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

Arman Fazeli
UC San Diego
afazelic@ucsd.edu

Hamed Hassani
University of Pennsylvania
hassani@seas.upenn.edu

Marco Mondelli
Stanford University
mondelli@stanford.edu

Alexander Vardy
UC San Diego
avardy@ucsd.edu

***Abstract*—We prove that, at least for the binary erasure channel, the polar-coding paradigm gives rise to codes that not only approach the Shannon limit but, in fact, 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. Specifically, for any fixed $\delta > 0$, we exhibit binary linear codes that ensure reliable communication at rates within $\varepsilon > 0$ of capacity with block length $n = O(1/\varepsilon^{2+\delta})$, construction complexity $\Theta(n)$, and encoding/decoding complexity $\Theta(n \log n)$.**

***Index Terms*—Non-asymptotic information theory, finite blocklength, polar codes, large kernels, scaling exponent.**

I. 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 $P_{\rm e} > 0$, there exists a binary code of rate at least $I(W) - \varepsilon$ which enables communication over W with probability of error at most $P_{\rm e}$. Ever since the publication of Shannon's famous paper [28], 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 $P_{\rm e} > 0$ 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.

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 $\Theta(n)$, and encoding/decoding complexity $\Theta(n \log n)$.

To establish this result, we use *polar coding*, invented by Arıkan [2] 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.

A. Background and context

A sequence of papers, starting with [6,29] in 1960s and culminating with [16,26], shows that for any discrete memoryless channel W and *any* code of length n and rate R that achieves error-probability $P_{\rm e}$ on W, we have

$$I(W) - R \ \geqslant\ \frac{\text{const}(P_{\rm e}, W)}{\sqrt{n}} \ -\ O\left(\frac{\log n}{n}\right), \qquad (1)$$

where the constant (which is given explicitly in [26]) depends on W and $P_{\rm e}$, but not on n. This immediately

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. The research of H. Hassani was supported by NSF-CRII award CCF-1755707. Marco Mondelli was supported in part by an Early Postdoc Mobility Fellowship from the Swiss National Science Foundation.

implies that if $n = O\left(1/\varepsilon^{\mu}\right)$, where $\varepsilon = I(W) - R$ is the gap to capacity, then $\mu \geqslant 2$.

It is well known [16,26] that the lower bound $\mu = 2$ is achieved by random linear codes. Unfortunately, random linear codes cannot be decoded efficiently. On general BMS channels, this task is NP-hard [4]. 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 [9] 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] for more details), so they are definitely not efficient. 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 [2] that runs in time $O(n \log n)$. It was shown in [11] 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. A sequence of papers [10,13,17,24] have 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 $3.579 \leqslant \mu \leqslant 4.714$.

One possible approach is to improve the successive-cancellation decoding algorithm. In particular, the successive cancellation *list decoder* proposed in [31] empirically provides a significant improvement in performance. However, [23] 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. 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 [18] 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 [8] 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 [8]. Pfister and Urbanke [25] 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 [12] 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.

Spatially-coupled LDPC codes: Achieve the capacity of any BMS channel under a belief-propagation decoding algorithm [20] that runs in linear time. A simple heuristic argument yields that the scaling exponent of these codes is roughly 3 (see [21, 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 [19] that runs in time $O(n^{\omega})$. While there has been empirical and analytical evidence that the performance of Reed-Muller codes on the BEC is close to that of random codes [14,22], no bounds on the scaling exponent of RM codes are known.

B. 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.

Theorem 1: Consider transmission over a binary erasure channel W with capacity $I(W)$. Fix $P_{\mathrm{e}} \in (0,1)$ and an arbitrary $\delta > 0$. Then, for all $R < I(W)$, there exists a sequence of binary linear codes of rate R that guarantee error probability at most P_{e} on the channel W, and whose block length n satisfies

$$n \leqslant \frac{\beta}{\big(I(W) - R\big)^{\mu}} \qquad \text{with} \quad \mu \leqslant 2 + \delta, \tag{2}$$

where $\beta = \big(1 + 2P_{\mathrm{e}}^{-0.01}\big)^3$ is a universal constant. Moreover, the codes in this sequence have construction complexity $\Theta(n)$ and encoding/decoding complexity $\Theta(n \log n)$.

