General Strong Polarization

Jarosław Błasiok*
John A. Paulson School of Engineering and Applied Sciences, Harvard University, 33 Oxford Street, Cambridge, MA 02138, USA.
jblasiok@g.harvard.edu

Venkatesan Guruswami†
Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
venkatg@cs.cmu.edu

Preetum Nakkiran‡
John A. Paulson School of Engineering and Applied Sciences, Harvard University, 33 Oxford Street, Cambridge, MA 02138, USA.
preetum@cs.harvard.edu

Atri Rudra§
Computer Science and Engineering Department, University at Buffalo, SUNY, Buffalo, NY 14260, USA.
atri@buffalo.edu

Madhu Sudan¶
Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University, 33 Oxford Street, Cambridge, MA 02138, USA.
madhu@cs.harvard.edu

ABSTRACT

Arıkan's exciting discovery of polar codes has provided an altogether new way to efficiently achieve Shannon capacity. Given a (constant-sized) invertible matrix M, a family of polar codes can be associated with this matrix and its ability to approach capacity follows from the *polarization* of an associated $[0,1]$-bounded martingale, namely its convergence in the limit to either 0 or 1 with probability 1. Arıkan showed appropriate polarization of the martingale associated with the matrix $G_2 = \left(\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right)$ to get capacity achieving codes. His analysis was later extended to all matrices M which satisfy an obvious necessary condition for polarization.

While Arıkan's theorem does not guarantee that the codes achieve capacity at small blocklengths (specifically in length which is a polynomial in $1/\varepsilon$ where ε is the difference between the capacity of a channel and the rate of the code), it turns out that a "strong" analysis of the polarization of the underlying martingale would lead to such constructions. Indeed for the martingale associated with G_2 such a strong polarization was shown in two independent works ([Guruswami and Xia, IEEE IT '15] and [Hassani et al., IEEE IT'14]), thereby resolving a major theoretical challenge associated with the efficient attainment of Shannon capacity.

In this work we extend the result above to cover martingales associated with all matrices that satisfy the necessary condition for (weak) polarization. In addition to being vastly more general, our proofs of strong polarization are (in our view) also much simpler and modular. Key to our proof is a notion of *local polarization* that only depends on the evolution of the martingale in a single time step. We show that local polarization always implies strong polarization. We then apply relatively simple reasoning about conditional entropies to prove local polarization in very general settings. Specifically, our result shows strong polarization over all prime fields and leads to efficient capacity-achieving source codes for compressing arbitrary i.i.d. sources, and capacity-achieving channel codes for arbitrary symmetric memoryless channels.

CCS CONCEPTS

• **Mathematics of computing → Coding theory**;

KEYWORDS

Polar codes

ACM Reference Format:
Jarosław Błasiok, Venkatesan Guruswami, Preetum Nakkiran, Atri Rudra, and Madhu Sudan. 2018. General Strong Polarization. In *Proceedings of 50th Annual ACM SIGACT Symposium on the Theory of Computing (STOC'18).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3188745.3188816

1 INTRODUCTION

Polar codes, proposed in Arıkan's remarkable work [2], gave a fresh information-theoretic approach to construct linear codes that achieve the Shannon capacity of symmetric channels, together with efficient encoding and decoding algorithms. About a decade after their discovery, there is now a vast and extensive body of work on polar coding spanning hundreds of papers, and polar codes are also being considered as one of the candidates for use in 5G wireless (e.g., see [6] and references therein). The underlying concept of polarizing transforms has emerged as a versatile tool to successfully attack a diverse collection of information-theoretic problems

*Supported by ONR grant N00014-15-1-2388.
†Research supported in part by NSF grant CCF-1422045.
‡Work supported in part by a Simons Investigator Award, NSF Awards CCF 1565641 and CCF 1715187, and the NSF Graduate Research Fellowship Grant No. DGE1144152.
§Research supported in part by NSF grants CCF-1717134 and CCF-1763481.
¶Work supported in part by a Simons Investigator Award and NSF Awards CCF 1565641 and CCF 1715187.

STOC'18, June 25–29, 2018, Los Angeles, CA, USA

ACM ISBN 978-1-4503-5559-9/18/06…$15.00
https://doi.org/10.1145/3188745.3188816

beyond the original channel and source coding applications, including wiretap channels [15], the Slepian-Wolf, Wyner-Ziv, and Gelfand-Pinsker problems [13], broadcast channels [8], multiple access channels [1, 20], and interference networks [22]. We recommend the survey by Şaşoğlu [19] for a nice treatment of the early work on polar codes.

The algorithmic interest in polar codes emerges from a consequence shown in the works [9, 10, 12] who show that this approach leads to a family of codes of rate $C - \varepsilon$ for transmission over a channel of (Shannon) capacity C, where the block length of the code and the decoding time grow only polynomially in $1/\varepsilon$. In contrast, for all previous constructions of codes, the decoding algorithms required time exponential in $1/\varepsilon$. Getting a polynomial running time in $1/\varepsilon$ was arguably one of the most important theoretical challenges in the field of algorithmic coding theory, and polar codes were the first to overcome this challenge. The analyses of polar codes turn into questions about *polarizations* of certain *martingales*. The vast class of polar codes alluded to in the previous paragraph all build on polarizing martingales, and the results of [9, 10, 12] show that for one of the families of polar codes, the underlying martingale polarizes "extremely fast" — a notion we refer to as *strong polarization* (which we will define shortly).

The primary goal of this work is to understand the process of polarization of martingales, and in particular to understand when does a martingale polarize strongly. In attempting to study this question, we come up with a local notion of polarization and show that this local notion is sufficient to imply strong polarization. Applying this improved understanding to the martingales arising in the study of polar codes we show that a simple necessary condition for weak polarization of such martingales is actually sufficient for strong polarization. This allows us to extend the results of [9, 10, 12] to a broad class of codes and show essentially that all polarizing codes lead to polynomial convergence to capacity. Below we formally describe the notion of polarization of martingales and our results.

1.1 Polarization of $[0,1]$-Martingales

Our interest is mainly in the (rate of) polarization of a specific family of martingales that we call the Arıkan martingales. We will define these objects later, but first describe the notion of polarization for general $[0,1]$-bounded martingales. Recall that a sequence of random variables $X_0, \ldots, X_t, \ldots$ is said to be a *martingale* if for every t and $a_0, \ldots, a_t$ it is the case that $\mathbb{E}[X_{t+1} | X_0 = a_0, \ldots, X_t = a_t] = a_t$. We say that that a martingale is $[0,1]$-*bounded* (or simply a $[0,1]$-martingale) if $X_t \in [0,1]$ for all $t \geq 0$.

