Reed-Muller codes polarize

Emmanuel Abbe*† and Min Ye†
*Mathematics Institute and the School of Computer and Communication Sciences at EPFL, Switzerland
Program in Applied and Computational Mathematics in Princeton University, USA
†Department of Electrical Engineering in Princeton University, USA

Abstract—**Reed-Muller (RM) codes were introduced in 1954 and have long been conjectured to achieve Shannon's capacity on symmetric channels. The activity on this conjecture has recently been revived with the emergence of polar codes. RM codes and polar codes are generated by the same matrix $G_m = \left[\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right]^{\otimes m}$ but using different subset of rows. RM codes select simply rows having largest weights. Polar codes select instead rows having the largest conditional mutual information proceeding top to down in G_m; while this is a more elaborate and channel-dependent rule, the top-to-down ordering allows Arıkan to show that the conditional mutual information polarizes, and this gives directly a capacity-achieving code on any symmetric channel. RM codes are yet to be proved to have such a property, despite the recent success for the erasure channel.**

In this paper, we connect RM codes to polarization theory. We show that proceeding in the RM code ordering, i.e., not top-to-down but from the lightest to the heaviest rows in G_m, the conditional mutual information again polarizes. We further demonstrate that it does so faster than for polar codes. This implies that G_m contains another code, different than the polar code and called here the twin-RM code, that is provably capacity-achieving on any symmetric channel. This gives in particular a necessary condition for RM codes to achieve capacity on symmetric channels. It further gives a sufficient condition if the rows with largest conditional mutual information correspond to the heaviest rows, i.e., if the twin-RM code is the RM code. We demonstrate here that the two codes are at least similar and give further evidence that they are indeed the same.

Index Terms—**Reed-Muller codes; polar codes; Shannon theory; capacity-achieving codes**

I. Introduction

Reed-Muller codes were introduced in 1954 by Muller [1] and studied shortly after by Reed [2]. They are among the first and simplest codes to construct (evaluations of multivariate polynomials of bounded degree) and have a wide range of applications in theoretical computer science, such as in [3]–[6]. Moreover, Reed-Muller codes have long been conjectured to achieve Shannon's capacity on symmetric channels.[1] This was recently settled in [7], [8] for the special case of the binary erasure channel (BEC), and in [9], [10] for special cases of extremal rates on both the BEC and the binary symmetric channel (BSC). The general conjecture of achieving capacity on the BSC and more generally any binary memoryless symmetric (BMS) channel[2] at constant rate remains widely open to date.

The research activity on RM codes has resurged in part due to the development of polar codes [11], [12]. Both RM codes and polar codes are generated by selecting subset of rows from the same base matrix $G_m = \left[\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right]^{\otimes m}$. Polar codes select the rows by tracking the conditional mutual information of each row given the past rows when proceeding top to down in G_m (see Section II-B for precise definitions). In this specific ordering, Arıkan was able to show a polarization result [11], i.e., that most of the rows have a conditional mutual information that tend to either 0 or 1. This in turn implies fairly directly that the code resulting from keeping the high conditional mutual information rows is capacity-achieving on any BMS channel.

A first drawback of polar codes is that the code construction, i.e., identifying the rows having high conditional mutual information, is non-trivial. In particular, there is to date no known explicit characterization of the row selection except for the BEC. This is however not an algorithmic limitation as there are known efficient algorithms that approximate arbitrarily closely the values of the conditional mutual information for each row [13]. Two more important drawbacks are first that polar codes are not universal [14], as their row selection is channel dependent, and then that their scaling law is sub-optimal compared to that of random codes [15] and likely of RM codes [16], making their error probability at short block length not as competitive as could be [17]. On the flip side, polar codes benefit from a powerful analytical framework, the polarization framework [11], recently strengthened in [18], [19], which allows to give performance guarantees, as well as an efficient successive decoding algorithm. Their performance at short block length has also been improved with the addition of outer codes and list decoding algorithms [20]. With these attributes, polar codes are to enter soon the 5G standards [21].

On the other side, RM codes benefit from a simple and universal code construction: selecting the heaviest rows is trivial and depends only on the capacity of the channel and not the actual channel. Further, it is already known that RM codes would have an optimal scaling law *if* they were proved to be capacity-achieving [16]. Performance improvements over polar codes at short block length were also demonstrated in [17]. On the flip side, the main challenges of RM codes are (i) their analytical framework, with the difficulty of obtaining performance guarantees, (ii) the absence of an efficient decoding algorithm that succeeds up to capacity for the constant rate regime.

[1] See [7] for accounts on this conjecture.

[2] Recall that a BMS channel is a channel $W : \{0,1\} \to \mathcal{Y}$ such that there is a permutation π on the output alphabet $\mathcal{Y}$ satisfying i) $\pi^{-1} = \pi$ and ii) $W(y|1) = W(\pi(y)|0)$ for all $y \in \mathcal{Y}$.

2575-8454/19/$31.00 ©2019 IEEE
DOI 10.1109/FOCS.2019.00026

A. Recent progress

As mentioned earlier, progress has recently been made on both points (i) and (ii). We mention briefly here a few references for decoding algorithms [2], [22]–[24], as this is not the focus of this paper, and also refer to our parallel paper with a decoding algorithm [25] for a more detailed discussion of those.

We now discuss performance guarantees. In [7], the case of the BEC is settled by exploiting results on the threshold of monotone Boolean functions, benefiting from the fact that the events of decoding failures for erasures correspond to monotone properties of Boolean functions. With this link, general results from Boolean function analysis [26]–[28] come to rescue and allow to close the conjecture for the BEC. While this gives an elegant proof, it has the downside of relying on a "Hammer" result for monotone functions [26], [29] that does not seem to generalize easily beyond erasures due to the loss of the monotonicity property. The approach of [9] relies instead on the polynomial characterization of RM codes (whose codewords are the evaluation vectors of bounded degree multivariate Boolean polynomials) and on the weight enumerator of RM codes. A downside of that approach is that it is currently not reaching the constant rate regime; although some recent progress towards that goal was made in [10].

Moreover, none of the above seem to shed light on the connection between polar and RM codes, which despite being relatives are analyzed very differently. Attempts to connect RM codes to polar code was made in [30], using the double conditional rank measure in relation to the algebraic approach to polarization [31], with conjectures based on this approach left in [30]. Another work investigating the relationship between RM and polar codes used the polynomial formalism and studied the generating monomials of polar codes [32].

B. This paper

Considering the developments so far, it may appear that the simplicity of the RM code construction fires back in the complexity of their analysis, in contrast to polar codes, where a more elaborate construction allows to benefit from the powerful polarization framework.

The first goal of this paper is to show that this is not a necessary limitation, and that RM codes benefit too from a polarization phenomenon, slightly different but potentially more effective than that of polar codes. We view RM codes as the evaluation of multivariate polynomials and make use of the classical recursive Plotkin construction[3] $(\boldsymbol{u}, \boldsymbol{u}+\boldsymbol{v})$ [33], which is similar in nature to the recursive construction of polar codes. Together with the establishment of an ordering on the conditional mutual information of RM codes, we derive a new polarization result for the RM code ordering.

This in turn gives rise to a new code, called here the twin-RM code, obtained by selecting rows with high conditional mutual information in the RM code ordering. We can prove that this twin-RM code is capacity-achieving on *any* BMS channel, as conjectured for RM codes. We further show that the twin-RM code is indeed closely related to the RM code in the following sense.

First the polarization of conditional mutual information under RM ordering is a necessary condition for RM codes to achieve capacity over any BMS channel.[4] Moreover, the capacity results for the twin-RM code also give a sufficient condition for capacity results on the RM code *if* the rows with largest conditional mutual information correspond mostly to the heaviest rows, i.e., if the twin-RM code is equivalent to the RM code. We give here a relaxed version of the latter, showing that the twin-RM code has similarity with the RM code, and verify that it is indeed exactly the RM code up to dimension 16 (for the BSC). Note that in the contrary case, i.e., if the twin-RM code were not equivalent to the RM code, then this would imply that the RM code does not achieve capacity on all BMS channels.

II. Background

A. RM codes

Consider the polynomial ring $\mathbb{F}_2[Z_1, Z_2, \ldots, Z_m]$ of m variables over $\mathbb{F}_2$. Since $Z^2 = Z$ in $\mathbb{F}_2$, the following set of 2^m monomials forms a basis of $\mathbb{F}_2[Z_1, Z_2, \ldots, Z_m]$:

$$\{\prod_{i\in A} Z_i : A \subseteq [m]\}, \text{ where } \prod_{i\in\emptyset} Z_i := 1.$$

Next associate to every subset $A \subseteq [m]$ a row vector $\boldsymbol{v}_m(A)$ of length 2^m, whose components $\boldsymbol{v}_m(A, \boldsymbol{z})$ are indexed by binary vectors $\boldsymbol{z} = (z_1, z_2, \ldots, z_m) \in \{0,1\}^m$,

$$\boldsymbol{v}_m(A, \boldsymbol{z}) = \prod_{i\in A} z_i, \tag{1}$$

i.e., $\boldsymbol{v}_m(A, \boldsymbol{z})$ is the evaluation of the monomial $\prod_{i\in A} Z_i$ at $\boldsymbol{z}$. For $0 \le r \le m$, the set of vectors

$$\{\boldsymbol{v}_m(A) : A \subseteq [m], |A| \le r\}$$

forms a basis of the r-th order Reed-Muller code $\mathcal{R}(m, r)$ of length $n := 2^m$ and dimension $\sum_{i=0}^{r} \binom{m}{i}$.

Definition 1. *The r-th order Reed-Muller code $\mathcal{R}(m, r)$ code is defined as the following set of binary vectors*

$$\mathcal{R}(m, r) := \Big\{ \sum_{A\subseteq[m], |A|\le r} u(A)\boldsymbol{v}_m(A) : u(A) \in \{0,1\} \\ \textit{for all } A \subseteq [m], |A| \le r \Big\}.$$

In this paper, we prove a polarization result in the Reed-Muller code ordering. To that end, we define a total order on all the subsets of $[m]$ as follows:

Definition 2 (Total order)**.** *For $A = \{a_1, a_2, \ldots, a_{|A|}\}, B = \{b_1, b_2, \ldots, b_{|B|}\} \subseteq [m]$, where $a_1 > a_2 > \cdots > a_{|A|}$ and*

[3] Any d-degree polynomial can be decomposed with two $(d-1)$-degree polynomials as $f(x^d) = x_d f_1(x^{d-1}) + f_0(x^{d-1})$.

[4] If RM codes achieves capacity, then the conditional entropies in the RM ordering must tend to 0 for the heavy-weight rows and 1 for the light-weight rows, so polarization must happen (besides in the critical window).

Example 1. *We write out a basis of $\mathcal{R}(3,3)$ as follows:*

$$\begin{array}{cccccccccc}
(z_1,z_2,z_3) & (1,1,1) & (1,1,0) & (1,0,1) & (1,0,0) & (0,1,1) & (0,1,0) & (0,0,1) & (0,0,0) \\
A=\{3,2,1\} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
A=\{2,1\} & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
A=\{3,1\} & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
A=\{3,2\} & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
A=\{1\} & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
A=\{2\} & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\
A=\{3\} & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\
A=\emptyset & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{array},$$

where the first row lists the index $\boldsymbol{z}$ of each component, and the second to the last rows are $\boldsymbol{v}_3(A), A \subseteq [3]$.