A couple of 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 -0.01. We point out that we could have similarly chosen any other negative constant as the exponent of P_{e}. 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.

To prove Theorem 1, we will show that there exist $\ell\times\ell$ binary kernels, with quasi-linear encoding and decoding complexity, 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 [2,27,30], we refer to [7] for specific details.

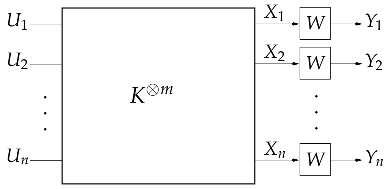
C. A primer on polar codes

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 [2], correspond to

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

However, it was shown in [18] 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 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 below:



(4)

To understand what polarization means in this context, we consider a number of channels associated with this transformation (see also Chapter 5 of [27] and Chapter 2.4 of [12]). 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 $i\in[n]$, 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

$$\begin{aligned} W_i\big(\boldsymbol{y},\boldsymbol{v}|\,u_i\big) &\overset{\text{def}}{=} \frac{1}{2^{n-1}} \sum_{\boldsymbol{u}'\in\{0,1\}^{n-i}} W^*\Big(\boldsymbol{y}\,\Big|\,(\boldsymbol{v},u_i,\boldsymbol{u}')\Big) \\ &= \frac{1}{2^{n-1}} \sum_{\boldsymbol{u}'\in\{0,1\}^{n-i}} W^n\Big(\boldsymbol{y}\,\Big|\,(\boldsymbol{v},u_i,\boldsymbol{u}')K^{\otimes m}\Big), \end{aligned} \tag{5}$$

where $(\cdot,\cdot)$ stands for concatenation. We say that W_i is the *i-th bit-channel*. It is easy to see that $W_i\big(\boldsymbol{y},\boldsymbol{v}|\,u_i\big)$ 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 [2] is that, as n grows, the n bit-channels W_i defined in (5) 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

$$\begin{aligned} I(W) &\overset{\text{def}}{=} \frac{1}{2}\sum_{y\in\mathscr{Y}}\sum_{x\in\{0,1\}} W(y|x)\log_2\frac{W(y|x)}{\frac{1}{2}W(y|0)+\frac{1}{2}W(y|1)}, \\ Z(W) &\overset{\text{def}}{=} \sum_{y\in\mathscr{Y}}\sqrt{W(y|0)W(y|1)}. \end{aligned} \tag{6}$$

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 [2, Theorem 1] can be informally stated as follows. 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 $1-I(W)$. An (n,k) polar code is constructed by selecting a set $\mathcal{A}$ of k good bit-channels to carry the information bits, while the input to all the other bit-channels is frozen to zeros.

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, [12, Section 3.4] and [8]) 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. 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 (5) with $n=\ell$ and $W^*(\boldsymbol{y}|\boldsymbol{u}) = W^\ell\big(\boldsymbol{y}\,\big|\,\boldsymbol{u}K\big)$. We refer to the set of ℓ polynomials $\big\{f_{K,1}(z), f_{K,2}(z), \ldots, f_{K,\ell}(z)\big\}$ 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. 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:

$$Z_{m+1} \,=\, f_{K,B_m}(Z_m), \qquad \text{for } B_m \sim \mathsf{Uniform}[\ell], \quad (7)$$

with the initial condition $Z_0 = z$. One can view (7) as a stochastic process on an infinite binary tree, where in each step we take one of the ℓ available branches with uniform probability. The polarization theorem then follows from the martingale convergence theorem, which in this case implies that

$$\lim_{m\to\infty} Z_m(1-Z_m) \,=\, 0.$$

This shows that the erasure probabilities $p_i(z)$ of the ℓ^m bit-channels polarize to either 0 or 1 as $m \to \infty$. Furthermore, since the matrix $K^{\otimes m}$ is nonsingular, it is easy to see that the polar transform in (4) preserves capacity. 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.