Definition 1.1 (Weak Polarization). *A $[0,1]$-martingale sequence $X_0, X_1, \ldots, X_t, \ldots$ is defined to be* weakly polarizing *if $\lim_{t\to\infty}\{X_t\}$ exists with probability 1, and this limit is either 0 or 1 (and so the limit is a Bernoulli random variable with expectation X_0).*

Thus a polarizing martingale does not converge to a single value with probability 1, but rather converges to one of its extreme values. For the applications to constructions of polar codes, we need more explicit bounds on the rates of convergence leading to the notions of (regular) polarization and strong polarization defined below in Definition 1.3 and 1.4 respectively.

Definition 1.2 ((τ,ε)-Polarization). *For functions $\tau, \varepsilon : \mathbb{Z}^+ \to \mathbb{R}^{\geq 0}$, a $[0,1]$-martingale sequence $X_0, X_1, \ldots X_t, \ldots$ is defined to be (τ,ε)-*polarizing *if for all t we have*

$$\Pr(X_t \in (\tau(t), 1-\tau(t))) < \varepsilon(t).$$

Definition 1.3 (Regular Polarization). *A $[0,1]$-martingale sequence $X_0, X_1, \ldots, X_t, \ldots$ is defined to be* regular polarizing *if for all constant $\gamma > 0$, there exist $\varepsilon(t) = o(1)$, such that X_t is $(\gamma^t, \varepsilon(t))$-polarizing.*

We refer to the above as being "sub-exponentially" close to the limit (since it holds for every $\gamma > 0$). While weak polarization by itself is an interesting phenomenon, regular polarization (of Arıkan martingales) leads to capacity-achieving codes (though without explicit bounds on the length of the code as a function of the gap to capacity) and thus regular polarization is well-explored in the literature and tight necessary and sufficient conditions are known for regular polarization of Arıkan martingales [3, 14].

To get codes of block length polynomially small in the gap to capacity, an even stronger notion of polarization is needed, where we require that the sub-exponential closeness to the limit happens with *all but exponentially small probability*. We define this formally next.

Definition 1.4 (Strong Polarization). *A $[0,1]$-martingale sequence $X_0, X_1, \ldots, X_t, \ldots$ is defined to be* strongly polarizing *if for all $\gamma > 0$ there exist $\eta < 1$ and $\beta < \infty$ such that martingale X_t is $(\gamma^t, \beta \cdot \eta^t)$-polarizing.*

In contrast to the rich literature on regular polarization, results on strong polarization are quite rare, reflecting a general lack of understanding of this phenomenon. Indeed (roughly) an Arıkan martingale can be associated with every invertible matrix over any finite field $\mathbb{F}_q$, and the only matrix for which strong polarization is known is $G_2 = \left(\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right)$ [9, 10, 12].[1]

Part of the reason behind the lack of understanding of strong polarization is that polarization is a "limiting phenomenon" in that one tries to understand $\lim_{t\to\infty} X_t$, whereas most stochastic processes, and the Arıkan martingales in particular, are defined by local evolution, i.e., one that relates X_{t+1} to X_t. The main contribution of this work is to give a local definition of polarization (Definition 1.5) and then showing that this definition implies strong polarization (Theorem 1.6). Later we show that Arıkan martingales polarize locally whenever they satisfy a simple condition that is necessary even for weak polarization. As a consequence we get strong polarization for all Arıkan martingales for which previously only regular polarization was known.

1.2 Results I: Local Polarization and Implication

Before giving the definition of local polarization, we give some intuition using the following martingale: Let $Z_0 = 1/2$, and $Z_{t+1} =$

[1] An exception is the work by Pfister and Urbanke [18] who showed that for the *q-ary erasure channel* for large enough q, the martingale associated with a $q \times q$ Reed-Solomon based matrix proposed in [17] polarizes strongly. A recent (unpublished) work [7] shows that for the binary erasure channel, martingales associated with large *random* matrices polarize strongly. Both these results obtain an optimal value of η for (specific/random) large matrices. However, they only apply to the erasure channel, which is simple to error correct via Gaussian elimination and therefore not really reflective of the general capacity-achieving power of polar codes.

$Z_t + Y_{t+1}2^{-(t+2)}$ where $Y_1, \ldots, Y_t, \ldots$ are chosen uniformly and independently from $\{-1, +1\}$. Clearly this sequence is not polarizing (the limit of Z_t is uniform in $[0, 1]$). One reason why this happens is that as time progresses, the martingale slows down and stops varying much. We would like to prevent this, but this is also inevitable if a martingale is polarizing. In particular, a polarizing martingale would be slowed at the boundary and cannot vary much. The first condition in our definition of local polarization insists that this be the only reason a martingale slows down (we refer to this as *variance in the middle*).

Next we consider what happens when a martingale is close to the boundary. For this part consider a martingale $Z_0 = 1/2$ and $Z_{t+1} = Z_t + \frac{1}{2}Y_{t+1}\min\{Z_t, 1 - Z_t\}$. This martingale does polarize and even shows regular polarization, but it can also be easily seen that the probability that $Z_t \in [\frac{1}{2} \cdot 2^{-t}, 1 - \frac{1}{2} \cdot 2^{-t}]$ is one (whereas we would like probability of say $Z_t \in [10^{-t}, 1 - 10^{-t}]$ to go to 0). So this martingale definitely does not show strong polarization. This is so since even in the best case the martingale is approaching the boundary at a fixed exponential rate, and not a sub-exponential one. To overcome this obstacle we require that when the martingale is close to the boundary, with a fixed constant probability it should get much closer in a single step (a notion we refer to as *suction at the ends*).

The definition below makes the above requirements precise.