$b_1 > b_2 > \cdots > b_{|B|}$, *we write $A < B$ if either of the following two conditions is satisfied:*

1) *$|A| > |B|$;*
2) *$|A| = |B|$, and there is an integer $i \in \{1,2,\dots,|A|\}$ such that $a_j = b_j \forall j < i$ and $a_i < b_i$.*

It is easy to check that for any two sets $A, B \subseteq [m]$, one of the following three relations must hold: $A < B, A = B$ or $A > B$. Therefore, this is indeed a total order on all the subsets of $[m]$. Note that condition 1) ensures that picking the 'largest' sets 'layer by layer' (i.e., with all sets of the same cardinality together) gives the RM code. Condition 2) says how to order the rows within a layer (e.g., if the code dimension requires breaking a layer), but any ordering resulting from a permutation of the elements in $[m]$ would be equivalent. We pick this convention as we will see the m-th element as the 'new element' when running the forthcoming inductions.

For $m = 3$, the rows in Example 1 are listed in the increasing order of the set A. Let $(U_A^{(m)} : A \subseteq [m])$ be 2^m i.i.d. Bernoulli-$1/2$ random variables. We use the shorthand notation $U_{<A}^{(m)} := (U_{A'}^{(m)} : A' \subseteq [m], A' < A)$ and $U^{(m)} := (U_A^{(m)} : A \subseteq [m])$. Next we define another n i.i.d. Bernoulli-$1/2$ random variables $X_{\boldsymbol{z}}^{(m)}, \boldsymbol{z} \in \{0,1\}^m$ by

$$(X_{\boldsymbol{z}}^{(m)}, \boldsymbol{z} \in \{0,1\}^m) := \sum_{A \subseteq [m]} U_A^{(m)} \boldsymbol{v}_m(A).$$

We transmit $X_{\boldsymbol{z}}^{(m)}, \boldsymbol{z} \in \{0,1\}^m$, through n independent copies of a BMS channel $W : \{0,1\} \to \mathcal{Y}$, and we denote the corresponding channel outputs as $Y_{\boldsymbol{z}}^{(m,W)}, \boldsymbol{z} \in \{0,1\}^m$. Let $X^{(m)} := (X_{\boldsymbol{z}}^{(m)} : \boldsymbol{z} \in \{0,1\}^m)$ and $Y^{(m,W)} := (Y_{\boldsymbol{z}}^{(m,W)} : \boldsymbol{z} \in \{0,1\}^m)$. For instance, if W is simply the binary symmetric channel (BSC), then $Y_{\boldsymbol{z}}^{(m,W)} = X_{\boldsymbol{z}}^{(m)}$ with probability $(1-p)$ and $Y_{\boldsymbol{z}}^{(m,W)} = X_{\boldsymbol{z}}^{(m)} \oplus 1$ with probability p.

In general, since W is symmetric and $(X_{\boldsymbol{z}}^{(m)}, \boldsymbol{z} \in \{0,1\}^m)$ are also i.i.d. Bernoulli-$1/2$ random variables, we have for all $\boldsymbol{z} \in \{0,1\}^m$, $H(X_{\boldsymbol{z}}^{(m)} | Y_{\boldsymbol{z}}^{(m,W)}) = 1 - I(W)$, and therefore $H(U^{(m)} | Y^{(m,W)}) = H(X^{(m)} | Y^{(m,W)}) = nH(X_{\boldsymbol{z}}^{(m)} | Y_{\boldsymbol{z}}^{(m,W)}) = n(1 - I(W))$, where $H(\cdot|\cdot)$ is the conditional entropy defined as $H(X|Y) = -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p_{X|Y}(x|y) p_Y(y) \log p_{X|Y}(x|y)$ and $I(\cdot)$ is the channel capacity (or the symmetric capacity for non-BMS channels) defined as $I(W) = (1/2) \sum_{x \in \mathbb{F}_2, y \in \mathcal{Y}} W(y|x) \log \frac{W(y|x)}{\sum_{u \in \mathbb{F}_2} W(y|u)}$. Thus, if X is Bernoulli-$1/2$ and Y is drawn from $W(\cdot|X)$, we have $I(W) = 1 - H(X|Y)$, and further,

$$\sum_{A \subseteq [m]} H(U_A^{(m)} | Y^{(m,W)}, U_{<A}^{(m)}) = H(U^{(m)} | Y^{(m,W)}) = H(X^{(m)} | Y^{(m,W)}) = n(1 - I(W)),$$

which means that the sum of all the conditional entropies on the left gives exactly the total conditional entropy of the original channel (i.e., the entropy is preserved). For convenience, we use the notation

$$H_A^{(m,W)} := H(U_A^{(m)} | Y^{(m,W)}, U_{<A}^{(m)}). \tag{2}$$

From now on, we omit to specify W from the notation $H_A^{(m,W)}, Y^{(m,W)}$ and $Y_{\boldsymbol{z}}^{(m,W)}$ when the underlying channel is not important, i.e., we write them as $H_A^{(m)}, Y^{(m)}$ and $Y_{\boldsymbol{z}}^{(m)}$. Therefore,

$$\sum_{A \subseteq [m]} H_A^{(m)} = n(1 - I(W)). \tag{3}$$

We also define the channel $W_A^{(m)}$ as the binary-input channel that takes $U_A^{(m)}$ as input and $Y^{(m)}, U_{<A}^{(m)}$ as outputs, i.e., $W_A^{(m)}$ is the channel seen by the successive decoder when decoding $U_A^{(m)}$.

In order to state our main results, we also need the definition of the Bhattacharyya parameter. Let (X, Y) be a pair of random variables such that X has Bernoulli-$1/2$ distribution and Y takes values from a finite alphabet $\mathcal{Y}$. The Bhattacharyya parameter is defined as

$$Z(X|Y) := \sum_{y \in \mathcal{Y}} \sqrt{P_{Y|X}(y|0) P_{Y|X}(y|1)}.$$

Similarly to $H_A^{(m)}$, we use the shorthand notation $Z_A^{(m)} = Z_A^{(m,W)} := Z(U_A^{(m)} | Y^{(m,W)}, U_{<A}^{(m)})$.

B. Polarization

The polar coding transform is given by the following $n \times n$ matrix

$$G_m := \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}^{\otimes m}, \tag{4}$$

where $\otimes$ is the Kronecker product and $n = 2^m$. Let (U_1, U_2) be a pair of i.i.d. uniform random variables, and let $(X_1, X_2) = (U_1, U_2)G_1$. Then transmit X_1 and X_2 through two copies of W. Under successive decoder, this transforms two copies of W into a "worse" channel $W^- : U_1 \to Y_1, Y_2$ and a "better" channel $W^+ : U_2 \to U_1, Y_1, Y_2$. This statement can be quantified with conditional entropies as follows:

$$H(U_1|Y_1, Y_2) \geq H(X_1|Y_1) \geq H(U_2|U_1, Y_1, Y_2), \quad (5)$$

together with the entropy-preservation equation

$$H(U_1|Y_1, Y_2) + H(U_2|U_1, Y_1, Y_2) = 2H(X_1|Y_1). \quad (6)$$

Similar relations among the Bhattacharyya parameters were also proved in [11, Proposition 5]:

$$Z(U_2|Y_1, Y_2, U_1) = (Z(X_1|Y_1))^2, \quad (7)$$

$$Z(U_1|Y_1, Y_2) \geq Z(X_1|Y_1). \quad (8)$$

Moreover, if $H(X_1|Y_1)$ is bounded away from 0 and 1, then the gap between $H(U_1|Y_1, Y_2)$ and $H(X_1|Y_1)$ is bounded away from 0, and so does the gap between $H(X_1|Y_1)$ and $H(U_2|U_1, Y_1, Y_2)$. (By (6), these two gaps are the same.) In other words, if W is neither noiseless nor completely noisy, then W^- is strictly worse than W, and W^+ is strictly better. The rigorous statement is as follows.

Lemma 1 ([11]). *Let (X_1, Y_1) and (X_2, Y_2) be an independent pair of random variables, where X_1 and X_2 take values in $\{0, 1\}$. For all $\epsilon > 0$, there is $\delta(\epsilon) > 0$ such that $H(X_1|Y_1), H(X_2|Y_2) \in (\epsilon, 1 - \epsilon)$ implies $H(X_1 + X_2|Y_1, Y_2) \geq \max(H(X_1|Y_1), H(X_2|Y_2)) + \delta(\epsilon)$.*

The polar coding scheme consists of applying the polar matrix G_m to n i.i.d. uniform random variables and transmitting the results through n copies of W. This amounts to iteratively applying the "+" and "−" polar transforms to W, and the power of this is that almost all the resulting bit-channels seen by the successive decoder become either noiseless or completely noisy. We next state this formally; let $\{\widetilde{H}_A^{(m)}\}_{A \subseteq [m]}$ (resp., $\{\widetilde{Z}_A^{(m)}\}_{A \subseteq [m]}$) be the conditional entropy (resp., Bhattacharyya parameter) of each row given all the past rows when decoding top to down in G_m.

Theorem 0 (**Polarization of polar codes** [11]). *For every BMS channel W, almost all elements in the set $\{\widetilde{H}_A^{(m)}\}_{A \subseteq [m]}$ are close to either 0 or 1 when m is large. More precisely, for any $0 < \epsilon < 1/10$ and any $\delta > 0$, there is a constant $M(\epsilon, \delta)$ such that for every $m > M(\epsilon, \delta)$,*

$$\frac{\left|\left\{A \subseteq [m] : \widetilde{H}_A^{(m)} > 1 - \epsilon\right\} \cup \left\{A \subseteq [m] : \widetilde{Z}_A^{(m)} < \delta\right\}\right|}{2^m}$$
$$\geq 1 - o(1).$$

III. Main results

Our main results are summarized in the following theorems.

Theorem 1 (**Polarization of RM codes**). *For every BMS channel W, almost all elements in the set $\{H_A^{(m)}\}_{A \subseteq [m]}$ are close to either 0 or 1 when m is large. More precisely, for any $0 < \epsilon < 1/10$ and any $\delta > 0$, there is a constant $M(\epsilon, \delta)$ such that for every $m > M(\epsilon, \delta)$,*

$$\frac{\left|\left\{A \subseteq [m] : H_A^{(m)} > 1 - \epsilon\right\} \cup \left\{A \subseteq [m] : Z_A^{(m)} < \delta\right\}\right|}{2^m}$$
$$\geq 1 - o(1).$$

As mentioned above, the basis vectors of RM codes $\{\boldsymbol{v}_m(A) : A \subseteq [m]\}$ are exactly the row vectors of the polar matrix G_m in (4). However, these rows are arranged in different orders for RM codes and polar codes, which makes the polarization of RM codes fundamentally different from that of polar codes.

As an immediate consequence of Theorem 1, or more precisely its quantitative version in Theorem 7 that allows us to pick $\delta_n = \text{poly}(1/n)$, we can construct a family of capacity-achieving codes as follows.

Theorem 2 (**Twin-RM codes achieve capacity**). *For a BMS channel W and $\delta_n > 0$, let*

$$\mathcal{G}(m, \delta_n) := \left\{A \subseteq [m] : Z_A^{(m)} < \delta_n\right\}$$

and define the family of twin-RM codes from the codewords

$$\mathcal{T}(m, \delta_n) :=$$
$$\left\{\sum_{A \in \mathcal{G}} u(A)\boldsymbol{v}_m(A) : u(A) \in \{0, 1\} \text{ for all } A \in \mathcal{G}(m, \delta_n)\right\},$$

where $\boldsymbol{v}_m(A)$ is defined in (1). Then, taking $\delta_n = n^{-1-\eta}$, $\eta > 0$, $\mathcal{T}(m, \delta_n)$ achieves the capacity of W under successive decoding.

This theorem tells us that we can construct capacity achieving codes using successive decoder under the RM ordering (i.e., the ordering defined by the weights of the rows in G_m). Note that none of the above give algorithmic results.

To establish the above, we need the following notion of ordering between the different conditional entropies in the RM ordering, which also exhibits some of the similarity between the RM and twin-RM codes.