II. 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 (7) polarizes. In other words, given a small $\varepsilon > 0$, how fast does the quantity $\mathbb{P}\{Z_m \in [\varepsilon, 1-\varepsilon]\}$ 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 $Z_m \in [\varepsilon, 1-\varepsilon]$ if and only if Y_m is lower-bounded by $\varepsilon^\alpha(1-\varepsilon)^\alpha$. Therefore, by Markov inequality, we have

$$\mathbb{P}\big\{Z_m \in [\varepsilon, 1-\varepsilon]\big\} \;\leqslant\; \frac{\mathbb{E}[g_\alpha(Z_m)]}{\varepsilon^\alpha(1-\varepsilon)^\alpha} \qquad (8)$$

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

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

Proceeding along these lines, we eventually conclude that

$$\mathbb{E}[g_\alpha(Z_m)] \;\leqslant\; \big(\lambda^*_{\alpha,K}\big)^m, \qquad (9)$$

where

$$\lambda^*_{\alpha,K} \;\triangleq\; \sup_{z\in(0,1)} \frac{1}{\ell} \, \frac{\displaystyle\sum_{i=1}^{\ell} \Big(f_{K,i}(z)\big(1 - f_{K,i}(z)\big)\Big)^\alpha}{\big(z(1-z)\big)^\alpha}. \qquad (10)$$

Step 2: Sharp transitions in the polarization behavior. We show that with probability at least $1 - O(1/\ell)$ over the random choice of a nonsingular $\ell \times \ell$ binary kernel K, we have

$$\lambda^*_{\alpha,K} \;\leqslant\; \ell^{-1/2+5\alpha} \,. \qquad (11)$$

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

$$\begin{aligned} &f_{K,i}(z) \leqslant \ell^{-(2+\log\ell)}, && \text{for } z \leqslant \frac{i}{\ell} - 5\ell^{-1/2}\log\ell, \\ &f_{K,i}(z) \geqslant 1 - \ell^{-(2+\log\ell)}, && \text{for } z \geqslant \frac{i}{\ell} + 5\ell^{-1/2}\log\ell, \end{aligned} \qquad (12)$$

with probability at least $1 - O(1/\ell)$ over the random choice of K.

Let us now go back to (10) and try to use this "sharpness" property of the polarization behavior in order to upper-bound $\lambda^*_{\alpha,K}$. In fact, let us only evaluate the term on the RHS of (10) at the single point $z = 1/2$, rather than taking the supremum over all $z \in (0,1)$. Using the "sharpness" property in (12), it is not difficult to see that for $z = 1/2$, this term will be of order

$$\ell^{-1/2}\log\ell + \ell^{-\alpha(2+\log\ell)} \;\leqslant\; \ell^{-1/2\,+\,5\alpha}, \qquad (13)$$

for all sufficiently large ℓ. With some more effort, we will establish in [7] that, in fact, the upper bound in (13) is valid for all $z \in (0,1)$ rather than at the single point $z = 1/2$.

Step 3: Finite-length scaling law. We can derive the finite-length scaling law for polar codes using the results of the previous two steps. From (8), (9), and (11), we conclude that

$$\mathbb{P}\big\{Z_m \in [\varepsilon, 1-\varepsilon]\big\} \;=\; O\Big(\varepsilon^{-\alpha}\big(\ell^{-1/2+5\alpha}\big)^m\Big). \qquad (14)$$

Denote the desired error probability by P_{e}, and set $\varepsilon = P_{\mathrm{e}}/n = P_{\mathrm{e}}\ell^{-m}$ in (14). Then we have

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

where δ can be made arbitrarily small by choosing a small enough α (and sufficiently large ℓ). 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 $\mathbb{P}\{Z_m \geqslant 1 - P_{\mathrm{e}}\ell^{-m}\}$ 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.