Definition 1.5 (Local Polarization)**.** *A $[0, 1]$-martingale sequence $X_0, \ldots, X_j, \ldots$, is* locally polarizing *if the following conditions hold:*

1. **(Variance in the middle):** *For every $\tau > 0$, there is a $\theta = \theta(\tau) > 0$ such that for all j, we have: If $X_j \in (\tau, 1 - \tau)$ then $\mathbb{E}[(X_{j+1} - X_j)^2 | X_j] \geq \theta$.*
2. **(Suction at the ends):** *There exists an $\alpha > 0$, such that for all $c < \infty$, there exists a $\tau = \tau(c) > 0$, such that:*
   (a) *If $X_j \leq \tau$ then $\Pr[X_{j+1} \leq X_j / c | X_j] \geq \alpha$.*
   (b) *Similarly, if $1 - X_j \leq \tau$ then $\Pr[(1 - X_{j+1} \leq (1 - X_j)/c | X_j] \geq \alpha$.*

   We refer to condition (a) above as Suction at the low end *and condition (b) as* Suction at the high end.

When we wish to be more explicit, we refer to the sequence as $(\alpha, \tau(\cdot), \theta(\cdot))$-locally polarizing.

As such this definition is neither obviously sufficient for strong polarization, nor is it obviously satisfiable by any interesting martingale. In the rest of the paper, we address these concerns. Our first technical contribution is a general theorem connecting local polarization to strong polarization.

Theorem 1.6 (Local vs. Strong Polarization). *If a $[0, 1]$-martingale sequence $X_0, \ldots, X_t, \ldots$, is locally polarizing, then it is also strongly polarizing.*

It remains to show that the notion of local polarization is not vacuous. Next, we show that in fact Arıkan martingales polarize locally (under simple necessary conditions). First we give some background on Polar codes.

1.3 The Arıkan Martingale and Polar Codes

The setting of polar codes considers an arbitrary *symmetric memoryless channel* and yields codes that aim to achieve the *capacity* of this channel. Given a finite field $\mathbb{F}_q$, and output alphabet $\mathcal{Y}$, recall that a q-ary channel $C_{Y|Z}$ is a probabilistic function from $\mathbb{F}_q$ to $\mathcal{Y}$ or equivalently it is given by q probability distributions $\{C_{Y|\alpha}\}_{\alpha \in \mathbb{F}_q}$ supported on $\mathcal{Y}$. A memoryless channel maps $\mathbb{F}_q^n$ to $\mathcal{Y}^n$ by acting independently (and identically) on each coordinate. A symmetric channel is a memoryless channel where for every $\alpha, \beta \in \mathbb{F}_q$ there is a bijection $\sigma : \mathcal{Y} \to \mathcal{Y}$ such that for every $y \in \mathcal{Y}$ it is the case that $C_{Y=y|\alpha} = C_{Y=\sigma(y)|\beta}$, and moreover for any pair $y_1, y_2 \in \mathcal{Y}$, we have $\sum_{x \in \mathbb{F}_q} C_{Y=y_1|x} = \sum_{x \in \mathbb{F}_q} C_{Y=y_2|x}$ (see, for example, [4, Section 7.2]). As shown by Shannon every memoryless channel has a finite capacity, denoted Capacity$(C_{Y|Z})$. For symmetric channels, this is the mutual information between the input Z and output Y where Z is drawn uniformly from $\mathbb{F}_q$ and Y is drawn from $C_{Y|Z}$ given Z.

Given any q-ary memoryless channel $C_{Y|Z}$ and invertible matrix $M \in \mathbb{F}_q^{k \times k}$, the theory of polar codes implicitly defines a martingale, which we call the Arıkan martingale associated with $(M, C_{Y|Z})$ and studies its polarization. (An additional contribution of this work is that we give an explicit compact definition of this martingale, see Definition 3.1. Since we do not need this definition for the purposes of this section, we defer it for Section 3). The consequences of regular polarization are described by the following remarkable theorem. (Below we use $M \otimes N$ to denote the tensor product (or the Kronecker product) of the matrix M and N. Further, we use $M^{\otimes t}$ to denote the tensor of a matrix M with itself t times.)

Theorem 1.7 (Implied by Arikan [2]). *Let C be a q-ary symmetric memoryless channel and let $M \in \mathbb{F}_q^{k \times k}$ be an invertible matrix. If the Arıkan martingale associated with (M, C) polarizes regularly, then given $\varepsilon > 0$ and $c < \infty$ there is a t_0 such that for every $t \geq t_0$ there is a code $C \subseteq \mathbb{F}_q^n$ for $n = k^t$ of dimension at least* $(\text{Capacity}(C) - \varepsilon) \cdot n$ *such that C is an affine code generated by the restriction of $(M^{-1})^{\otimes t}$ to a subset of its rows and an affine shift. Moreover there is a polynomial time decoding algorithm for these codes that has failure probability bounded by n^{-c}.*[2]

For $n = 2^t$, Arıkan and Telatar [3] proved that the martingale associated with the matrix $G_2 = \left(\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right)$, polarizes regularly over any binary input symmetric channel (Arıkan's original paper [2] proved a weaker form of regular polarization with $\tau(t) < 2^{-5t/4}$ which also sufficed for decoding error going to 0). Subsequent work generalized this to other matrices with the work of Korada, Şaşoğlu, and Urbanke [14] giving a precise characterization of matrices M for which the Arıkan martingale polarizes (again over binary input channels). We will refer to such matrices as *mixing*.

Definition 1.8 (Mixing Matrix)**.** *A matrix $M \in \mathbb{F}_q^{k \times k}$ is said to be* mixing*, if it is invertible and none of the permutations of the rows of M yields an upper triangular matrix, i.e., for every permutation $\pi : [k] \to [k]$ there exists $i, j \in [k]$ with $j < \pi(i)$ such that $M_{i,j} \neq 0$.*

[2] We remark that the encoding and decoding are not completely uniform as described above, since the subset of rows and the affine shift that are needed to specify the code are only guaranteed to exist. In the case of additive channels, where the shift can be assumed to be zero, the work of Tal and Vardy [21] (or [10, Sec. V]) removes this non-uniformity by giving a polynomial time algorithm to find the subset.

It is not too hard to show that the Arıkan martingale associated with non-mixing matrices do not polarize (even weakly). In contrast [14] shows that every mixing matrix over $\mathbb{F}_2$ polarizes regularly. Mori and Tanaka [17] show that the same result holds for all prime fields, and give a slightly more complicated criterion that characterizes (regular) polarization for general fields. (These works show that the decoding failure probability of the resulting polar codes is at most 2^{-n^β} for some positive β determined by the structure of the mixing matrix — this follows from an even stronger decay in the first of the two parameters in the definition of polarization. However, they do *not* show strong polarization which is what we achieve.)

As alluded to earlier, strong polarization leads to even more effective code constructions and this is captured by the following theorem.

