigcirc}))}{2\psi(H(W))} < 0.818.$$

Let $H(W) = x$ and $E(W) = y = 9x(1-x)/7$. Then $H(W^{\bigcirc}) = x^2 - y^2/12 = (169x^2 + 54x^3 - 27x^4)/196$, while $H(W^{\square}) = 2x - x^2 + y^2/12 = (392x - 169x^2 - 54x^3 + 27x^4)/196$. The statement we want to prove boils down to

$$\frac{\psi\left(\frac{169x^2+54x^3-27x^4}{196}\right) + \psi\left(\frac{392x-169x^2-54x^3+27x^4}{196}\right)}{2\psi(x)} < 0.818 \tag{2}$$

for $0 < x < 1$. One can verify this numerically.

If $y > 9x(1-x)/7$, then $H(W^{\bigcirc})$ and $H(W^{\square})$ are more separated than when $y = 9x(1-x)/7$. Hence $\psi(H(W^{\square})) + \psi(H(W^{\bigcirc}))$ is smaller than the numerator of formula (2) as ψ is convex. Hence the quotient is still smaller than 0.818.

C.1 Suboptimality

The eigenvalue 0.818 is not optimal in two aspects. For one, the eigenfunction ψ is not optimal. A common practice is to run power iteration

$$\psi_0(x) \coloneqq (x(1-x))^{0.7},$$
$$\psi_{k+1}(x) \coloneqq \frac{\psi_k\big(\frac{169x^2+54x^3-27x^4}{196}\big) + \psi_k\big(\frac{392x-169x^2-54x^3+27x^4}{196}\big)}{2\max\psi_k}$$

until ψ_k converges (note: use spline). Then the limit of ψ_k will induce a smaller eigenvalue. See [WLVG22, Section III.C] for more details.

On the other, the trapping region we used is $Q(W) \geqslant 9/7$. A better value is $2\sqrt{7}-4$. But even $Q(W) \geqslant 2\sqrt{7}-4$ is not optimal. A smaller trapping region is $y \geqslant x(1-x)(1.66-0.38x(1-x))$. But even that is not optimal. So in the next appendix, we will use power iteration to find the optimal trapping region.

D Numerical Trapping Region

In this appendix, we want to find the optimal (smallest) trapping region. In this appendix, W is a balanced TEC and $x = H(W)$ and $y = E(W)$. Recall that x and y determine W uniquely; define

$$h_\bigcirc(x,y) \coloneqq H(W^\bigcirc) = x^2 - \frac{y^2}{12}, \qquad e_\bigcirc(x,y) \coloneqq E(W^\bigcirc) = 2xy - \frac{2y^2}{3},$$
$$h_\square(x,y) \coloneqq H(W^\square) = 2x - x^2 + \frac{y^2}{12}, \qquad e_\square(x,y) \coloneqq E(W^\square) = 2y - 2xy - \frac{2y^2}{3}.$$

We call the lower boundary of a trapping region the *inner bound* and the upper boundary of a trapping region the *outer bound*. For instance, $y = (2\sqrt{7}-4)x(1-x)$ and $y = x(1-x)(1.66-0.38x \times (1-x))$ are inner bounds; $y = 2x(1-x)$ and $y = x(1-x)(2-2x(1-x)/3)$ are outer bounds. Paraphrased, the mission of this appendix is to find the optimal inner and outer bounds.

D.1 Inner bound

Suppose $y = \varphi(x)$ is an inner bound, i.e., φ is such that the children of a TEC above the curve $y = \varphi(x)$ will still be above the same curve. Then the definition translates into:

$$e_\bigcirc(x,\varphi(x)) \geqslant \varphi\big(h_\bigcirc(x,\varphi(x))\big), \qquad e_\square(x,\varphi(x)) \geqslant \varphi\big(h_\square(x,\varphi(x))\big).$$

In other words,

$$\varphi(x) \leqslant e_\bigcirc\big(h_\bigcirc^{-1}(x),\varphi(h_\bigcirc^{-1}(x))\big), \qquad \varphi(x) \leqslant e_\square\big(h_\square^{-1}(x),\varphi(h_\square^{-1}(x))\big).$$

where $h_\bigcirc^{-1}$ and $h_\square^{-1}$ are inverse functions of $x \mapsto h_\bigcirc(x,\varphi(x))$ and $x \mapsto h_\square(x,\varphi(x))$, respectively.