Definition 3. *For $A = \{a_1, a_2, \dots, a_{|A|}\}, B = \{b_1, b_2, \dots, b_{|B|}\} \subseteq [m]$, $A \neq B$, where $a_1 < a_2 < \cdots < a_{|A|}$ and $b_1 < b_2 < \cdots < b_{|B|}$, we define*

$$A \prec B \text{ if and only if } |A| \geq |B| \text{ and } a_i \leq b_i, \forall i \leq |B|.$$

Note that we set the above to be $A \prec B$ and not $A \succ B$ as this also gives $[m]$ as the 'first' set and $\emptyset$ as the 'last' set, as for the RM code ordering.

Theorem 3 (**Partial order**). *If $A \prec B$, then $H_A^{(m)} \geq H_B^{(m)}$.*

According to Theorem 2 and Theorem 3, the twin-RM code $\mathcal{T}(m, \delta_n)$ tend to select sets A with small cardinality, which is similar to RM codes (that exactly selects sets with the smallest cardinality). However, we can not establish here whether this is exactly the RM code or not. We do give a positive indication by proving that this is exactly the RM code up to $n = 16$ for the BSC in Section VI-F.

IV. Proof outline

In order to explain the main ideas of the proof, we introduce the following definition.

Definition 4 (Increasing chain of sets)**.** *Let $A_0 = \emptyset$ and $A_m = [m]$. We say that $A_0 \subseteq A_1 \subseteq A_2 \subseteq \cdots \subseteq A_m$ is an increasing chain of sets if $|A_i| = i$ for all $i = 0, 1, 2, \dots, m$.*

A main step in our argument consist in proving the following two theorems:

Theorem 4 (**Monotonicity of RM entropies on chains.**)**.** *For every BMS channel W, every $m > 0$ and every increasing chain of sets $\emptyset = A_0 \subseteq A_1 \subseteq A_2 \subseteq \cdots \subseteq A_m = [m]$, we have*

$$H_{A_0}^{(m)} \le H_{A_1}^{(m)} \le H_{A_2}^{(m)} \le \cdots \le H_{A_m}^{(m)}.$$

Theorem 5 (**RM polarization on chains.**)**.** *For every BMS channel W and every $\epsilon > 0$, there is a constant $D(\epsilon)$ (which is independent of m and W) such that for every $m > 0$ and every increasing chain of sets $\emptyset = A_0 \subseteq A_1 \subseteq A_2 \subseteq \cdots \subseteq A_m = [m]$,*

$$\left|\left\{i \in \{0, 1, \dots, m\} : \epsilon < H_{A_i}^{(m)} < 1 - \epsilon\right\}\right| \le D(\epsilon).$$

In order to prove these two theorems, we only need to show two results. We establish first an interlacing property:

Lemma 2 (**Interlacing of RM entropies**)**.**

$$H_{A_i}^{(m+1)} \le H_{A_i}^{(m)} \le H_{A_{i+1}}^{(m+1)} \quad \forall i \in \{0, 1, \dots, m\}. \quad (9)$$

Second, we prove a separation property of non-extremal entropies:

Lemma 3 (**Separation of RM entropies**)**.** *For any $\epsilon > 0$, there is $\delta(\epsilon) > 0$ such that for any increasing chain of sets and any $i \in \{0, 1, \dots, m\}$,*

$$H_{A_i}^{(m)} \in (\epsilon, 1 - \epsilon)$$

implies that

$$H_{A_i}^{(m)} - H_{A_i}^{(m+1)} > \delta(\epsilon) \text{ and } H_{A_{i+1}}^{(m+1)} - H_{A_i}^{(m)} > \delta(\epsilon). \quad (10)$$

It is clear that Theorem 4 follows immediately from (9); see Fig. 1 for an illustration. Now we prove Theorem 5 using (10) and Theorem 4. By (10) we know that as long as $H_{A_i}^{(m)} > \epsilon$ and $H_{A_{i+1}}^{(m)} < 1-\epsilon$, we have $H_{A_{i+1}}^{(m)} - H_{A_i}^{(m)} > 2\delta$; see Fig. 1 for an illustration. Let j be the smallest index such that $H_{A_j}^{(m)} > \epsilon$, and let j' be the largest index such that $H_{A_{j'}}^{(m)} < 1 - \epsilon$. Then

$$\left|\left\{i \in \{0, 1, \dots, m\} : \epsilon < H_{A_i}^{(m)} < 1 - \epsilon\right\}\right| = j' - j + 1.$$

Since $H_{A_i}^{(m)}$ increases with i, we have

$$H_{A_{j'}}^{(m)} - H_{A_j}^{(m)} = \sum_{i=j}^{j'-1} (H_{A_{i+1}}^{(m)} - H_{A_i}^{(m)}) > 2(j' - j)\delta.$$

Since $H_{A_{j'}}^{(m)} - H_{A_j}^{(m)}$ is upper bounded by 1, we have $j' - j < \frac{1}{2\delta}$. Therefore,

$$\left|\left\{i \in \{0, 1, \dots, m\} : \epsilon < H_{A_i}^{(m)} < 1 - \epsilon\right\}\right| < \frac{1}{2\delta} + 1.$$

Thus we have proved Theorem 5 with the choice of $D(\epsilon) = \frac{1}{2\delta(\epsilon)} + 1$.

Now we are left to explain how to prove (9)–(10). The proof is divided into two steps. First, we prove (9)–(10) for the special case of $A_{i+1} = A_i \cup \{m + 1\}$. Then we show that by the symmetry of RM codes, $H_{A_i \cup \{j\}}^{(m+1)} \ge H_{A_i \cup \{m+1\}}^{(m+1)}$ for any $j \in [m] \setminus A_i$; see Lemma 4. Below we focus on the explanation of the first part.

To prove (9)–(10) for the special case of $A_{i+1} = A_i \cup \{m+1\}$, we use the recursive structure of RM code, and connect it back to that of polar codes. However, (9) is *not* a polar code triplet of the kind $W^- \le W \le W^+$ since we are working with the RM code ordering. The good news is that (9) gives in fact a larger spread than the one occurring for triplets of polar codes. The rest of this section is dedicated to explaining the precise meaning of previous phrase. With this connection in mind, (9)–(10) will be derived from (5) and Lemma 1.

We now derive (9)–(10). For a given BMS channel W, we denote the channel mapping from $U^{(m)}$ to $Y^{(m,W)}$ as $W^{(m)}$. Let us divide $Y^{(m+1)} := (Y_{\boldsymbol{z}}^{(m+1)} : \boldsymbol{z} = (z_1, z_2, \dots, z_{m+1}) \in \{0, 1\}^{m+1})$ into two subvectors:

$$\begin{aligned} &Y_{\text{odd}}^{(m+1)} := \\ &(Y_{\boldsymbol{z}}^{(m+1)} : \boldsymbol{z} = (z_1, z_2, \dots, z_{m+1}) \in \{0,1\}^{m+1}, z_{m+1} = 1), \\ &Y_{\text{even}}^{(m+1)} := \\ &(Y_{\boldsymbol{z}}^{(m+1)} : \boldsymbol{z} = (z_1, z_2, \dots, z_{m+1}) \in \{0,1\}^{m+1}, z_{m+1} = 0). \end{aligned} \quad (11)$$

The main observation is that the conditional distribution of $Y_{\text{odd}}^{(m+1)}$ given $(U_A^{(m+1)} + U_{A \cup \{m+1\}}^{(m+1)}, A \subseteq [m])$ is exactly $W^{(m)}$, and so is the conditional distribution of $Y_{\text{even}}^{(m+1)}$ given $(U_A^{(m+1)}, A \subseteq [m])$. To see this, we also divide the channel input random vector $X^{(m+1)}$ into two subvectors $X_{\text{odd}}^{(m+1)}$ and $X_{\text{even}}^{(m+1)}$ in the same way. Clearly, the output random vector $Y_{\text{odd}}^{(m+1)}$ only depends on $X_{\text{odd}}^{(m+1)}$, and $Y_{\text{even}}^{(m+1)}$ only depends on $X_{\text{even}}^{(m+1)}$. By definition, the random vector $X^{(m+1)}$ is the evaluation vector of the random polynomial $\sum_{A \subseteq [m+1]} U_A^{(m+1)} Z_A$, where Z_A is the shorthand notation of the monomial $\prod_{i \in A} Z_i$. When $m + 1 \in A$, the monomial Z_A is equal to 0 on all coordinates in $X_{\text{even}}^{(m+1)}$. Therefore, on all coordinates in $X_{\text{even}}^{(m+1)}$, we have $\sum_{A \subseteq [m+1]} U_A^{(m+1)} Z_A = \sum_{A \subseteq [m]} U_A^{(m+1)} Z_A$. As a consequence, the mapping from $(U_A^{(m+1)}, A \subseteq [m])$ to $X_{\text{even}}^{(m+1)}$ is exactly the same as the mapping from $(U_A^{(m)}, A \subseteq [m])$ to $X^{(m)}$, and thus the conditional distribution of $Y_{\text{even}}^{(m+1)}$ given $(U_A^{(m+1)}, A \subseteq [m])$ is exactly $W^{(m)}$. Next observe that for all $A \subseteq [m]$, we always have $Z_A = Z_{A \cup \{m+1\}}$ on all coordinates in $X_{\text{odd}}^{(m+1)}$.

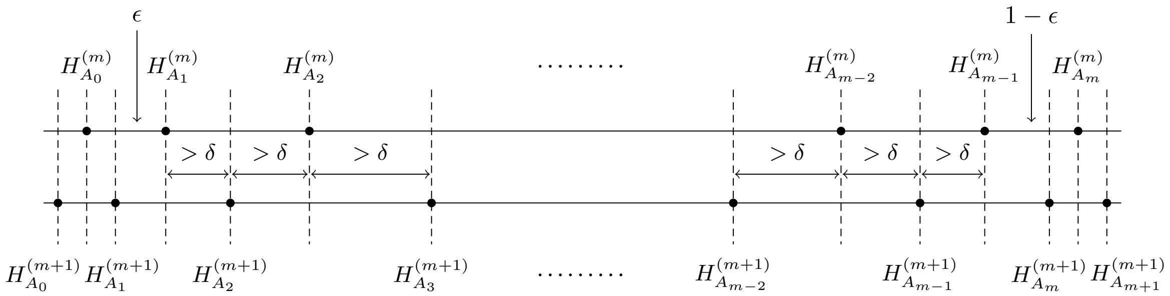

Fig. 1: Illustration of the interlacing property in (9) used in the proofs of Theorem 4 and Theorem 5.

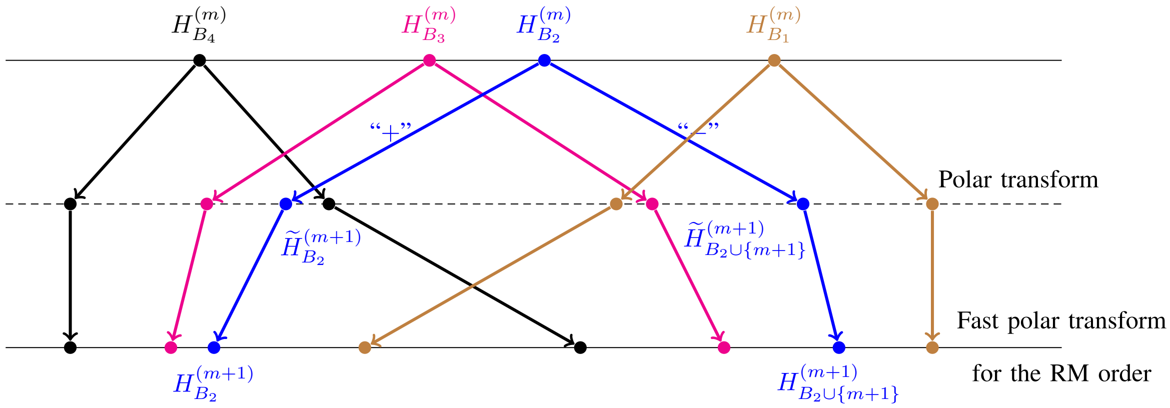

Fig. 2: The fast polar transform with block size 4. The dots on the second line are the results of the standard polar transform, and the dots on the third line are the results of fast polar transform. In the fast polar transform, the worse ("−") bit-channel in the standard polar transform gets even worse, and the better ("+") bit-channel in the standard polar transform gets even better. Therefore, the gap between $H_{B_i\cup\{m+1\}}^{(m+1)}$ and $H_{B_i}^{(m+1)}$ is always larger than the gap between $\widetilde{H}_{B_i\cup\{m+1\}}^{(m+1)}$ and $\widetilde{H}_{B_i}^{(m+1)}$. Intuitively, this explains why RM codes polarize and do so even faster than polar codes.