Theorem 1.9 ([2, 10, 12]). *Let C be a q-ary symmetric memoryless channel and let $M \in \mathbb{F}_q^{k\times k}$ be an invertible matrix. If the Arıkan martingale associated with (M, C) polarizes strongly, then for every c there exists $t_0(x) = O(\log x)$ such that for every $\varepsilon > 0$ and every $t \geq t_0(1/\varepsilon)$ there is an affine code C, that is generated by the rows of $(M^{-1})^{(\otimes t)}$ and an affine shift, with the property that the rate of C is at least $\mathrm{Capacity}(C) - \varepsilon$, and C can be encoded and decoded in time $O(n \log n)$ where $n = k^t$ and failure probability of the decoder is at most n^{-c}.*

This theorem is implicit in the works above, but for completeness we include a proof of this theorem in the full version of this paper. As alluded to earlier, the only Arıkan martingales that were known to polarize strongly were those where the underlying matrix was $G_2 = \left(\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right)$. Specifically Guruswami and Xia [10] and Hassani et al. [12] show strong polarization of the Arıkan martingale associated with this matrix over any binary input symmetric channel, and Guruswami and Velingker [9] extended to the case of q-ary input channels for prime q. By using the concept of local polarization we are able to extend these results to all mixing matrices.

1.4 Results II: Local Polarization of Arıkan Martingales

In our second main result, we show that every mixing matrix gives rise to an Arıkan martingale that is locally polarizing:

Theorem 1.10. *For every prime q, for every mixing matrix $M \in \mathbb{F}_q^{k\times k}$, and for every symmetric channel $C_{Y|Z}$ over $\mathbb{F}_q$, the associated Arıkan martingale sequence is locally polarizing.*

As a consequence of Theorems 1.9, 1.6, and 1.10, we have the following theorem.

Theorem 1.11. *For every prime q, every mixing matrix $M \in \mathbb{F}_q^{k\times k}$, every symmetric channel C over $\mathbb{F}_q$, and every $c < \infty$, there exists $t_0(x) = O(\log x)$ such that for every $\varepsilon > 0$, for every $t \geq t_0(1/\varepsilon)$, there is an affine code C, that is generated by the rows of $(M^{-1})^{(\otimes t)}$ and an affine shift, with the property that the rate of C is at least $\mathrm{Capacity}(C) - \varepsilon$, and C can be encoded and decoded in time $O(n \log n)$ where $n = k^t$ and failure probability of the decoder is at most n^{-c}.*

The above theorem shows that all polar codes associated with every mixing matrix achieves the Shannon capacity of a symmetric memoryless channel efficiently, thus, vastly expanding on the class of polar codes known to satisfy this condition.

Our primary motivation in this work is to develop a general approach to proving polarization that applies to all matrices (matching the simple necessary condition for polarization) and is strong enough for the desired coding theory conclusion (convergence to capacity at polynomial block lengths, the distinguishing feature of polar codes). At the same time, our proof is arguably simpler and brings to light exactly what drives strong polarization — namely some simple local polarization conditions that hold for the single step evolution. One concrete motivation to consider polar codes with different choice of mixing matrices M is that an appropriate choice can lead to decoding error probability of $\exp(-n^\beta)$ for any $\beta < 1$ (as opposed to $\beta < 1/2$ for G_2) [14, 17], where $n = k^t$ is the block length of the code.

1.5 Comparison with Previous Analyses of (Strong) Polarization

While most of the ingredients going into our eventual analysis of strong polarization are familiar in the literature on polar codes, our proofs end up being much simpler and modular. We describe some of the key steps in our proofs and contrast them with those in previous works.

Definition of Local Polarization. While we are not aware of a definition similar to local polarization being explicit in the literature before, such notions have been considered implicitly before. For instance, for the variation in the middle (where we require that $\mathbb{E}[(X_{t+1}-X_t)^2] \geq \theta$ if $X_t \in (\tau, 1-\tau)$) the previous analyses in [9, 10] required θ be quadratic in τ. Indeed this was the most significant technical hurdle in the analysis for prime case in [9]. In contrast, our requirement on the variation is very weak and qualitative, allowing any function $\theta(\tau) > 0$. Similarly, our requirement in the *suction at the ends* case is relative mild and qualitative. In previous analyses the requirements were of the form "if $X_t \leq \tau$ then $X_{t+1} \leq X_t^2$ with positive probability." This high demand on the suction case prevented the analyses from relying only on the local behavior of the martingale $X_0, \ldots, X_t, \ldots$ and instead had to look at other parameters associated with it which essentially depend on the entire sequence. (For the reader familiar with previous analyses, this is where the Bhattacharyya parameters enter the picture.) Our approach, in contrast, only requires arbitrarily large constant factor drop, and thereby works entirely with the local properties of X_t.

Local Polarization Implies Strong Polarization. Our proof that local polarization implies strong polarization is short (about 3 pages) and comes in two parts. The first part uses a simple variance argument to show that X_t is exponentially close (in t) to the limit except with probability exponentially small in t. The second part then amplifies X_t's proximity to $\{0, 1\}$ to sub-exponentially small values using the suction at the end guarantee of each local step, coupled with Doob's martingale inequality and standard concentration inequalities. Such a two-part breakdown of the analysis is not new; however, our technical implementation is more abstract, more general and more compact all at the same time.

Local Polarization of Arıkan martingales. We will elaborate further on the approach for this after defining the Arıkan martingales, but we can say a little bit already now: First we essentially reduce the analysis of the polarization of Arıkan martingale associated with an arbitrary mixing matrix M to the analysis when $M = G_2$. This reduction loses in the parameters $(\alpha, \tau(\cdot), \theta(\cdot))$ specifying the level of local polarization, but since our strong polarization theorem works for any function, such loss in performance does not hurt the eventual result. Finally, local polarization for the case where the matrix is G_2 is of course standard, but even here our proofs (which we include for completeness) are simpler since they follow from known entropic inequalities on sums of two independent random variables. We stress that even quantitatively weak forms of these inequalities meet our requirements of local polarization, and we do not need strong forms of such inequalities (like Mrs. Gerber's lemma for the binary case [5, 10] and an ad hoc one for the prime case [9]).