Suppose there exists an optimal inner bound and it is of the form $y = \varphi(x)$, i.e., φ is the greatest function such that the children of a TEC above the curve $y = \varphi(x)$ will still be above the same curve. Then the optimality of φ translates into

$$\varphi(x) = \min\Big(\quad e_\bigcirc\big(h_\bigcirc^{-1}(x),\varphi(h_\bigcirc^{-1}(x))\big) \quad , \quad e_\square\big(h_\square^{-1}(x),\varphi(h_\square^{-1}(x))\big) \quad \Big).$$

We do not know a priori if the optimal inner bound exists. But we can consider the following inductive definition

$$\begin{aligned}
\varphi_0(x) &\coloneqq 2x(1-x), \\
h_{\bigcirc k}(x) &\coloneqq h_\bigcirc(x, \varphi_k(x)), \\
h_{\Box k}(x) &\coloneqq h_\Box(x, \varphi_k(x)), \\
e_{\bigcirc k}(x) &\coloneqq e_\bigcirc\big(h_{\bigcirc k}^{-1}(x), \varphi_k(h_{\bigcirc k}^{-1}(x))\big), \\
e_{\Box k}(x) &\coloneqq e_\Box\big(h_{\Box k}^{-1}(x), \varphi_k(h_{\Box k}^{-1}(x))\big), \\
\varphi_{k+1}(x) &\coloneqq \min(e_{\bigcirc k}(x), e_{\Box k}(x)).
\end{aligned}$$

These functions can be and were implemented by linear splines with 10^5 nodes. The advantage of linear splines is that the inverse function of a linear spline is still a linear spline. Per our computation, φ_k converges as $k \to \infty$. So we suspect that the limit of φ_k is the optimal inner bound.

A plot of the limit of φ_k is in Figure 3.

D.2 Outer bound

An argument similar to the previous sub-appendix applies to outer bound. Suppose $y = \chi(x)$ is an outer bound, i.e., χ is such that the children of a TEC below the curve $y = \chi(x)$ will still be below the curve. The definition translates into

$$e_\bigcirc(x, \chi(x)) \leqslant \chi\big(h_\bigcirc(x, \chi(x))\big), \qquad e_\Box(x, \chi(x)) \leqslant \chi\big(h_\Box(x, \chi(x))\big).$$

In other words,

$$\chi(x) \geqslant e_\bigcirc\big(h_\bigcirc^{-1}(x), \chi(h_\bigcirc^{-1}(x))\big), \qquad \chi(x) \geqslant e_\Box\big(h_\Box^{-1}(x), \chi(h_\Box^{-1}(x))\big).$$

Suppose $y = \chi(x)$ is the optimal outer bound, then the optimality implies

$$\chi(x) = \max\Big(\; e_\bigcirc\big(h_\bigcirc^{-1}(x), \chi(h_\bigcirc^{-1}(x))\big) \;,\; e_\Box\big(h_\Box^{-1}(x), \chi(h_\Box^{-1}(x))\big) \; \Big).$$

This means that we can setup an inductive definition almost identical to the one above except that the last line will be with max:

$$\begin{aligned}
\chi_0(x) &\coloneqq 2x(1-x), \\
h_{\bigcirc k}(x) &\coloneqq h_\bigcirc(x, \chi_k(x)), \\
h_{\Box k}(x) &\coloneqq h_\Box(x, \chi_k(x)), \\
e_{\bigcirc k}(x) &\coloneqq e_\bigcirc\big(h_{\bigcirc k}^{-1}(x), \chi_k(h_{\bigcirc k}^{-1}(x))\big), \\
e_{\Box k}(x) &\coloneqq e_\Box\big(h_{\Box k}^{-1}(x), \chi_k(h_{\Box k}^{-1}(x))\big), \\
\chi_{k+1}(x) &\coloneqq \max(e_{\bigcirc k}(x), e_{\Box k}(x)).
\end{aligned}$$

We did the computation and the end result of χ_k is plotted in Figure 3.