Therefore, on all coordinates in $X_{\text{odd}}^{(m+1)}$, we have

$$\begin{aligned}
&\sum_{A\subseteq[m+1]} U_A^{(m+1)} Z_A \\
&= \sum_{A\subseteq[m]} U_A^{(m+1)} Z_A + \sum_{A\subseteq[m]} U_{A\cup\{m+1\}}^{(m+1)} Z_{A\cup\{m+1\}} \\
&= \sum_{A\subseteq[m]} U_A^{(m+1)} Z_A + \sum_{A\subseteq[m]} U_{A\cup\{m+1\}}^{(m+1)} Z_A \\
&= \sum_{A\subseteq[m]} (U_A^{(m+1)} + U_{A\cup\{m+1\}}^{(m+1)}) Z_A .
\end{aligned}$$

As a consequence, the mapping from $(U_A^{(m+1)} + U_{A\cup\{m+1\}}^{(m+1)}, A \subseteq [m])$ to $X_{\text{odd}}^{(m+1)}$ is exactly the same as the mapping from $(U_A^{(m)}, A \subseteq [m])$ to $X^{(m)}$, and thus the conditional distribution of $Y_{\text{odd}}^{(m+1)}$ given $(U_A^{(m+1)} + U_{A\cup\{m+1\}}^{(m+1)}, A \subseteq [m])$ is exactly $W^{(m)}$.

The recursive structure between the bit-channels $\{W_A^{(m)} : A \subseteq [m]\}$ in the m-th level and the bit-channels $\{W_A^{(m+1)} : A \subseteq [m+1]\}$ in the $(m+1)$-th level can in fact be described as a polarization procedure. More specifically, The bit-channels $\{W_A^{(m)} : A \subseteq [m]\}$ are divided into $m+1$ layers according to the cardinality of the set A: The i-th layer is $\{W_A^{(m)} : A \subseteq [m], |A| = i\}$, the sets with cardinality i, for $i = 0, 1, 2, \dots, m$. Then we take two copies of each layer $\{W_A^{(m)} : A \subseteq [m], |A| = i\}$ and perform the fast polar transform, which we will discuss in more detail below. The outcome of the "−" fast polar transform is the bit-channels $\{W_{A\cup\{m+1\}}^{(m+1)} : A \subseteq [m], |A| = i\}$ in the next level, and the outcome of the "+" fast polar transform is the bit-channels $\{W_A^{(m+1)} : A \subseteq [m], |A| = i\}$. From the perspective of the bit-channels $\{W_A^{(m+1)} : A \subseteq [m+1]\}$ in the $(m+1)$-th level, except for the 0-th and the $(m+1)$-th layers, each layer $\{W_A^{(m+1)} : A \subseteq [m+1], |A| = i\}$ is divided into two parts: The first part is $\{W_A^{(m+1)} : A \subseteq [m], |A| = i\}$, which is the "+" fast polar transform of the i-th layer $\{W_A^{(m)} : A \subseteq [m], |A| = i\}$ in the m-th level. The second part is $\{W_{A\cup\{m+1\}}^{(m+1)} : A \subseteq [m], |A| = i-1\}$, which is the "−" fast polar transform of the $(i-1)$-th layer

$\{W_A^{(m)} : A \subseteq [m], |A| = i-1\}$ in the m-th level. As for the 0-th and the $(m+1)$-th layers, each of them only contains a single bit-channel $W_\emptyset^{(m+1)}$ and $W_{[m+1]}^{(m+1)}$, respectively, where $W_\emptyset^{(m+1)}$ is the "+" polar transform of $W_\emptyset^{(m)}$, and $W_{[m+1]}^{(m+1)}$ is the "$-$" polar transform of $W_{[m]}^{(m)}$.

Next we explain the fast polar transform. Let us consider the bit-channels $\{W_A^{(m+1)} : A \subseteq [m+1]\}$ in the $(m+1)$-th level. According to the total order defined in Definition 2, the layers from top to down are the $(m+1)$-th layer, the m-th layer, $\dots$, all the way down to the 0-th layer. Within each layer, all the sets containing the element $m+1$ appear after those not containing this element. Now let $B_1 < B_2 < \dots < B_j$ be all the sets in the t-th layer of $[m]$, i.e., they are all the subsets of $[m]$ with cardinality t. Then by the discussion above we know that the following sequence of subsets

$$\begin{aligned} &B_1 \cup \{m+1\} < B_2 \cup \{m+1\} < \dots < B_j \cup \{m+1\} \\ &< B_1 < B_2 < \dots < B_j \end{aligned} \tag{12}$$

are consecutive according to the total order on the subsets of $[m+1]$. By definition, if a successive decoder decodes according to (12), then the bit-channels seen by this decoder are equivalent to

$$\begin{aligned} &W_{B_1\cup\{m+1\}}^{(m+1)}, W_{B_2\cup\{m+1\}}^{(m+1)}, \dots, W_{B_j\cup\{m+1\}}^{(m+1)}, \\ &W_{B_1}^{(m+1)}, W_{B_2}^{(m+1)}, \dots, W_{B_j}^{(m+1)}. \end{aligned}$$

In order to connect fast polar transform to the standard polar transform, we consider the following order of the sets in (12):

$$B_1 \cup \{m+1\}, B_1, B_2 \cup \{m+1\}, B_2, \dots, B_j \cup \{m+1\}, B_j. \tag{13}$$

Assuming that we are still given $(U_{<B_1\cup\{m+1\}}^{(m+1)}, Y^{(m+1)})$, but this time the successive decoder decodes in this order instead of the order in (12). We denote the bit-channels seen by this successive decoder as

$$\begin{aligned} &\widetilde{W}_{B_1\cup\{m+1\}}^{(m+1)}, \widetilde{W}_{B_1}^{(m+1)}, \widetilde{W}_{B_2\cup\{m+1\}}^{(m+1)}, \widetilde{W}_{B_2}^{(m+1)}, \\ &\qquad \dots, \widetilde{W}_{B_j\cup\{m+1\}}^{(m+1)}, \widetilde{W}_{B_j}^{(m+1)}. \end{aligned}$$

It is easy to check that $\widetilde{W}_{B_i\cup\{m+1\}}^{(m+1)}$ and $\widetilde{W}_{B_i}^{(m+1)}$ are "$-$" and "+" polar transforms of $W_{B_i}^{(m)}$, respectively. Then by (5) and Lemma 1, we know that

$$\widetilde{H}_{B_i\cup\{m+1\}}^{(m+1)} \ge H_{B_i}^{(m)} \ge \widetilde{H}_{B_i}^{(m+1)}, \tag{14}$$

and if $H_{B_i}^{(m)} \in (\epsilon, 1-\epsilon)$, then

$$\widetilde{H}_{B_i\cup\{m+1\}}^{(m+1)} - H_{B_i}^{(m)} > \delta \quad \text{and} \quad H_{B_i}^{(m)} - \widetilde{H}_{B_i}^{(m+1)} > \delta, \tag{15}$$

where $\widetilde{H} = 1 - I(\widetilde{W})$. Comparing the order in (12) and (13), we can see that every set that appears before $B_i \cup \{m+1\}$ in (12) also appears before $B_i \cup \{m+1\}$ in (13). Therefore, for every $i \in \{1, 2, \dots, j\}$, we have

$$H_{B_i\cup\{m+1\}}^{(m+1)} \ge \widetilde{H}_{B_i\cup\{m+1\}}^{(m+1)} \quad \text{and} \quad \widetilde{H}_{B_i}^{(m+1)} \ge H_{B_i}^{(m+1)}. \tag{16}$$

In other words, in the standard polar transform, we obtain a worse bit-channel through the "$-$" transform and a better one through the "+" transform. Then in the fast polar transform, we make the bit-channel obtained through the standard "$-$" polar transform even worse and the bit-channel obtained through the standard "+" polar transform even better. Therefore, the gap between $H_{B_i\cup\{m+1\}}^{(m+1)}$ and $H_{B_i}^{(m)}$ is even larger than the gap between $\widetilde{H}_{B_i\cup\{m+1\}}^{(m+1)}$ and $H_{B_i}^{(m)}$. Similarly, the gap between $H_{B_i}^{(m)}$ and $H_{B_i}^{(m+1)}$ is even larger than the gap between $H_{B_i}^{(m)}$ and $\widetilde{H}_{B_i}^{(m+1)}$; see Fig. 2 for an illustration. Combining this with (14)–(15), we have shown that (9)–(10) hold for any $A_{i+1} = A_i \cup \{m+1\}$.

By now, we have explained how to prove Theorem 4 and Theorem 5. The next step is to use these two theorems to prove Theorem 1. To that end, we need the following strengthened form of Theorem 5.

Theorem 6 (**Strong RM polarization on chains**). *For every BMS channel W and every $0 < \epsilon < 0.1$, any $\delta_n = \mathrm{poly}(1/n)$ and $0 < \gamma < 1$, there is a constant $M(\epsilon, \delta_n, \gamma)$ such that for every $m > M(\epsilon, \delta_n, \gamma)$ and every increasing chain of sets $\emptyset = A_0 \subseteq A_1 \subseteq A_2 \subseteq \dots \subseteq A_m = [m]$,*

$$\begin{aligned} &\Big|\Big\{i \in \{0,1,\dots,m\} : H_{A_i}^{(m)} > 1-\epsilon\Big\} \cup \\ &\Big\{i \in \{0,1,\dots,m\} : Z_{A_i}^{(m)} < \delta_n\Big\}\Big| \ge m - m^\gamma. \end{aligned} \tag{17}$$

The proof of this theorem mainly relies on the fact that the Bhattacharyya parameter is close to 0 if and only if the conditional entropy is close to 0. More precisely, the proof relies on the following two well-known inequalities in the polar coding literature (see Proposition 1 of [11] for a proof):

$$Z(X|Y) \ge H(X|Y), \tag{18}$$

$$(1 - H(X|Y))^2 \le 1 - (Z(X|Y))^2. \tag{19}$$

We switch from the conditional entropy in Theorem 5 to the Bhattacharyya parameter in Theorem 6 and Theorem 1 for two reasons: First, the Bhattacharyya parameter $Z(X|Y)$ is an upper bound on the error probability of the MAP decoder of X given Y, i.e. (see [11]),

$$P_e(X|Y) \le Z(X|Y). \tag{20}$$

This property makes it convenient for us to prove that the twin-RM codes achieve capacity (Theorem 2). Second, in the "+" polar transform, the evolution of Bhattacharyya parameters follows a square law (7). As a result, it is easier to obtain a better bound on the Bhattacharyya parameters than on the conditional entropy.