Some Weakness in our Analyses. We first point out two weaknesses in our analyses. First, in contrast to the result of Mori and Tanaka [17] who characterize the set of matrices that lead to regular polarization over all fields, we only get a characterization over prime fields. Second, our definition of strong polarization only allows us to bound the failure probability of decoding by an arbitrarily small polynomial in the block length whereas results such as those in [3] actually get exponentially small (2^{-n^β} for some $\beta > 0$) failure probability.

In both cases we do not believe that these limitations are inherent to our approach. In particular the extension to general fields will probably involve more care, but should not run into major technical hurdles. Reducing the failure probability will lead to new technical challenges, but we do believe they can be overcome. Specificially, this requires stronger suction which is not true for the Arıkan martingale if one considers a single step evolution, but it seems plausible that multiple steps (even two) might show strong enough suction! We hope to investigate this in future work.

Organization of the Rest of this Paper. In this extended abstract, we give a sketch of the proof of our main new contribution (Theorem 1.6) that local polarization implies strong polarization in Section 2. We formally define the Arıkan martingale in Section 3 and give a very brief overview of our proof that the Arıkan martingale locally polarizes in Section 4. The full version of this paper contains all proofs.

2 LOCAL TO GLOBAL POLARIZATION

In this section we sketch the proof of Theorem 1.6, which asserts that every locally polarizing $[0,1]$-martingale is also strongly polarizing. The proof of this statement is implemented in two main steps: first, we show that any locally polarizing martingale, is $((1-\frac{\nu}{2})^t, (1-\frac{\nu}{4})^t)$-polarizing for some constant ν depending only on the parameters α, τ, θ of local polarization. This means that, except with exponentially small probability, $\min\{X_{t/2}, 1-X_{t/2}\}$ is exponentially small in t, which we can use to ensure that X_s for all $\frac{t}{2} \le s \le t$ stays in the range where the conditions of *suction at the ends* apply (again, except with exponentially small failure probability). Finally, we show that if the martingale stays in the *suction at the ends* regime, it will polarize strongly — i.e. if we have a $[0,1]$-martingale, such that in each step it has probability at least α to decrease by a factor of C, we can deduce that at the end we have $\Pr(X_T > C^{-\alpha T/4}) \le \exp(-\Omega(\alpha T))$.

We start by showing that in the first $t/2$ steps we do get exponentially small polarization, with all but exponentially small failure probability. This is proved using a simple potential function $\min\{\sqrt{X_t}, \sqrt{1-X_t}\}$ which we show shrinks by a constant factor, $1-\nu$ for some $\nu > 0$, in expectation at each step. Previous analyses in [9, 10] tracked $\sqrt{X_t(1-X_t)}$ (or some tailormade algebraic functions [11, 16]) as potential functions, and relied on quantitatively strong forms of variance in the middle to demonstrate that the potential diminishes by a constant factor in each step. While such analyses can lead to sharper bounds on the parameter ν, which in turn translate to better *scaling exponents* in the polynomial convergence to capacity, e.g. see [11, Thm. 18] or [16, Thm. 1], these analyses are more complex, and less general.

Lemma 2.1. *If a $[0,1]$-martingale sequence $X_0, \ldots X_t, \ldots,$ is $(\alpha, \tau(\cdot), \theta(\cdot))$-locally polarizing, then there exist $\nu > 0$, depending only on α, τ, θ, such that $\mathbb{E}[\min(\sqrt{X_t}, \sqrt{1-X_t})] \le (1-\nu)^t$.*

We defer the proof of this lemma to the full version of this paper, but give a brief idea of the proof here. W.l.o.g. we may consider the case $X_j < 1/2$ and note that it suffices to prove that $\mathbb{E}[X_{j+1}] \le (1-\nu)X_j$. In turn this claim follows easily from the facts that (1) $\mathbb{E}[X_{j+1}] = X_j$, (2) The square-root function is strictly concave (and so $\mathbb{E}[\sqrt{X_{j+1}}] < \sqrt{X_j}$ unless $X_{j+1} = X_j$ deterministically), and (3) for an appropriate choice of the threshold τ_0, the variance in the middle (when $X_j \in (\tau_0, 1/2)$), as well as the suction at the ends (when $X_j \le \tau_0$), conditions guarantee that the standard deviation of X_{j+1} is a constant multiple of X_j for all choices of $X_j < 1/2$. The following corollary is immediate from Lemma 2.1 and Markov's inequality.

Corollary 2.2. *If a $[0,1]$-martingale sequence $X_0, \ldots X_t, \ldots,$ is $(\alpha, \tau(\cdot), \theta(\cdot))$-locally polarizing, then there exist $\nu > 0$, depending only on α, τ, θ, such that $\Pr[\min(X_{t/2}, 1-X_{t/2}) > \lambda(1-\frac{\nu}{2})^t] \le (1-\frac{\nu}{4})^t \frac{1}{\sqrt{\lambda}}$.*

The next lemma will be used to show that if a [0,1]-martingale indeed stays at all steps $s \ge \frac{t}{2}$ in the *suction at the ends* range, i.e. in each step it has constant probability α of dropping by some large constant factor C, then expect it to be $(C^{-\alpha t/8}, \exp(-\Omega(\alpha t)))$-polarized.

Lemma 2.3. *There exists $c < \infty$, such that for all K, α with $K\alpha \ge c$ the following holds: Let X_t be a martingale satisfying $\Pr(X_{t+1} < e^{-K} X_t | X_t) \ge \alpha$, where $X_0 \in (0,1)$. Then $\Pr(X_T > \exp(-\alpha K T/4)) \le \exp(-\Omega(\alpha T))$.*

We give a full proof of this lemma in the full version of this paper, and just sketch the idea here. Consider the random variable $Y_j = \log X_j$. On the one hand we have $Y_{j+1} \le Y_j - K$ with probability at least α, and on the other hand we have that the probability that $Y_{j+1} \ge Y_j + i$ is at most 2^{-i} (for every positive i). Intuitively this suggests $\mathbb{E}[Y_{j+1} - Y_j] \le -K\alpha + c$ for some absolute constant c and so $\mathbb{E}[Y_t] \le Y_0 - t(K\alpha - c)$. Concentration results then can be

applied to show that the probability that Y_t does not drop by half this amount is exponentially small.

Given Corollary 2.2 and Lemma 2.3, Theorem 1.6 is essentially immediate. The only additional ingredient is an application of Doob's inequality to assert that once "moderate" polarization has occurred, the probability of coming out of a "suction end" is exponentially small. Again the full version of this paper has the details.