D.3 Scaling exponent

Using the numerical limit of $\varphi_k(x)$ in place of $(2\sqrt{7}-4)x(1-x)$, we can strengthen Lemma 13 and Theorem 14. Details omitted, our final number is $\mu < 3.328$.

Theorem 15 (enhanced main theorem)**.** *Consider a pair of BECs treated as a TEC, or consider any TEC where $pqrst > 0$. The 2×2 matrix $\left[\begin{smallmatrix}1&0\\ \omega&1\end{smallmatrix}\right]$ over $\mathbb{F}_4$ induces a scaling exponent less than 3.328.*

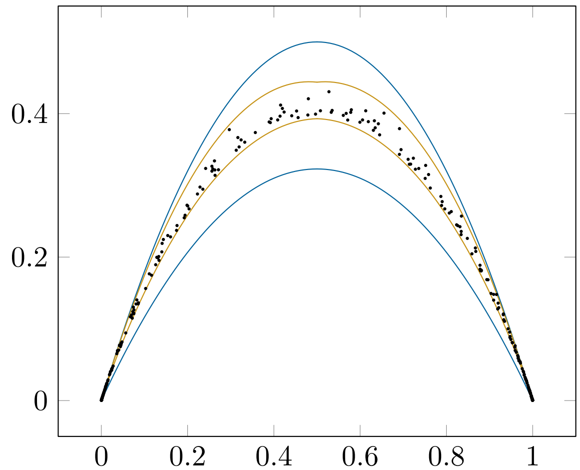

Figure 3: Horizontal axis: $x = H(W_n)$; vertical axis: $y = E(W_n)$. Curves from top to bottom: $2x(1-x)$, the numerical outer bound χ_{293}, the numerical inner bound φ_{120}, and $(2\sqrt{7}-4)x(1-x)$. Dots: the 10th-generation descendants of BEC(0.55).

E Simulations

In this appendix, we presents simulations in two figures. In Figure 3, we take a pair of BEC(0.55), treat them as W, compute $H(W_n)$ and $E(W_n)$ for $n = 10$, and compare it with trapping regions.

In Figure 4, we compute the expectation of $\psi(H(W_n))$ for W_n defined by $\left[\begin{smallmatrix}1&0\\1&1\end{smallmatrix}\right]$ and that defined by $\left[\begin{smallmatrix}1&0\\\omega&1\end{smallmatrix}\right]$. The result shows a clear advantage of $\left[\begin{smallmatrix}1&0\\\omega&1\end{smallmatrix}\right]$ over $\left[\begin{smallmatrix}1&0\\1&1\end{smallmatrix}\right]$.

References

[AMV22] Fariba Abbasi, Hessam Mahdavifar, and Emanuele Viterbo. "Hybrid Non-Binary Repeated Polar Codes". In: *IEEE Transactions on Wireless Communications* (2022), pp. 1–1. ISSN: 1558-2248. DOI: 10.1109/TWC.2022.3159807.

[Ari09] Erdal Arikan. "Channel Polarization: A Method for Constructing Capacity-Achieving Codes for Symmetric Binary-Input Memoryless Channels". In: *IEEE Transactions on Information Theory* 55.7 (July 2009), pp. 3051–3073. ISSN: 1557-9654. DOI: 10.1109/TIT.2009.2021379.

[AT09] Erdal Arikan and Emre Telatar. "On the rate of channel polarization". In: *2009 IEEE International Symposium on Information Theory*. June 2009, pp. 1493–1495. DOI: 10.1109/ISIT.2009.5205856.

[BBL20] Manan Bhandari, Ishan Bansal, and V. Lalitha. "On the Polarizing Behavior and Scaling Exponent of Polar Codes with Product Kernels". In: *2020 National Conference on Communications (NCC)*. Feb. 2020, pp. 1–6. DOI: 10.1109/NCC48643.2020.9056096.