Once we prove Theorem 6, we further use that there are $m!$ distinct increasing chains of sets for a given m. Let us fix m

and list all the $m!$ distinct increasing chains of sets as follows:

$$\begin{array}{c}\emptyset = A_0(1) \subseteq A_1(1) \subseteq A_2(1) \subseteq \cdots \subseteq A_m(1) = [m], \\ \emptyset = A_0(2) \subseteq A_1(2) \subseteq A_2(2) \subseteq \cdots \subseteq A_m(2) = [m], \\ \emptyset = A_0(3) \subseteq A_1(3) \subseteq A_2(3) \subseteq \cdots \subseteq A_m(3) = [m], \\ \vdots \quad \vdots \quad \vdots \quad \vdots \\ \emptyset = A_0(m!) \subseteq A_1(m!) \subseteq A_2(m!) \subseteq \cdots \subseteq A_m(m!) = [m].\end{array}$$

In Theorem 6, we have shown that among each increasing chain of sets, almost all the bit-channels becomes either completely noisy or noiseless. Let $\mathcal{A}$ be the collection of all the "bad" subsets of $[m]$, and let $\mathcal{S}$ be the collection of all the "bad" sets in the above $m!$ increasing chains (including multiplicity), where "bad" means that the set does not belong to the left-hand side of (17). Theorem 6 tells us that the "bad" sets in each chain is upper bounded by m^γ, so $|\mathcal{S}| \le m^\gamma m!$. On the other hand, notice that each subset with cardinality i appears $i!(m-i)!$ times in all the $m!$ increasing chains listed above, and that $i!(m-i)! \ge \lfloor m/2 \rfloor!(m - \lfloor m/2 \rfloor)!$ for all $i \in [m]$. Therefore, $|\mathcal{S}| \ge \lfloor m/2 \rfloor!(m - \lfloor m/2 \rfloor)!|\mathcal{A}|$. Combining the upper and lower bounds of $|\mathcal{S}|$, we obtain an upper bound on $|\mathcal{A}|$, and this proves the following more precise version of Theorem 1:

Theorem 7 (Polarization of RM codes). *For every BMS channel W, almost all elements in the set $\{H_A^{(m)}\}_{A \subseteq [m]}$ are close to either 0 or 1 when m is large. More precisely, for any $0 < \epsilon < 1/10$, any $\delta_n = \mathrm{poly}(1/n)$ and any $0 < \gamma < 1/2$, there is a constant $M(\epsilon, \delta_n, \gamma)$ such that for every $m > M(\epsilon, \delta_n, \gamma)$,*

$$\frac{\left|\left\{A \subseteq [m] : H_A^{(m)} > 1 - \epsilon\right\} \cup \left\{A \subseteq [m] : Z_A^{(m)} < \delta_n\right\}\right|}{2^m} \ge 1 - m^{\gamma - 1/2}.$$

Theorem 2 in turn follows directly from Theorem 7 (using (20)).

V. Open problems

Recently, Hassani et al. gave theoretical results showing that *if* RM codes achieve capacity, then they have an almost optimal scaling law over BSC channels under ML decoding [16], where optimal scaling law means that for a fixed linear code, the decoding error probability of ML decoder transitions from 0 to 1 as a function of the crossover probability of the BSC channel in the sharpest manner (i.e., comparable to random codes). In this paper, we have demonstrated that RM codes polarize faster than polar codes (see (16) and the discussion in Fig. 2) even though we stated our bound in Theorem 7 by exploiting the polar code bounds and therefore without the scaling-law improvement. An interesting direction would thus be to use directly the fast polarization of RM codes to prove that RM codes (or twin-RM codes) have a better scaling law than polar codes and/or an "optimal" scaling law.

Likewise, the scaling of δ_n that we obtain in Theorem 7 is stated now as polynomial due to the simplified proof, but we expect that an exponential scaling of $\exp(-n^{0.499})$ as obtained in [11], [12], [18] for polar codes should also be achievable.

Related to the above, the polarization framework has recently been generalized and strengthened due to the works of [18], [19]. It would be interesting to see if these can lead to further improvements of the bounds.

Also, this paper gives a second ordering of the matrix G_m that polarizes, i.e., the RM code ordering in addition to the polar code ordering (and the various other equivalent orderings that result from both of these). Are there many more[5] orderings that polarize? Is the RM code ordering "optimal" in some sense (e.g., for scaling-laws)?

Finally, we showed that the successive cancellation decoder in the RM code ordering achieves capacity on any BMS channel. While we exploited the recursive structure of RM codes in many parts of our proofs, we did not provide any computational bounds on the resulting successive decoding algorithm that exploits this structure. Can we also turn the successive decoding algorithm into an efficient one?

VI. Proofs

A. Two technical lemmas

We first need to establish some symmetry properties of RM codes and their impact on the conditional mutual information. Denote by S_m the symmetric group of order m. For $\pi \in S_m$ and $A \subseteq [m]$, define $\pi(A) := \{\pi(a) : a \in A\}$. Note that S_m is contained in the automorphism group of RM codes, as any degree $\le k$ polynomial is a degree $\le k$ polynomial under a relabelling of its variables.

Let $A \subseteq [m]$, and let $\mathcal{B}$ be a subset of the power set of $[m]$. For any $\pi \in S_m$, i.e., any relabelling of the elements of $[m]$, we have

$$\begin{aligned} &H(U_A^{(m)} | Y^{(m)}, \{U_B^{(m)} : B \in \mathcal{B}\}) \\ =&H(U_{\pi(A)}^{(m)} | Y^{(m)}, \{U_{\pi(B)}^{(m)} : B \in \mathcal{B}\}). \end{aligned} \tag{21}$$

This equality leads to the following lemma.

Lemma 4. *Let W be a BMS channel. Let $A \subset [m]$ and $i_1, i_2 \in [m]$ satisfy that $i_1, i_2 \notin A$ and $i_1 < i_2$. Then*

$$H_{A \cup \{i_1\}}^{(m)} \ge H_{A \cup \{i_2\}}^{(m)} \quad \text{and} \quad Z_{A \cup \{i_1\}}^{(m)} \ge Z_{A \cup \{i_2\}}^{(m)}.$$

Proof. Define $\pi \in S_m$ as

$$\pi(i) = i \text{ for all } i \ne i_1, i_2, \quad \pi(i_1) = i_2, \quad \pi(i_2) = i_1. \tag{23}$$

By (21), we have (22), where the inequality in (22) follows from the fact that

$$\begin{aligned} &\{\pi(B) : B \subseteq [m], B < (A \cup \{i_1\})\} \\ \subseteq&\{B : B \subseteq [m], B < (A \cup \{i_2\})\}. \end{aligned} \tag{24}$$

Indeed, if $B < (A \cup \{i_1\})$ and $i_1 \notin B$, then $\pi(B) \le B < (A \cup \{i_1\}) < (A \cup \{i_2\})$. If $B < (A \cup \{i_1\})$ and $i_1 \in B$, then $(B \setminus \{i_1\}) < A$, so

$$\begin{aligned} &\pi(B) = \pi((B \setminus \{i_1\}) \cup \{i_1\}) \\ =&\pi(B \setminus \{i_1\}) \cup \{i_2\} \le (B \setminus \{i_1\}) \cup \{i_2\} < A \cup \{i_2\}. \end{aligned}$$

[5] Clearly some ordering do not polarize, such as the down-to-top ordering in G_m.

$$\begin{aligned} H^{(m)}_{A\cup\{i_1\}} &= H\Big(U^{(m)}_{A\cup\{i_1\}}\Big|Y^{(m)}, \{U^{(m)}_B : B\subseteq[m], B<(A\cup\{i_1\})\}\Big) \\ &= H\Big(U^{(m)}_{\pi(A\cup\{i_1\})}\Big|Y^{(m)}, \{U^{(m)}_{\pi(B)} : B\subseteq[m], B<(A\cup\{i_1\})\}\Big) \\ &= H\Big(U^{(m)}_{A\cup\{i_2\}}\Big|Y^{(m)}, \{U^{(m)}_{\pi(B)} : B\subseteq[m], B<(A\cup\{i_1\})\}\Big) \\ &\ge H\Big(U^{(m)}_{A\cup\{i_2\}}\Big|Y^{(m)}, \{U^{(m)}_B : B\subseteq[m], B<(A\cup\{i_2\})\}\Big) = H^{(m)}_{A\cup\{i_2\}}, \end{aligned} \tag{22}$$

Therefore we have shown that $B < (A\cup\{i_1\})$ implies $\pi(B) < (A\cup\{i_2\})$, which is exactly the set containment in (24). This completes the proof of the lemma. Using Lemma 6 and the same reasoning as above, one can easily show that $Z^{(m)}_{A\cup\{i_1\}} \ge Z^{(m)}_{A\cup\{i_2\}}$. □

Lemma 5. *For every BMS channel W, every positive integer m, every $A\subseteq[m]$ and every $j\in[m+1]\setminus A$, we have the interlacing property:*

$$H^{(m+1)}_{A\cup\{j\}} \ge H^{(m)}_A \ge H^{(m+1)}_A, \tag{37}$$

$$Z^{(m+1)}_A \le \left(Z^{(m)}_A\right)^2, \qquad Z^{(m)}_A \le Z^{(m+1)}_{A\cup\{j\}}. \tag{38}$$

Moreover, for any $\epsilon>0$, there is $\delta(\epsilon)>0$ such that for any positive integer m, any $A\subseteq[m]$ and any $j\in[m+1]\setminus A$,

$$H^{(m)}_A \in (\epsilon, 1-\epsilon)$$

implies that

$$H^{(m)}_A - H^{(m+1)}_A > \delta(\epsilon) \text{ and } H^{(m+1)}_{A\cup\{j\}} - H^{(m)}_A > \delta(\epsilon). \tag{39}$$

Proof. Recall the definition of $Y^{(m+1)}_{\text{odd}}$ and $Y^{(m+1)}_{\text{even}}$ in (11). Let $y^{(m)} = (y^{(m)}_{\boldsymbol{z}} : \boldsymbol{z}\in\{0,1\}^m) \in \mathcal{Y}^n$ be a vector of length $n=2^m$ whose components take values in $\mathcal{Y}$, and this vector is indexed by $\boldsymbol{z}\in\{0,1\}^m$, which is similar to the random vector $Y^{(m)}$. Let $u^{(m)} = (u^{(m)}_A : A\subseteq[m]) \in \{0,1\}^n$ be a binary vector of length $n=2^m$, and this vector is indexed by $A\subseteq[m]$, which is similar to the random vector $U^{(m)}$. For $y^{(m)}\in\mathcal{Y}^n$, define the following three events:

$$\begin{aligned} \{Y^{(m)} = y^{(m)}\} &:= \{Y^{(m)}_{\boldsymbol{z}} = y^{(m)}_{\boldsymbol{z}} \text{ for all } \boldsymbol{z}\in\{0,1\}^m\}, \\ \{Y^{(m+1)}_{\text{odd}} = y^{(m)}\} &:= \{Y^{(m+1)}_{(\boldsymbol{z},1)} = y^{(m)}_{\boldsymbol{z}} \text{ for all } \boldsymbol{z}\in\{0,1\}^m\}, \\ \{Y^{(m+1)}_{\text{even}} = y^{(m)}\} &:= \{Y^{(m+1)}_{(\boldsymbol{z},0)} = y^{(m)}_{\boldsymbol{z}} \text{ for all } \boldsymbol{z}\in\{0,1\}^m\}, \end{aligned}$$

where for $\boldsymbol{z} = (z_1, z_2, \dots, z_m) \in \{0,1\}^m$, $(\boldsymbol{z},1) := (z_1, z_2, \dots, z_m, 1)$ and $(\boldsymbol{z},0) := (z_1, z_2, \dots, z_m, 0)$. According to the arguments below Equation (11), for any $y^{(m)}\in\mathcal{Y}^n$ and any $u^{(m)} = (u^{(m)}_A : A\subseteq[m]) \in \{0,1\}^n$,

$$\begin{aligned} &P\Big(\{Y^{(m+1)}_{\text{odd}} = y^{(m)}\}\Big|\{U^{(m+1)}_A + U^{(m+1)}_{A\cup\{m+1\}} = u^{(m)}_A \\ &\qquad\qquad \text{for all } A\subseteq[m]\}\Big) \\ =&P\Big(\{Y^{(m+1)}_{\text{even}} = y^{(m)}\}\Big|\{U^{(m+1)}_A = u^{(m)}_A \text{ for all } A\subseteq[m]\}\Big) \\ =&P\Big(\{Y^{(m)} = y^{(m)}\}\Big|\{U^{(m)}_A = u^{(m)}_A \text{ for all } A\subseteq[m]\}\Big). \end{aligned} \tag{40}$$