References

[1] A. Amraoui, A. Montanari, T. Richardson, and R. Urbanke, Finite-length scaling for iteratively decoded LDPC ensembles. *IEEE Trans. Inform. Theory*, **55** (February 2009), 473–498.

[2] E. Arikan, Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels. *IEEE Trans. Inform. Theory 55*, **7** (July 2009).

[3] E. Arikan and I.E. Telatar, On the rate of channel polarization. In *Proc. of the IEEE Int. Symposium on Inform. Theory* (Seoul, South Korea, July 2009), pp. 1493–1495.

[4] E.R. Berlekamp, R.J. McEliece, and H.C.A. van Tilborg, On the inherent intractability of certain coding problems, *IEEE Trans. Inform. Theory*, **24** (May 1978), 384–386.

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

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

[7] A. Fazeli, S. H. Hassani, M. Mondelli, and A. Vardy, Binary linear codes with optimal scaling and quasi-linear complexity. Available: `http://arxiv.org/abs/1711.01339`.

[8] 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, October 2014), pp. 797–804.

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

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

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

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

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

[14] H. Hassani, S. Kudekar, Y. Polyanskiy, O. Ordentlich, and R. Urbanke, Almost Optimal Scaling of Reed-Muller Codes on BEC and BSC Channels, in *Proc. IEEE Intern. Symp. Information Theory*, Vail, CO, June 2018.

[15] S.H. Hassani, R. Mori, T. Tanaka, and R.L. Urbanke, Rate-dependent analysis of the asymptotic behavior of channel polarization. *IEEE Trans. Inform. Theory 59*, 4 (Apr. 2013), 2267–2276.

[16] M. Hayashi, Information spectrum approach to second-order coding rate in channel coding. *IEEE Trans. Inform. Theory 55*, 11 (Nov. 2009), 4947–4966.

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

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

[19] S. Kudekar, S. Kumar, M. Mondelli, H. D. Pfister, E. Şaşoğlu, and R. Urbanke, R, Reed-Muller codes achieve capacity on erasure channels, *IEEE Trans. Inform. Theory 63*, 7 (July 2017), 4298–4316.

[20] S. Kudekar, T. J. Richardson, and R. L. Urbanke, Spatially coupled ensembles universally achieve capacity under belief propagation, *IEEE Trans. Inform. Theory 59*, 12 (Dec. 2013), 7761–7813.

[21] M. Mondelli, S.H. Hassani, and R.L. Urbanke, How to achieve the capacity of asymmetric channels, accepted to *IEEE Trans. Inform. Theory*, January 2018; also arxiv.org/abs/1406.7373.

[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. 62*, 9 (Sept. 2014).

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

[24] 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. Inform. Theory*, vol. 62, no. 12, pp. 6698–6712, December 2016

[25] H. D. Pfister, and R. Urbanke, Near-optimal finite-length scaling for polar codes over large alphabets May 2016. [Online]. Available: `http://arxiv.org/abs/1605.01997`.

[26] Y. Polyanskiy, H. V. Poor, and S. Verdú, Channel coding rate in the finite block-length regime, *IEEE Trans. Inform. Theory 56*, 5 (May 2010), 2307–2359.

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

[28] C.E. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.*, **27**, (April/October 1948), 379–423 and 623–656.

[29] V. Strassen, Asymptotische abschätzungen in Shannon's informationstheorie, In *Trans. 3rd Prague Conf. Inf. Theory* (1962).

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

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

[32] A. Vardy and Y. Be'ery, Maximum-likelihood soft decision decoding of BCH codes, *IEEE Transactions on Information Theory*, 1994 Mar,40(2), pp. 546–554.