[BFSTV17] Sarit Buzaglo, Arman Fazeli, Paul H. Siegel, Veeresh Taranalli, and Alexander Vardy. "Permuted successive cancellation decoding for polar codes". In: *2017 IEEE International Symposium on Information Theory (ISIT)*. June 2017, pp. 2618–2622. DOI: 10.1109/ISIT.2017.8007003.

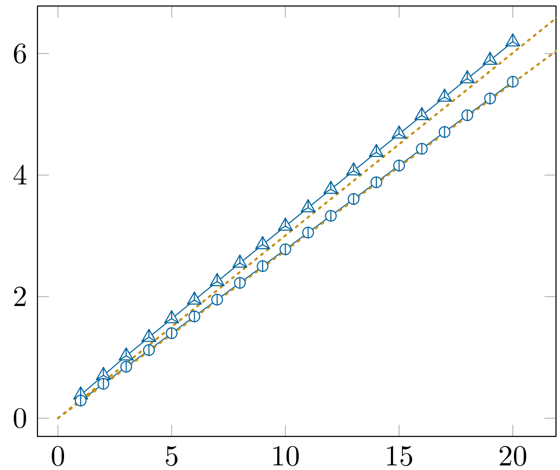

Figure 4: Horizontal axis: generation (that is, n); vertical axis: $-\log_2$ of the expectation of $\psi(H(W_n))/\psi(H(W_0))$, where $\psi(x) \coloneqq (x(1-x))^{0.7}$. Triangle marks: BEC(0.55) polarized by $\left[\begin{smallmatrix}1&0\\\omega&1\end{smallmatrix}\right]$; circle marks: BEC(0.55) polarized by $\left[\begin{smallmatrix}1&0\\1&1\end{smallmatrix}\right]$. Dotted lines: the lines of slopes 1/3.328 (above) and 1/3.627 (below). This reaffirms that $\mu \approx 3.627$ is accurate for $\left[\begin{smallmatrix}1&0\\1&1\end{smallmatrix}\right]$ and $\mu < 3.328$ is an overestimate for $\left[\begin{smallmatrix}1&0\\\omega&1\end{smallmatrix}\right]$.

[BGLB20] Valerio Bioglio, Frédéric Gabry, Ingmar Land, and Jean-Claude Belfiore. "Multi-Kernel Polar Codes: Concept and Design Principles". In: *IEEE Transactions on Communications* 68.9 (Sept. 2020), pp. 5350–5362. ISSN: 1558-0857. DOI: 10.1109/TCOMM.2020.3006212.

[BGNRS22] Jarosław Błasiok, Venkatesan Guruswami, Preetum Nakkiran, Atri Rudra, and Madhu Sudan. "General Strong Polarization". In: *J. ACM* 69.2 (Mar. 2022). ISSN: 0004-5411. DOI: 10.1145/3491390. URL: https://doi.org/10.1145/3491390.

[BGS18] Jaroslaw Blasiok, Venkatesan Guruswami, and Madhu Sudan. "Polar Codes with Exponentially Small Error at Finite Block Length". In: *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2018)*. Ed. by Eric Blais, Klaus Jansen, José D. P. Rolim, and David Steurer. Vol. 116. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2018, 34:1–34:17. ISBN: 978-3-95977-085-9. DOI: 10.4230/LIPIcs.APPROX-RANDOM.2018.34. URL: http://drops.dagstuhl.de/opus/volltexte/2018/9438.

[BKB04] D. Baron, M.A. Khojastepour, and R.G. Baraniuk. "How quickly can we approach channel capacity?" In: *Conference Record of the Thirty-Eighth Asilomar Conference on Signals, Systems and Computers, 2004.* Vol. 1. Nov. 2004, 1096–1100 Vol.1. DOI: 10.1109/ACSSC.2004.1399310.

[CBM18] Peiyao Chen, Baoming Bai, and Xiao Ma. *Two-Stage Polarization-Based Nonbinary Polar Codes for 5G URLLC*. 2018. DOI: 10.48550/ARXIV.1801.08059. URL: https://arxiv.org/abs/1801.08059.