Since the two vectors $(U^{(m+1)}_A + U^{(m+1)}_{A\cup\{m+1\}} : A\subseteq[m])$ and $(U^{(m+1)}_A : A\subseteq[m])$ are independent, $(Y^{(m+1)}_{\text{odd}}, \{U^{(m+1)}_A + U^{(m+1)}_{A\cup\{m+1\}} : A\subseteq[m]\})$ and $(Y^{(m+1)}_{\text{even}}, \{U^{(m+1)}_A : A\subseteq[m]\})$ are also independent. By (40), we also obtain (25). Therefore, for any $A\subseteq[m]$, we have (26), where equality (a) in (26) holds because $(Y^{(m+1)}_{\text{odd}}, \{U^{(m+1)}_A + U^{(m+1)}_{A\cup\{m+1\}} : A\subseteq[m]\})$ and $(Y^{(m+1)}_{\text{even}}, \{U^{(m+1)}_A : A\subseteq[m]\})$ are independent. It is also clear that we have (27), so we obtain (28). According to the ordering of sets defined in Definition 2, it is easy to verify (29). Therefore, we have (30). Combining (30) with (28), we have $H^{(m+1)}_{A\cup\{m+1\}} \ge H^{(m)}_A \ge H^{(m+1)}_A$. Then by Lemma 4, for any $j\in[m+1]\setminus A$, we have $H^{(m+1)}_{A\cup\{j\}} \ge H^{(m+1)}_{A\cup\{m+1\}} \ge H^{(m)}_A \ge H^{(m+1)}_A$. This completes the proof of (37).

Next we prove (38). Let

$$\begin{aligned} &X_1 := U^{(m+1)}_A, \quad X_2 := U^{(m+1)}_A + U^{(m+1)}_{A\cup\{m+1\}}, \\ &Y_1 := \Big(Y^{(m+1)}_{\text{even}}, \{U^{(m+1)}_{A'} : A'\subseteq[m], A'<A\}\Big), \\ &Y_2 := \Big(Y^{(m+1)}_{\text{odd}}, \{U^{(m+1)}_{A'} + U^{(m+1)}_{A'\cup\{m+1\}} : A'\subseteq[m], A'<A\}\Big), \\ &X := U^{(m)}_A, \quad Y := \Big(Y^{(m)}, \{U^{(m)}_{A'} : A'\subseteq[m], A'<A\}\Big). \end{aligned}$$

Then (X_1, Y_1) and (X_2, Y_2) are i.i.d., and they have the same distribution as (X, Y). By (7) we have (31). According to (29) and Lemma 6, we obtain (32). By (8), we have (33). Combining (33) with (29) and Lemma 6, we obtain (34). By Lemma 4, for any $j\in[m+1]\setminus A$, we further have that $Z^{(m+1)}_{A\cup\{j\}} \ge Z^{(m+1)}_{A\cup\{m+1\}} \ge Z^{(m)}_A$. Combining this with (32), we complete the proof of (38).

Now we prove (39). For every $\epsilon>0$, we use the same $\delta(\epsilon)>0$ as in Lemma 1. We assume $H^{(m)}_A \in (\epsilon, 1-\epsilon)$ and use Lemma 1 to prove (39) under this assumption. Since (X_1, Y_1) and (X_2, Y_2) are i.i.d. with the same distribution as (X, Y), we have

$$H(X_1|Y_1) = H(X_2|Y_2) = H(X|Y) = H^{(m)}_A \in (\epsilon, 1-\epsilon).$$

According to Lemma 1,

$$H(X_1+X_2|Y_1, Y_2) \ge H^{(m)}_A + \delta(\epsilon). \tag{41}$$

We also have (35). Therefore by (26) and (41), we have (36). Combining (36) with (30) and Lemma 4, we conclude that for any $j\in[m+1]\setminus A$,

$$\begin{aligned} &H^{(m)}_A - H^{(m+1)}_A > \delta(\epsilon), \\ &H^{(m+1)}_{A\cup\{j\}} - H^{(m)}_A \ge H^{(m+1)}_{A\cup\{m+1\}} - H^{(m)}_A > \delta(\epsilon). \end{aligned}$$

$$
\begin{aligned}
H_A^{(m)} &= H\Big(U_A^{(m)}\Big|Y^{(m)}, U_{<A}^{(m)}\Big) = H\Big(U_A^{(m+1)}\Big|Y_{\text{even}}^{(m+1)}, \{U_{A'}^{(m+1)} : A' \subseteq [m], A' < A\}\Big) \\
&= H\Big(U_A^{(m+1)} + U_{A\cup\{m+1\}}^{(m+1)}\Big|Y_{\text{odd}}^{(m+1)}, \{U_{A'}^{(m+1)} + U_{A'\cup\{m+1\}}^{(m+1)} : A' \subseteq [m], A' < A\}\Big).
\end{aligned} \tag{25}
$$

$$
\begin{aligned}
&H\Big(U_A^{(m+1)}\Big|U_{A\cup\{m+1\}}^{(m+1)}, Y^{(m+1)}, \{U_{A'}^{(m+1)} : A' \subseteq [m], A' < A\}, \{U_{A'\cup\{m+1\}}^{(m+1)} : A' \subseteq [m], A' < A\}\Big) \\
&\quad + H\Big(U_{A\cup\{m+1\}}^{(m+1)}\Big|Y^{(m+1)}, \{U_{A'}^{(m+1)} : A' \subseteq [m], A' < A\}, \{U_{A'\cup\{m+1\}}^{(m+1)} : A' \subseteq [m], A' < A\}\Big) \\
&= H\Big(U_A^{(m+1)}, U_{A\cup\{m+1\}}^{(m+1)}\Big|Y^{(m+1)}, \{U_{A'}^{(m+1)} : A' \subseteq [m], A' < A\}, \{U_{A'\cup\{m+1\}}^{(m+1)} : A' \subseteq [m], A' < A\}\Big) \\
&= H\Big(U_A^{(m+1)}, U_A^{(m+1)} + U_{A\cup\{m+1\}}^{(m+1)}\Big|Y_{\text{even}}^{(m+1)}, \{U_{A'}^{(m+1)} : A' \subseteq [m], A' < A\}, \\
&\qquad\qquad\qquad\qquad\qquad\qquad Y_{\text{odd}}^{(m+1)}, \{U_{A'}^{(m+1)} + U_{A'\cup\{m+1\}}^{(m+1)} : A' \subseteq [m], A' < A\}\Big) \\
&\overset{(a)}{=} H\Big(U_A^{(m+1)}\Big|Y_{\text{even}}^{(m+1)}, \{U_{A'}^{(m+1)} : A' \subseteq [m], A' < A\}\Big) \\
&\quad + H\Big(U_A^{(m+1)} + U_{A\cup\{m+1\}}^{(m+1)}\Big|Y_{\text{odd}}^{(m+1)}, \{U_{A'}^{(m+1)} + U_{A'\cup\{m+1\}}^{(m+1)} : A' \subseteq [m], A' < A\}\Big) \\
&= 2H_A^{(m)},
\end{aligned} \tag{26}
$$

$$
\begin{aligned}
&H\Big(U_A^{(m+1)}\Big|U_{A\cup\{m+1\}}^{(m+1)}, Y^{(m+1)}, \{U_{A'}^{(m+1)} : A' \subseteq [m], A' < A\}, \{U_{A'\cup\{m+1\}}^{(m+1)} : A' \subseteq [m], A' < A\}\Big) \\
&\leq H\Big(U_A^{(m+1)}\Big|Y_{\text{even}}^{(m+1)}, \{U_{A'}^{(m+1)} : A' \subseteq [m], A' < A\}\Big) = H_A^{(m)},
\end{aligned} \tag{27}
$$

$$
\begin{aligned}
&H\Big(U_A^{(m+1)}\Big|U_{A\cup\{m+1\}}^{(m+1)}, Y^{(m+1)}, \{U_{A'}^{(m+1)} : A' \subseteq [m], A' < A\}, \{U_{A'\cup\{m+1\}}^{(m+1)} : A' \subseteq [m], A' < A\}\Big) \\
&\leq H_A^{(m)} \leq H\Big(U_{A\cup\{m+1\}}^{(m+1)}\Big|Y^{(m+1)}, \{U_{A'}^{(m+1)} : A' \subseteq [m], A' < A\}, \{U_{A'\cup\{m+1\}}^{(m+1)} : A' \subseteq [m], A' < A\}\Big).
\end{aligned} \tag{28}
$$

$$
\begin{aligned}
&\Big(\{A \cup \{m+1\}\} \cup \{A' : A' \subseteq [m], A' < A\} \cup \{A' \cup \{m+1\} : A' \subseteq [m], A' < A\}\Big) \\
&\subseteq \{A' : A' \subseteq [m+1], A' < A\}, \\
&\{A' : A' \subseteq [m+1], A' < (A \cup \{m+1\})\} \subseteq \{A' : A' \subseteq [m], A' < A\} \cup \{A' \cup \{m+1\} : A' \subseteq [m], A' < A\}.
\end{aligned} \tag{29}
$$

$$
\begin{aligned}
&H_A^{(m+1)} \leq H\Big(U_A^{(m+1)}\Big|U_{A\cup\{m+1\}}^{(m+1)}, Y^{(m+1)}, \{U_{A'}^{(m+1)} : A' \subseteq [m], A' < A\}, \{U_{A'\cup\{m+1\}}^{(m+1)} : A' \subseteq [m], A' < A\}\Big), \\
&H_{A\cup\{m+1\}}^{(m+1)} \geq H\Big(U_{A\cup\{m+1\}}^{(m+1)}\Big|Y^{(m+1)}, \{U_{A'}^{(m+1)} : A' \subseteq [m], A' < A\}, \{U_{A'\cup\{m+1\}}^{(m+1)} : A' \subseteq [m], A' < A\}\Big).
\end{aligned} \tag{30}
$$

$$
\begin{aligned}
&Z\left(U_A^{(m+1)}\Big|Y^{(m+1)}, \{U_{A'}^{(m+1)} : A' \subseteq [m], A' < A\}, \{U_{A'\cup\{m+1\}}^{(m+1)} : A' \subseteq [m], A' < A\}, U_{A\cup\{m+1\}}^{(m+1)}\right) \\
&= Z(X_1|Y_1, Y_2, X_1 + X_2) = (Z(X|Y))^2 = \left(Z_A^{(m)}\right)^2.
\end{aligned} \tag{31}
$$

$$
\begin{aligned}
&Z_A^{(m+1)} = Z(U_A^{(m+1)}|Y^{(m+1)}, U_{<A}^{(m+1)}) \\
&\leq Z(U_A^{(m+1)}|Y^{(m+1)}, \{U_{A'}^{(m+1)} : A' \subseteq [m], A' < A\}, \{U_{A'\cup\{m+1\}}^{(m+1)} : A' \subseteq [m], A' < A\}, U_{A\cup\{m+1\}}^{(m+1)}) = \left(Z_A^{(m)}\right)^2.
\end{aligned} \tag{32}
$$

$$
\begin{aligned}
&Z\left(U_{A\cup\{m+1\}}^{(m+1)}\Big|Y^{(m+1)}, \{U_{A'}^{(m+1)} : A' \subseteq [m], A' < A\}, \{U_{A'\cup\{m+1\}}^{(m+1)} : A' \subseteq [m], A' < A\}\right) \\
&= Z(X_1 + X_2|Y_1, Y_2) \geq Z(X|Y) = Z_A^{(m)}.
\end{aligned} \tag{33}
$$