3 ARIKAN MARTINGALE

We now formally describe the Arıkan martingale associated with an invertible matrix $M \in \mathbb{F}_q^{k\times k}$ and a channel $C_{Y|Z}$. Briefly, this martingale measures at time t, the conditional entropy of a random variable $\mathbf{A'}_\mathbf{i}$, conditioned on the values of a vector of variables $\mathbf{B'}$ and on the values of $\mathbf{A'}_\mathbf{j}$ for $\mathbf{j}$ smaller than $\mathbf{i}$ for a random choice of the index $\mathbf{i}$. Here $\mathbf{A'}$ is a vector of k^t random variables taking values in $\mathbb{F}_q$ while $\mathbf{B'} \in \mathcal{Y}^{k^t}$. The exact construction of the joint distribution of these $2k^t$ variables is the essence of the Arıkan construction of codes, and we describe it shortly. The hope with this construction is that eventually (for large values of t) the conditional entropies are either very close to 0, or very close to $\log q$ for most choices of $\mathbf{i}$.

Before proceeding with a description of this joint distribution, as well as how the choice of $\mathbf{i}$ evolves with time, we will fix some notation. (A full description of our notational conventions is included in the full version of this paper). Below, we index vectors in $\mathbb{F}_q^{k^t}$ with t-tuples $\mathbf{i} \in [k]^t$. Let $\prec$ denote the lexicographic order on these t-tuples, i.e., $\mathbf{i} = (i_1, \ldots, i_t) \prec \mathbf{j} = (j_1, \ldots, j_t)$ if $i_\ell < j_\ell$ for the least index $\ell \in [t]$ for which $i_\ell \neq j_\ell$. We somewhat abuse the indexing notation, using $\mathbf{X}_{\prec \mathbf{i}}$ to mean the set of variables $\{\mathbf{X_j} : \mathbf{j} \prec \mathbf{i}\}$. We use notation $\mathbf{X}_{[\mathbf{i},\cdot]}$ to denote the slice of coordinates of $\mathbf{X}$ with prefix $\mathbf{i}$.

When $t = 1$, the process starts with k independent and identical pairs of variables $\{(A_i, B_i)\}_{i\in[k]}$ where $A_i \sim \mathbb{F}_q$ and $B_i \sim C_{Y|Z=A_i}$. (So each pair corresponds to an independent input/output pair from transmission of a uniformly random input over the channel $C_{Y|Z}$.) Let $\mathbf{A} = (A_1, \ldots, A_k)$ and $\mathbf{B'} = (B_1, \ldots, B_k)$, and note that the conditional entropies $H(A_i|\mathbf{A}_{<i}, \mathbf{B'})$ are all equal, and this entropy, divided by $\log_2 q$, will be our value of X_0. On the other hand, if we now let $\mathbf{A'} = \mathbf{A} \cdot M$ then the conditional entropies $H(\mathbf{A'}_i|\mathbf{A'}_{<i}, \mathbf{B'})$ are no longer equal (for most, and in particular for all mixing, matrices M). On the other hand, conservation of conditional entropy on application of an invertible transformation tells us that $\mathbb{E}_{i\sim[k]}[H(\mathbf{A'}_i|\mathbf{A'}_{<i}, \mathbf{B'})/\log_2 q] = X_0$. Thus letting $X_1 = H(\mathbf{A'}_i|\mathbf{A'}_{<i}, \mathbf{B'})/\log_2 q$ (for random i) gives us the martingale at time $t = 1$.

While this one step of multiplication by M *differentiates* among the k (previously identical) random variables, it doesn't yet polarize. The hope is by iterating this process one can get polarization[3]. But to get there we need to describe how to iterate this process. This iteration is conceptually simple and illustrated in Figure 1 (though notationally still complex). Roughly the idea is that at the beginning of stage t, we have defined a joint distribution of k^t dimensional vectors $(\mathbf{A}, \mathbf{B})$ along with a multi-index $\mathbf{i} \in [k]^t$. We now sample k independent and identically distributed pairs of these random variables $\{(\mathbf{A}^{(\ell)}, \mathbf{B}^{(\ell)})\}_{\ell\in[k]}$ and view $(\mathbf{A}^{(\ell)})_{\ell\in[k]}$ as a $k^t \times k$ matrix which we multiply by M to get a new $k^t \times k$ matrix. Flattening this matrix into a k^{t+1}-dimensional vector gives us a sample from the distribution of $\mathbf{A'} \in \mathbb{F}_q^{k^{t+1}}$. $\mathbf{B'}$ is simply the concatenation of all the vectors $(\mathbf{B}^{(\ell)})_{\ell\in[k]}$. And finally the new index $\mathbf{j} \in [k]^{t+1}$ is simply obtained by extending $\mathbf{i} \in [k]^t$ with a $(t+1)$th coordinate distributed uniformly at random in $[k]$. X_{t+1} is now defined to be $H(\mathbf{A'_j}|\mathbf{A'}_{<\mathbf{j}}, \mathbf{B'})/\log_2 q$. The formal description is below.

Definition 3.1 (Arıkan martingale). *Given an invertible matrix $M \in \mathbb{F}_q^{k\times k}$ and a channel description $C_{Y|Z}$ for $Z \in \mathbb{F}_q, Y \in \mathcal{Y}$, the Arıkan-martingale $X_0, \ldots X_t, \ldots$ associated with it is defined as follows. For every $t \in \mathbb{N}$, let D_t be the distribution on pairs $\mathbb{F}_q^{k^t} \times \mathcal{Y}^{k^t}$ described inductively below:*

A sample (A, B) from D_0 supported on $\mathbb{F}_q \times \mathcal{Y}$ is obtained by sampling $A \sim \mathbb{F}_q$, and $B \sim C_{B|A}$. For $t \geq 1$, a sample $(\mathbf{A'}, \mathbf{B'}) \sim D_t$ supported on $\mathbb{F}_q^{k^t} \times \mathcal{Y}^{k^t}$ is obtained as follows:

- *Draw k independent samples $(\mathbf{A}^{(1)}, \mathbf{B}^{(1)}), \ldots, (\mathbf{A}^{(k)}, \mathbf{B}^{(k)}) \sim D_{t-1}$.*
- *Let $\mathbf{A'}$ be given by $\mathbf{A'}_{[\mathbf{i},\cdot]} = (\mathbf{A}^{(1)}_\mathbf{i}, \ldots, \mathbf{A}^{(k)}_\mathbf{i}) \cdot M$ and $\mathbf{B'} = (\mathbf{B}^{(1)}, \mathbf{B}^{(2)}, \ldots \mathbf{B}^{(k)})$.*