[Chi14] Mao-Ching Chiu. "Non-binary polar codes with channel symbol permutations". In: *2014 International Symposium on Information Theory and its Applications*. Oct. 2014, pp. 433–437.

[CKW19] Semih Cayci, Toshiaki Koike-Akino, and Ye Wang. “Nonbinary Polar Coding for Multi-level Modulation”. In: *2019 Optical Fiber Communications Conference and Exhibition (OFC)*. 2019, pp. 1–3.

[CLMS21] Eduardo Camps, Hiram H. López, Gretchen L. Matthews, and Eliseo Sarmiento. “Polar Decreasing Monomial-Cartesian Codes”. In: *IEEE Transactions on Information Theory* 67.6 (June 2021), pp. 3664–3674. ISSN: 1557-9654. DOI: 10.1109/TIT.2020.3047624.

[FHMV21] Arman Fazeli, Hamed Hassani, Marco Mondelli, and Alexander Vardy. “Binary Linear Codes With Optimal Scaling: Polar Codes With Large Kernels”. In: *IEEE Transactions on Information Theory* 67.9 (Sept. 2021), pp. 5693–5710. ISSN: 1557-9654. DOI: 10.1109/TIT.2020.3038806.

[FV14] Arman Fazeli and Alexander Vardy. “On the scaling exponent of binary polarization kernels”. In: *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. Sept. 2014, pp. 797–804. DOI: 10.1109/ALLERTON.2014.7028536.

[GB14] Dina Goldin and David Burshtein. “Improved Bounds on the Finite Length Scaling of Polar Codes”. In: *IEEE Transactions on Information Theory* 60.11 (Nov. 2014), pp. 6966–6978. ISSN: 1557-9654. DOI: 10.1109/TIT.2014.2359197.

[GHU12] Ali Goli, S. Hamed Hassani, and Rüdiger Urbanke. “Universal bounds on the scaling behavior of polar codes”. In: *2012 IEEE International Symposium on Information Theory Proceedings*. July 2012, pp. 1957–1961. DOI: 10.1109/ISIT.2012.6283641.

[GRY22] Venkatesan Guruswami, Andrii Riazanov, and Min Ye. “Arıkan meets Shannon: Polar codes with near-optimal convergence to channel capacity”. In: *IEEE Transactions on Information Theory* (2022), pp. 1–1. ISSN: 1557-9654. DOI: 10.1109/TIT.2022.3146786.

[GX15] Venkatesan Guruswami and Patrick Xia. “Polar Codes: Speed of Polarization and Polynomial Gap to Capacity”. In: *IEEE Transactions on Information Theory* 61.1 (Jan. 2015), pp. 3–16. ISSN: 1557-9654. DOI: 10.1109/TIT.2014.2371819.

[HAU10] S. Hamed Hassani, Kasra Alishahi, and Rudiger Urbanke. “On the scaling of polar codes: II. The behavior of un-polarized channels”. In: *2010 IEEE International Symposium on Information Theory*. June 2010, pp. 879–883. DOI: 10.1109/ISIT.2010.5513585.

[HAU14] Seyed Hamed Hassani, Kasra Alishahi, and Rüdiger L. Urbanke. “Finite-Length Scaling for Polar Codes”. In: *IEEE Transactions on Information Theory* 60.10 (Oct. 2014), pp. 5875–5898. ISSN: 1557-9654. DOI: 10.1109/TIT.2014.2341919.

[Hay09] Masahito Hayashi. “Information Spectrum Approach to Second-Order Coding Rate in Channel Coding”. In: *IEEE Transactions on Information Theory* 55.11 (Nov. 2009), pp. 4947–4966. ISSN: 1557-9654. DOI: 10.1109/TIT.2009.2030478.

[HMTU13] S. Hamed Hassani, Ryuhei Mori, Toshiyuki Tanaka, and Rüdiger L. Urbanke. “Rate-Dependent Analysis of the Asymptotic Behavior of Channel Polarization”. In: *IEEE Transactions on Information Theory* 59.4 (Apr. 2013), pp. 2267–2276. ISSN: 1557-9654. DOI: 10.1109/TIT.2012.2228295.