$$\begin{aligned} &Z^{(m+1)}_{A\cup\{m+1\}} = Z(U^{(m+1)}_{A\cup\{m+1\}}|Y^{(m+1)}, U^{(m+1)}_{<(A\cup\{m+1\})}) \\ &\ge Z\left(U^{(m+1)}_{A\cup\{m+1\}}\middle|Y^{(m+1)}, \{U^{(m+1)}_{A'} : A'\subseteq[m], A'<A\}, \{U^{(m+1)}_{A'\cup\{m+1\}} : A'\subseteq[m], A'<A\}\right) \ge Z^{(m)}_A. \end{aligned} \tag{34}$$

$$\begin{aligned} &H(X_1+X_2|Y_1,Y_2) \\ =&H\left(U^{(m+1)}_{A\cup\{m+1\}}\middle|Y^{(m+1)}, \{U^{(m+1)}_{A'} : A'\subseteq[m], A'<A\}, \{U^{(m+1)}_{A'\cup\{m+1\}} : A'\subseteq[m], A'<A\}\right). \end{aligned} \tag{35}$$

$$\begin{aligned} &H\left(U^{(m+1)}_{A\cup\{m+1\}}\middle|Y^{(m+1)}, \{U^{(m+1)}_{A'} : A'\subseteq[m], A'<A\}, \{U^{(m+1)}_{A'\cup\{m+1\}} : A'\subseteq[m], A'<A\}\right) - H^{(m)}_A \\ =&H^{(m)}_A - H\left(U^{(m+1)}_{A}\middle|U^{(m+1)}_{A\cup\{m+1\}}, Y^{(m+1)}, \{U^{(m+1)}_{A'} : A'\subseteq[m], A'<A\}, \{U^{(m+1)}_{A'\cup\{m+1\}} : A'\subseteq[m], A'<A\}\right) \\ \ge&\delta(\epsilon). \end{aligned} \tag{36}$$

This completes the proof of the lemma. □

Lemma 6 (Lemma 1.8 in [34])**.** *Let (X, Y, Y') be a triple of discrete random variables, where X has Bernoulli-$1/2$ distribution. Then*

$$Z(X|Y, Y') \le Z(X|Y).$$

This lemma can be proved by a straightforward application of the Cauchy-Schwarz inequality.

B. Proof of Theorem 6

Without loss of generality, assume that $\delta_n = n^{-d}$ for some positive constant d. If $A \subseteq B \subseteq [m]$ and $|B| = |A| + 1$, then by (38),

$$Z^{(m)}_A \le Z^{(m+1)}_B \le \left(Z^{(m)}_B\right)^2. \tag{42}$$

For an increasing chain of sets $\emptyset = A_0 \subseteq A_1 \subseteq A_2 \subseteq \cdots \subseteq A_m = [m]$, (42) implies that

$$Z^{(m)}_{A_0} \le Z^{(m)}_{A_1} \le Z^{(m)}_{A_2} \cdots \le Z^{(m)}_{A_m}.$$

For a given $0 < \epsilon < 0.1$, define i_1 as the largest integer between 0 and m such that $H^{(m)}_{A_{i_1}} < \epsilon$, and define i_2 as the smallest integer between 0 and m such that $H^{(m)}_{A_{i_2}} > 1-\epsilon$. According to Theorem 4, $H^{(m)}_{A_i} < \epsilon$ for all $i \le i_1$ and $H^{(m)}_{A_i} > 1-\epsilon$ for all $i \ge i_2$. By Theorem 5, we know that

$$i_2 - i_1 - 1 \le D(\epsilon). \tag{43}$$

Since $H^{(m)}_{A_{i_1}} < \epsilon < 0.1$, by (19) we obtain that

$$Z^{(m)}_{A_{i_1}} < 1/2. \tag{44}$$

According to (42),

$$\log_2(Z^{(m)}_{A_i}) \le 2\log_2(Z^{(m)}_{A_{i+1}}),$$

and so

$$\log_2(Z^{(m)}_{A_i}) \le 2^j\log_2(Z^{(m)}_{A_{i+j}}).$$

For a given $0 < \gamma < 1$, define $i_3 := \lfloor i_1 - \frac{1}{2}m^\gamma\rfloor$. If $i_3 \ge 0$, then

$$\log_2(Z^{(m)}_{A_{i_3}}) \le 2^{m^\gamma/2}\log_2(Z^{(m)}_{A_{i_1}}) \le -2^{m^\gamma/2} \le -dm,$$

where the second inequality follows from (44) and the last inequality holds when m is large enough. Therefore, for all $i \le i_3$,

$$Z^{(m)}_{A_i} \le Z^{(m)}_{A_{i_3}} \le 2^{-dm} = n^{-d} = \delta_n.$$

Thus we have

$$\{0,1,\dots,i_3\} \subseteq \left\{i \in \{0,1,\dots,m\} : Z^{(m)}_{A_i} < \delta_n\right\},$$

$$\left\{i \in \{0,1,\dots,m\} : H^{(m)}_{A_i} > 1-\epsilon\right\} = \{i_2, i_2+1, \dots, m\}.$$

Combining this with (43), we obtain that

$$\begin{aligned} &\Big|\left\{i \in \{0,1,\dots,m\} : H^{(m)}_{A_i} > 1-\epsilon\right\}\cup \\ &\qquad\qquad \left\{i \in \{0,1,\dots,m\} : Z^{(m)}_{A_i} < \delta_n\right\}\Big| \\ &\ge i_3 + 1 + m - i_2 + 1 \ge i_1 - \frac{1}{2}m^\gamma + m - i_2 + 1 \\ &\ge m - \frac{1}{2}m^\gamma - D(\epsilon) \ge m - m^\gamma, \end{aligned}$$

where the last inequality holds when m is large enough.

On the other hand, if $i_3 < 0$, then $i_1 < \frac{1}{2}m^\gamma$, and by (43), $i_2 < \frac{1}{2}m^\gamma + D(\epsilon) + 1$. Therefore

$$\begin{aligned} &\Big|\left\{i \in \{0,1,\dots,m\} : H^{(m)}_{A_i} > 1-\epsilon\right\}\cup \\ &\qquad\qquad \left\{i \in \{0,1,\dots,m\} : Z^{(m)}_{A_i} < \delta_n\right\}\Big| \\ &\ge m - i_2 + 1 \ge m - \frac{1}{2}m^\gamma - D(\epsilon) \ge m - m^\gamma. \end{aligned}$$

This completes the proof of Theorem 6.

C. Proof of Theorem 7

We first observe that there is a one-to-one mapping between increasing chains of sets and permutations on $[m]$. Indeed, given $\pi \in S_m$, we can obtain an increasing chain of sets $\emptyset = A_0 \subseteq A_1 \subseteq A_2 \subseteq \cdots \subseteq A_m = [m]$ by setting $A_i = \{\pi(1), \pi(2), \dots, \pi(i)\}$ for all $i \in [m]$. On the other hand, given an increasing chain of sets $\emptyset = A_0 \subseteq A_1 \subseteq A_2 \subseteq \cdots \subseteq A_m = [m]$, we can obtain a permutation $\pi \in S_m$ by setting $\pi(i) = A_i \setminus A_{i-1}$ for all $i \in [m]$. Thus there are $m!$ distinct increasing chains of sets for a given m. Let us fix m and list all the $m!$ distinct increasing chains of sets as follows:

$$\begin{aligned}\emptyset = A_0(1) \subseteq A_1(1) \subseteq A_2(1) \subseteq \cdots \subseteq A_m(1) = [m],\\ \emptyset = A_0(2) \subseteq A_1(2) \subseteq A_2(2) \subseteq \cdots \subseteq A_m(2) = [m],\\ \emptyset = A_0(3) \subseteq A_1(3) \subseteq A_2(3) \subseteq \cdots \subseteq A_m(3) = [m],\\ \vdots \quad \vdots \quad \vdots \quad \vdots \qquad\qquad\\ \emptyset = A_0(m!) \subseteq A_1(m!) \subseteq A_2(m!) \subseteq \cdots \subseteq A_m(m!) = [m].\end{aligned}$$

Notice that for every $i \in \{0, 1, 2, \dots, m\}$, $|A_i(1)| = |A_i(2)| = |A_i(3)| = \cdots = |A_i(m!)| = i$. There are $\frac{m!}{i!(m-i)!}$ subsets of $[m]$ with cardinality i. By symmetry, each of them appears the same number of times in $(A_i(1), A_i(2), A_i(3), \dots, A_i(m!))$. Thus each subset with cardinality i appears $i!(m-i)!$ times in $(A_i(1), A_i(2), A_i(3), \dots, A_i(m!))$. In other words, each subset $A \subseteq [m]$ appears $|A|!(m-|A|)!$ times in $(A_i(j) : i \in \{0, 1, 2, \dots, m\}, j \in [m!])$.

For any $0 < \epsilon < 0.1$, define

$$\mathcal{S}(\epsilon) := \Big\{(i,j) : i \in \{0,1,2,\dots,m\}, j \in [m!], H^{(m)}_{A_i(j)} \le 1-\epsilon, Z^{(m)}_{A_i(j)} \ge \delta_n\Big\}.$$

Then by Theorem 6, we know that for any $0 < \gamma < 1/2$ and any given $j \in [m!]$,

$$\Big|\Big\{i \in \{0,1,\dots,m\} : H^{(m)}_{A_i(j)} \le 1-\epsilon, Z^{(m)}_{A_i(j)} \ge \delta_n\Big\}\Big| \le m^{\gamma} \quad \text{for all } m > M(\epsilon,\gamma).$$

Consequently, for all $m > M(\epsilon, \gamma)$,

$$|\mathcal{S}(\epsilon)| \le (m!)m^{\gamma}. \tag{45}$$

We further define

$$\mathcal{A}(\epsilon) := \Big\{A \subseteq [m] : H^{(m)}_A \le 1-\epsilon, Z^{(m)}_A \ge \delta_n\Big\}.$$

By the arguments above, we have

$$|\mathcal{S}(\epsilon)| = \sum_{A \subseteq \mathcal{A}(\epsilon)} |A|!(m-|A|)!.$$

It is easy to see that $i!(m-i)! \ge \lfloor m/2 \rfloor!(m - \lfloor m/2 \rfloor)!$ for all $i \in \{0, 1, 2, \dots, m\}$. Therefore

$$|\mathcal{S}(\epsilon)| \ge \lfloor m/2 \rfloor!(m - \lfloor m/2 \rfloor)!|\mathcal{A}(\epsilon)|.$$

Combining this with (45), we obtain that

$$|\mathcal{A}(\epsilon)| \le \binom{m}{\lfloor m/2 \rfloor} m^{\gamma}.$$

Consequently,

$$\frac{|\mathcal{A}(\epsilon)|}{2^m} \le m^{\gamma} \frac{\binom{m}{\lfloor m/2 \rfloor}}{2^m}.$$

By Stirling's formula,

$$\frac{\binom{m}{\lfloor m/2 \rfloor}}{2^m} = \sqrt{\frac{2}{\pi m}}(1 + o_m(1)). \tag{46}$$

Since $\sqrt{2/\pi} < 1$, we conclude that for all $m > M(\epsilon, \gamma)$,

$$\frac{|\mathcal{A}(\epsilon)|}{2^m} \le m^{\gamma - 1/2}.$$

This completes the proof of Theorem 7.