Then, the sequence X_t is defined as follows: For each $t \in \mathbb{N}$, sample $i_t \in [k]$ iid uniformly. Let $\mathbf{j} = (i_1, \ldots, i_t)$ and let $X_t := H(\mathbf{A_j}|\mathbf{A}_{<\mathbf{j}}, \mathbf{B})/\log_2 q$, where the entropies are with respect to the distribution $(\mathbf{A}, \mathbf{B}) \sim D_t$. [4]

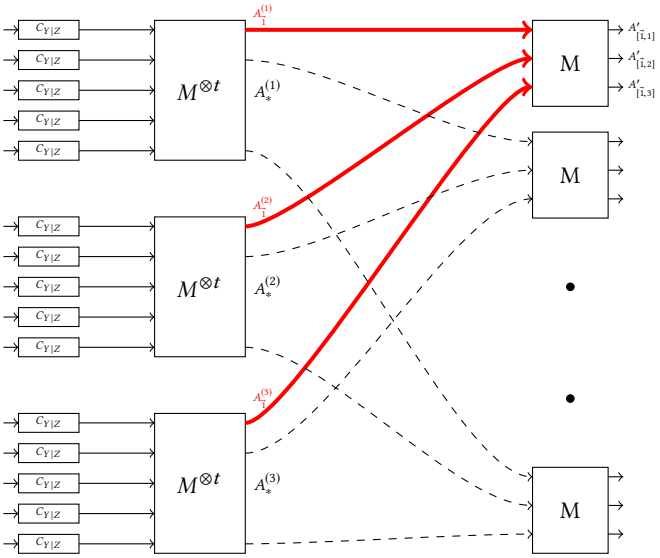

Figure 1: Figure showing evolution of Arıkan martingale for 3×3 matrix M.

[3] In the context of Polar coding, *differentiation* and *polarization* are good events, and hence our "hope."

[4] We stress that the only randomness in the evolution of X_t is in the choice of $i_1, \ldots, i_t, \ldots$. The process of sampling $\mathbf{A}$ and $\mathbf{B}$ is only used to define the distributions for which we consider the conditional entropies $H(\mathbf{A_j}|\mathbf{A}_{<\mathbf{j}}, \mathbf{B})$.

Figure 1 illustrates the definition by highlighting the construction of the vector $\mathbf{A}'$, and in particular highlights the recursive nature of the construction.

It is easy (and indeed no different than in the case $t = 1$) to show that $\mathbb{E}[X_{t+1}|X_t] = X_t$ and so the Arıkan martingale is indeed a martingale. This is asserted below and proved in the full version of this paper.

Proposition 3.2. *For every matrix M and channel $C_{Y|Z}$, the Arıkan martingale is a martingale and in particular a $[0,1]$-martingale.*

Finally, we remark that based on the construction it is not too hard to see that if M were an identity matrix, or more generally a non-mixing matrix, then X_t would deterministically equal X_0. (There is no differentiation and thus no polarization.) The thrust of this paper is to show that in all other cases we have strong polarization.

4 PROOF OVERVIEW OF LOCAL POLARIZATION

Here we describe the overall structure of the proof of Theorem 1.10, which states that the Arıkan martingale is locally polarizing.

Theorem 1.10. *For every prime q, for every mixing matrix $M \in \mathbb{F}_q^{k\times k}$, and for every symmetric channel $C_{Y|Z}$ over $\mathbb{F}_q$, the associated Arıkan martingale sequence is locally polarizing.*

We describe the main ideas for the case $t = 1$. (All other steps are similar.) Let $\mathbf{A} \in \mathbb{F}_q^k$ be a random vector, and let W be an arbitrary random variable such that the entries of $\mathbf{A}$ are independent and identically-distributed, conditioned on W. In what follows, let $\overline{H}(\cdot|\cdot) = H(\cdot|\cdot)/\log_2 q$. Let $X_0 := \overline{H}(\mathbf{A}_1|W)$ be the conditional entropy of each entry of $\mathbf{A}$.

Now let $\mathbf{A}' := \mathbf{A} \cdot M$ be the variables obtained in the next step of polarization. Local polarization of the Arıkan martingale boils down to showing that for a random index $i \in [k]$, the conditional entropies of the transformed variables $X_1 \sim \overline{H}(\mathbf{A}'_i|\mathbf{A}'_{<i}, W)$ satisfy the local polarization conditions (of variance in the middle and suction at the ends).

Our analysis follows roughly two steps: First we focus on a simple case $M = G_2 = \left(\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right)$. This turns our attentions to the quantities $\overline{H}(A_1 + A_2|W)$ and $\overline{H}(A_2|W, A_1 + A_2)$. Variance in the middle in this case corresponds to proving that $\overline{H}(A_1 + A_2|W) - \overline{H}(A_1|W)$ is bounded away from zero if $\overline{H}(A_1, W) \in (\tau, 1-\tau)$. This statement is already explicit in the literature (see [5, Lemma 4.2]) and follows relatively easily from the concavity of the (conditional) entropy. Suction at the ends also follows easily from the properties of entropy. Specifically in the low-end after conditioning on W, low entropy implies that A_i takes on some value with very high probability, and the event that it does not take on this modal value happens with low probability. With $A_1 + A_2$ the probability the unlikely event is the probability that at least one of them takes on the unlikely value, and this probability is roughly twice the probability that any one of them takes on the unlikely value. This fact in turn immediately implies that $\overline{H}(A_2|W, A_1 + A_2)$ is much smaller than $\overline{H}(A_1|W)$. The full version of this paper argues this explicitly. Finally for the suction at the high end, we use elementary Fourier analysis and note that the non-zero Fourier coefficients of the distribution of A_1 (after conditioning on W) can be used to estimate its entropy, and these Fourier coefficients square (exactly) when considering $A_1 + A_2$, showing that $\overline{H}(A_1 + A_2|W)$ is much closer to 1 than $\overline{H}(A_1|W)$. Again this is formally shown in the full version.