[KMTU10] Satish Babu Korada, Andrea Montanari, Emre Telatar, and Rüdiger Urbanke. “An empirical scaling law for polar codes”. In: *2010 IEEE International Symposium on Information Theory*. June 2010, pp. 884–888. DOI: 10.1109/ISIT.2010.5513579.

[KŞU10] Satish Babu Korada, Eren Şaşoğlu, and Rüdiger Urbanke. "Polar Codes: Characterization of Exponent, Bounds, and Constructions". In: *IEEE Transactions on Information Theory* 56.12 (Dec. 2010), pp. 6253–6264. ISSN: 1557-9654. DOI: 10.1109/TIT.2010.2080990.

[LH06] Ingmar Land and Johannes Huber. "Information Combining". In: *Foundations and Trends® in Communications and Information Theory* 3.3 (2006), pp. 227–330. ISSN: 1567-2190. DOI: 10.1561/0100000013. URL: http://dx.doi.org/10.1561/0100000013.

[Lin21] Liping Lin. "On the Construction of the Kernel Matrix by Primitive BCH Codes for Polar Codes". In: *Communications and Network* 14.1 (2021), pp. 23–35.

[LY21] Guan-Chen Liu and Qi-Yue Yu. *Non-Binary Polar Coded System for the Two-User Multiple-Access Channel*. 2021. DOI: 10.48550/ARXIV.2111.03839. URL: https://arxiv.org/abs/2111.03839.

[MHU15] Marco Mondelli, S. Hamed Hassani, and Rüdiger L. Urbanke. "Scaling Exponent of List Decoders With Applications to Polar Codes". In: *IEEE Transactions on Information Theory* 61.9 (Sept. 2015), pp. 4838–4851. ISSN: 1557-9654. DOI: 10.1109/TIT.2015.2453315.

[MHU16] Marco Mondelli, S. Hamed Hassani, and Rüdiger L. Urbanke. "Unified Scaling of Polar Codes: Error Exponent, Scaling Exponent, Moderate Deviations, and Error Floors". In: *IEEE Transactions on Information Theory* 62.12 (Dec. 2016), pp. 6698–6712. ISSN: 1557-9654. DOI: 10.1109/TIT.2016.2616117.

[MT10] Ryuhei Mori and Toshiyuki Tanaka. "Non-binary polar codes using Reed-Solomon codes and algebraic geometry codes". In: *2010 IEEE Information Theory Workshop*. Aug. 2010, pp. 1–5. DOI: 10.1109/CIG.2010.5592755.

[MT14] Ryuhei Mori and Toshiyuki Tanaka. "Source and Channel Polarization Over Finite Fields and Reed–Solomon Matrices". In: *IEEE Transactions on Information Theory* 60.5 (May 2014), pp. 2720–2736. ISSN: 1557-9654. DOI: 10.1109/TIT.2014.2312181.

[Nas16] Rajai Nasser. "An Ergodic Theory of Binary Operations—Part I: Key Properties". In: *IEEE Transactions on Information Theory* 62.12 (Dec. 2016), pp. 6931–6952. ISSN: 1557-9654. DOI: 10.1109/TIT.2016.2616642.

[PB13] Woomyoung Park and Alexander Barg. "Polar Codes for Q-Ary Channels, $q = 2^r$". In: *IEEE Transactions on Information Theory* 59.2 (2013), pp. 955–969. DOI: 10.1109/TIT.2012.2219035.

[PPV10] Yury Polyanskiy, H. Vincent Poor, and Sergio Verdu. "Channel Coding Rate in the Finite Blocklength Regime". In: *IEEE Transactions on Information Theory* 56.5 (May 2010), pp. 2307–2359. ISSN: 1557-9654. DOI: 10.1109/TIT.2010.2043769.

[PSL16] Noam Presman, Ofer Shapira, and Simon Litsyn. "Mixed-Kernels Constructions of Polar Codes". In: *IEEE Journal on Selected Areas in Communications* 34.2 (Feb. 2016), pp. 239–253. ISSN: 1558-0008. DOI: 10.1109/JSAC.2015.2504278.

[PU16] Henry D. Pfister and Rüdiger Urbanke. "Near-optimal finite-length scaling for polar codes over large alphabets". In: *2016 IEEE International Symposium on Information Theory (ISIT)*. July 2016, pp. 215–219. DOI: 10.1109/ISIT.2016.7541292.

[RWLP22] Constantin Runge, Thomas Wiegart, Diego Lentner, and Tobias Prinz. *Multilevel Binary Polar-Coded Modulation Achieving the Capacity of Asymmetric Channels*. 2022. arXiv: 2202.04010 [cs.IT].

[Şaş12] Eren Şaşoğlu. "Polar codes for discrete alphabets". In: *2012 IEEE International Symposium on Information Theory Proceedings*. July 2012, pp. 2137–2141. DOI: 10.1109/ISIT.2012.6283740.

[Sav21] Valentin Savin. "Non-Binary Polar Codes for Spread-Spectrum Modulations". In: *2021 11th International Symposium on Topics in Coding (ISTC)*. Aug. 2021, pp. 1–5. DOI: 10.1109/ISTC49272.2021.9594166.

[SP11] Aria G. Sahebi and S. Sandeep Pradhan. "Multilevel polarization of polar codes over arbitrary discrete memoryless channels". In: *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. Sept. 2011, pp. 1718–1725. DOI: 10.1109/Allerton.2011.6120375.

[SSSH13] Mathis Seidl, Andreas Schenk, Clemens Stierstorfer, and Johannes B. Huber. "Polar-Coded Modulation". In: *IEEE Transactions on Communications* 61.10 (Oct. 2013), pp. 4108–4119. ISSN: 1558-0857. DOI: 10.1109/TCOMM.2013.090513.130433.

[ŞTA09] Eren Şaşoğlu, Emre Telatar, and Erdal Arikan. "Polarization for arbitrary discrete memoryless channels". In: *2009 IEEE Information Theory Workshop*. Oct. 2009, pp. 144–148. DOI: 10.1109/ITW.2009.5351487.

[Tro21] Grigorii Trofimiuk. "Shortened Polarization Kernels". In: *2021 IEEE Globecom Workshops (GC Wkshps)*. Dec. 2021, pp. 1–6. DOI: 10.1109/GCWkshps52748.2021.9681982.

[TT21] Grigorii Trofimiuk and Peter Trifonov. "Window Processing of Binary Polarization Kernels". In: *IEEE Transactions on Communications* 69.7 (July 2021), pp. 4294–4305. ISSN: 1558-0857. DOI: 10.1109/TCOMM.2021.3072730.

[Wan21] Hsin-Po Wang. *Complexity and Second Moment of the Mathematical Theory of Communication*. 2021. arXiv: 2107.06420 [cs.IT].

[WD21] Hsin-Po Wang and Iwan M. Duursma. "Polar Codes' Simplicity, Random Codes' Durability". In: *IEEE Transactions on Information Theory* 67.3 (2021), pp. 1478–1508.

[WLVG22] Hsin-Po Wang, Ting-Chun Lin, Alexander Vardy, and Ryan Gabrys. *Sub-4.7 Scaling Exponent of Polar Codes*. 2022. DOI: 10.48550/ARXIV.2204.11683. URL: https://arxiv.org/abs/2204.11683.

[YFV19] Hanwen Yao, Arman Fazeli, and Alexander Vardy. "Explicit Polar Codes with Small Scaling Exponent". In: *2019 IEEE International Symposium on Information Theory (ISIT)*. 2019, pp. 1757–1761. DOI: 10.1109/ISIT.2019.8849741.

[YS18] Peihong Yuan and Fabian Steiner. "Construction and Decoding Algorithms for Polar Codes based on 2×2 Non-Binary Kernels". In: *2018 IEEE 10th International Symposium on Turbo Codes Iterative Information Processing (ISTC)*. Dec. 2018, pp. 1–5. DOI: 10.1109/ISTC.2018.8625284.