These lemmas immediately suffice to prove local polarization when $M = G_2$. The general case is not much harder. It turns out that using the fact that entropy is preserved under invertible transforms we can argue that the conditional entropies $\overline{H}(\mathbf{A}'_j|\mathbf{A}'_{<j})$ are preserved if we replace M by DM for an invertible lower triangular matrix D. Furthermore we also have that $\overline{H}(Y|\mathbf{Z}) = \overline{H}(Y|\mathbf{T}\cdot\mathbf{Z})$ for any variable Y, vector $\mathbf{Z}$ and invertible matrix $\mathbf{T}$. Using these simple facts we establish that w.l.o.g. we can consider a matrix M such that there exists a j, j' and non-zero $\alpha, \alpha' \in \mathbb{F}_q$ such that $\overline{H}(\mathbf{A}'_j|\mathbf{A}'_{<j}) \geq \overline{H}(A_1+\alpha A_2)$, and $\overline{H}(\mathbf{A}'_{j'}|\mathbf{A}'_{<j'}) \leq \overline{H}(A_1|A_1+\alpha' A_2)$. This allows us to use the analysis in the 2×2 case to argue that the conditions of local polarization are met by the Arikan martingale associated with any mixing matrix. (We note that the parameters of local polarization do drop by factors that are polynomial in k since we only show the existence of such j, j'). The full details of this step may be found in the full version of this paper, culminating in the proof of Theorem 1.10.

REFERENCES

[1] Emmanuel Abbe and Emre Telatar. 2012. Polar Codes for the m–User Multiple Access Channel. *IEEE Transactions on Information Theory* 58, 8 (2012), 5437–5448.

[2] Erdal Arıkan. 2009. Channel Polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels. *IEEE Transactions on Information Theory* (July 2009), 3051–3073.

[3] Erdal Arıkan and Emre Telatar. 2009. On the Rate of Channel Polarization. In *Proceedings of 2009 IEEE International Symposium on Information Theory*. 1493–1495.

[4] Thomas M. Cover and Joy A. Thomas. 2005. *Elements of Information Theory* (2nd ed.). John Wiley and Sons, Hoboken, NJ, USA.

[5] Eren Şaşoğlu. 2012. Polarization and Polar Codes. *Foundations and Trends in Communications and Information Theory* 8, 4 (2012), 259–381. https://doi.org/10.1561/0100000041

[6] Furkan Ercan, Carlo Condo, Seyyed Ali Hashemi, and Warren J. Gross. 2017. On Error-Correction Performance and Implementation of Polar Code List Decoders for 5G. In *55th Annual Allerton Conference on Communication, Control, and Computing*.

[7] Arman Fazeli, S. Hamed Hassani, Marco Mondelli, and Alexander Vardy. 2017. Binary linear codes with optimal scaling and quasi-linear complexity. *Personal communication* (October 2017).

[8] Naveen Goela, Emmanuel Abbe, and Michael Gastpar. 2013. Polar codes for broadcast channels. In *Proceedings of the 2013 IEEE International Symposium on Information Theory, Istanbul, Turkey, July 7-12, 2013*. 1127–1131.

[9] Venkatesan Guruswami and Ameya Velingker. 2015. An Entropy Sumset Inequality and Polynomially Fast Convergence to Shannon Capacity Over All Alphabets. In *Proceedings of 30th Conference on Computational Complexity*. 42–57.

[10] Venkatesan Guruswami and Patrick Xia. 2015. Polar Codes: Speed of Polarization and Polynomial Gap to Capacity. *IEEE Trans. Information Theory* 61, 1 (2015), 3–16. Preliminary version in Proc. of FOCS 2013.

[11] S.H. Hassani, K. Alishahi, and R. Urbanke. 2014. Finite-Length Scaling for Polar Codes. *Information Theory, IEEE Transactions on* PP, 99 (2014), 1–1. https://doi.org/10.1109/TIT.2014.2341919

[12] Seyed Hamed Hassani, Kasra Alishahi, and Rüdiger L. Urbanke. 2014. Finite-Length Scaling for Polar Codes. *IEEE Trans. Information Theory* 60, 10 (2014), 5875–5898. https://doi.org/10.1109/TIT.2014.2341919

[13] Satish Babu Korada. 2010. Polar codes for Slepian-Wolf, Wyner-Ziv, and Gelfand-Pinsker. In *Proceedings of the 2010 IEEE Information Theory Workshop*. 1–5.

[14] Satish Babu Korada, Eren Sasoglu, and Rüdiger L. Urbanke. 2010. Polar Codes: Characterization of Exponent, Bounds, and Constructions. *IEEE Transactions on Information Theory* 56, 12 (2010), 6253–6264.

[15] Hessam Mahdavifar and Alexander Vardy. 2011. Achieving the Secrecy Capacity of Wiretap Channels Using Polar Codes. *IEEE Transactions on Information Theory* 57, 10 (2011), 6428–6443.

[16] Marco Mondelli, S. Hamed Hassani, and Rüdiger L. Urbanke. 2016. Unified Scaling of Polar Codes: Error Exponent, Scaling Exponent, Moderate Deviations, and Error Floors. *IEEE Trans. Information Theory* 62, 12 (2016), 6698–6712. https://doi.org/10.1109/TIT.2016.2616117

[17] Ryuhei Mori and Toshiyuki Tanaka. 2014. Source and Channel Polarization Over Finite Fields and Reed-Solomon Matrices. *IEEE Trans. Information Theory* 60, 5 (2014), 2720–2736.

[18] Henry D. Pfister and Rüdiger L. Urbanke. 2016. Near-optimal finite-length scaling for polar codes over large alphabets. In *IEEE International Symposium on Information Theory, ISIT*. 215–219.

[19] Eren Sasoglu. 2012. Polarization and Polar Codes. *Foundations and Trends in Communications and Information Theory* 8, 4 (2012), 259–381.

[20] Eren Sasoglu, Emre Telatar, and Edmund M. Yeh. 2013. Polar Codes for the Two-User Multiple-Access Channel. *IEEE Transactions on Information Theory* 59, 10 (2013), 6583–6592.

[21] Ido Tal and Alexander Vardy. 2013. How to Construct Polar Codes. *IEEE Transactions on Information Theory* 59, 10 (Oct 2013), 6562–6582.

[22] Lele Wang and Eren Sasoglu. 2014. Polar coding for interference networks. In *2014 IEEE International Symposium on Information Theory, Honolulu, HI, USA, June 29 - July 4, 2014*. 311–315. https://doi.org/10.1109/ISIT.2014.6874845