D. Proof of Theorem 2

We first show that the code rate of $\mathcal{T}(m, \delta_n)$ approaches the channel capacity $I(W)$, i.e.,

$$|\mathcal{G}(m, \delta_n)| \ge 2^m(I(W) - o(1)).$$

By (3), for all $0 < \epsilon < 1$, we have

$$(1-\epsilon)\Big|\Big\{A \subseteq [m] : H^{(m)}_A > 1-\epsilon\Big\}\Big| < \sum_{A \subseteq [m]} H^{(m)}_A = 2^m(1 - I(W)).$$

Therefore,

$$\Big|\Big\{A \subseteq [m] : H^{(m)}_A > 1-\epsilon\Big\}\Big| < \frac{1}{1-\epsilon}2^m(1 - I(W)).$$

According to Theorem 7, for $0 < \epsilon < 0.1$, $0 < \gamma < 1/2$ and $m > M(\epsilon, \gamma)$,

$$\begin{aligned}&|\mathcal{G}(m, \delta_n)| \ge\\ &\Big|\Big\{A \subseteq [m] : H^{(m)}_A > 1-\epsilon\Big\} \cup \Big\{A \subseteq [m] : Z^{(m)}_A < \delta_n\Big\}\Big|\\ &- \Big|\Big\{A \subseteq [m] : H^{(m)}_A > 1-\epsilon\Big\}\Big|\\ &\ge 2^m(1 - m^{\gamma-1/2}) - \frac{1}{1-\epsilon}2^m(1 - I(W)).\end{aligned}$$

The last line can be made arbitrarily close to $2^m I(W)$ if we set ϵ to be small enough and m to be large enough. Thus the code rate of $\mathcal{T}(m, \delta_n)$ approaches $I(W)$.

Next we prove that the decoding error of $\mathcal{T}(m, \delta_n)$ goes to 0 under the successive decoder that is similar to the one used for polar codes, i.e., we decode $U^{(m)}_A$ one by one using the channel outputs $Y^{(m)}$ and the previously decoded inputs $U^{(m)}_{<A}$. The decoding order is from small to large sets according to the order defined in Section I, i.e., we decode U_A before decoding U_B if $A < B$. According to (20), for every $A \in \mathcal{G}(m, \delta_n)$, the error probability of decoding $U^{(m)}_A$ from $Y^{(m)}$ and $U^{(m)}_{<A}$ is at most

$$P_e(U^{(m)}_A | Y^{(m)}, U^{(m)}_{<A}) \le Z(U^{(m)}_A | Y^{(m)}, U^{(m)}_{<A}) = Z^{(m)}_A < \delta_n.$$

By the union bound, the error probability of decoding the whole codeword under successive decoder is at most $n\delta_n \to 0$. Thus we conclude that the code $\mathcal{T}(m, \delta_n)$ achieves the capacity of W.

E. Proof of Theorem 3

Let $A \prec B$. Define $A' := \{a_1, \dots, a_{|B|}\}$ and note that by assumption, A' is pointwise smaller than B.

We first apply (37) repeatedly to obtain

$$H^{(m)}_A \ge H^{(m)}_{\{a_1,\dots,a_{|A|-1}\}} \ge H^{(m)}_{\{a_1,\dots,a_{|A|-2}\}} \ge \cdots \ge H^{(m)}_{A'}.$$

We then apply Lemma 4 repeatedly to obtain

$$\begin{aligned}H^{(m)}_{A'} = H^{(m)}_{\{a_1,\dots,a_{|B|-1},a_{|B|}\}} &\ge H^{(m)}_{\{a_1,\dots,a_{|B|-1},b_{|B|}\}}\\ &\ge H^{(m)}_{\{a_1,\dots,b_{|B|-1},b_{|B|}\}} \ge \cdots \ge H^{(m)}_B.\end{aligned}$$

Therefore $H^{(m)}_A \ge H^{(m)}_{A'} \ge H^{(m)}_B$.

F. Twin-RM code is the same as RM code up to $n = 16$ for BSC

We show that the twin-RM code is the same as the RM code up to $n = 16$ for BSC. Our claim follows immediately from the following proposition:

Proposition 1. *For BSC channels and $m \le 4$, if two subsets $A, B \subseteq [m]$ satisfy that $|A| > |B|$, then $H_A^{(m)} \ge H_B^{(m)}$.*

Proof. The cases of $m \le 2$ are trivial, so we only prove the cases of $m = 3$ and $m = 4$. Let us start with $m = 3$. We only need to show that $H_{[3]}^{(3)} \ge H_{[2]}^{(3)}, H_{\{2,3\}}^{(3)} \ge H_{\{1\}}^{(3)}, H_{\{3\}}^{(3)} \ge H_{\emptyset}^{(3)}$. In Section IV, we already showed that $W_{[3]}^{(3)}$ and $W_{[2]}^{(3)}$ are the "$-$" and "$+$" polar transforms of $W_{[2]}^{(2)}$, respectively, and that $W_{\{3\}}^{(3)}$ and $W_{\emptyset}^{(3)}$ are the "$-$" and "$+$" polar transforms of $W_{\emptyset}^{(2)}$, respectively. Therefore, $H_{[3]}^{(3)} \ge H_{[2]}^{(3)}$ and $H_{\{3\}}^{(3)} \ge H_{\emptyset}^{(3)}$ follow immediately (and this extends to any dimension, i.e., the first and last transitions are always ordered due to polar codes). Now let us prove

$$H_{\{2,3\}}^{(3)} \ge H_{\{1\}}^{(3)} \tag{47}$$

using the equivalence between source and channel coding. Suppose that $X_1, X_2, \dots, X_8$ are i.i.d. Bernoulli-p random variables, where p is the crossover probability of the BSC channel. Let $Y_1 = \sum_{i=1}^{8} X_i, Y_2 = X_1 + X_2 + X_3 + X_4, Y_3 = X_1 + X_2 + X_5 + X_6, Y_4 = X_1 + X_3 + X_5 + X_7, Y_5 = X_1 + X_2$. Then (47) is equivalent to $H(Y_4|Y_1, Y_2, Y_3) \ge H(Y_5|Y_1, Y_2, Y_3, Y_4)$. Notice that both X_1 and X_2 appear in Y_1, Y_2, Y_3. Therefore,

$$\begin{aligned}
&H(Y_4|Y_1, Y_2, Y_3) \\
=&H(X_1 + X_3 + X_5 + X_7|Y_1, Y_2, Y_3) \\
=&H(X_2 + X_3 + X_5 + X_7|Y_1, Y_2, Y_3) \\
=&H(Y_4 + Y_5|Y_1, Y_2, Y_3) \ge H(Y_4 + Y_5|Y_1, Y_2, Y_3, Y_4) \\
=&H(Y_5|Y_1, Y_2, Y_3, Y_4).
\end{aligned}$$

This completes the proof of (47).

For the case of $m = 4$, we only need to show that $H_{[4]}^{(4)} \ge H_{[3]}^{(4)}, H_{\{2,3,4\}}^{(4)} \ge H_{\{1,2\}}^{(4)}, H_{\{3,4\}}^{(4)} \ge H_{\{1\}}^{(4)}, H_{\{4\}}^{(4)} \ge H_{\emptyset}^{(4)}$. In particular, $H_{[4]}^{(4)} \ge H_{[3]}^{(4)}$ and $H_{\{4\}}^{(4)} \ge H_{\emptyset}^{(4)}$ follow immediately from the discussions in Section IV, so we only need to show the other two inequalities. We still use the equivalence between source and channel coding. Suppose that $X_1, X_2, \dots, X_{16}$ are i.i.d. Bernoulli-p random variables, where p is the crossover probability of the BSC channel. Let

$$\begin{aligned}
&Y_1 = \sum_{i=1}^{16} X_i, \\
&Y_2 = X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8, \\
&Y_3 = X_1 + X_2 + X_3 + X_4 + X_9 + X_{10} + X_{11} + X_{12}, \\
&Y_4 = X_1 + X_2 + X_5 + X_6 + X_9 + X_{10} + X_{13} + X_{14}, \\
&Y_5 = X_1 + X_3 + X_5 + X_7 + X_9 + X_{11} + X_{13} + X_{15}, \\
&Y_6 = X_1 + X_2 + X_3 + X_4.
\end{aligned}$$

Then $H_{\{3,4\}}^{(4)} \ge H_{\{1\}}^{(4)}$ is equivalent to $H(Y_5|Y_1, Y_2, Y_3, Y_4) \ge H(Y_6|Y_1, Y_2, Y_3, Y_4, Y_5)$. Notice that both X_1 and X_2 appear in Y_1, Y_2, Y_3, Y_4. Therefore,

$$\begin{aligned}
&H(Y_5|Y_1, Y_2, Y_3, Y_4) \\
=&H(X_1 + X_3 + X_5 + X_7 + X_9 + X_{11} \\
&\qquad + X_{13} + X_{15}|Y_1, Y_2, Y_3, Y_4) \\
=&H(X_2 + X_3 + X_5 + X_7 + X_9 + X_{11} \\
&\qquad + X_{13} + X_{15}|Y_1, Y_2, Y_3, Y_4).
\end{aligned}$$

Similarly, both X_3 and X_4 appear in Y_1, Y_2, Y_3, and neither of them appears in Y_4. Therefore,

$$\begin{aligned}
&H(X_2 + X_3 + X_5 + X_7 + X_9 + X_{11} \\
&\qquad + X_{13} + X_{15}|Y_1, Y_2, Y_3, Y_4) \\
=&H(X_2 + X_4 + X_5 + X_7 + X_9 + X_{11} \\
&\qquad + X_{13} + X_{15}|Y_1, Y_2, Y_3, Y_4).
\end{aligned}$$

Thus we conclude that

$$\begin{aligned}
&H(Y_5|Y_1, Y_2, Y_3, Y_4) \\
=&H(X_2 + X_4 + X_5 + X_7 + X_9 + X_{11} \\
&\qquad + X_{13} + X_{15}|Y_1, Y_2, Y_3, Y_4) \\
=&H(Y_5 + Y_6|Y_1, Y_2, Y_3, Y_4) \\
\ge&H(Y_5 + Y_6|Y_1, Y_2, Y_3, Y_4, Y_5) \\
=&H(Y_6|Y_1, Y_2, Y_3, Y_4, Y_5).
\end{aligned}$$

This completes the proof of $H_{\{3,4\}}^{(4)} \ge H_{\{1\}}^{(4)}$. $H_{\{2,3,4\}}^{(4)} \ge H_{\{1,2\}}^{(4)}$ can be proved in the same way, and we omit its proof here. □

References

[1] D. E. Muller, "Application of boolean algebra to switching circuit design and to error detection," *Transactions of the IRE professional group on electronic computers*, no. 3, pp. 6–12, 1954.

[2] I. Reed, "A class of multiple-error-correcting codes and the decoding scheme," *Transactions of the IRE Professional Group on Information Theory*, vol. 4, no. 4, pp. 38–49, 1954.

[3] W. Gasarch, "A survey on private information retrieval," *Bulletin of the EATCS*, vol. 82, no. 72-107, p. 113, 2004.

[4] A. Bogdanov and E. Viola, "Pseudorandom bits for polynomials," *SIAM Journal on Computing*, vol. 39, no. 6, pp. 2464–2486, 2010.

[5] A. Bhattacharyya, S. Kopparty, G. Schoenebeck, M. Sudan, and D. Zuckerman, "Optimal testing of Reed-Muller codes," in *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. IEEE, 2010, pp. 488–497.

[6] B. Barak, P. Gopalan, J. Håstad, R. Meka, P. Raghavendra, and D. Steurer, "Making the long code shorter," in *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*. IEEE, 2012, pp. 370–379.

[7] S. Kudekar, S. Kumar, M. Mondelli, H. D. Pfister, E. Şaşolu, and R. Urbanke, "Reed–Muller codes achieve capacity on erasure channels," *IEEE Transactions on Information Theory*, vol. 63, no. 7, pp. 4298–4316, 2017.

[8] S. Kudekar, S. Kumar, M. Mondelli, H. D. Pfister, E. Şaşoğlu, and R. Urbanke, "Reed-Muller codes achieve capacity on erasure channels," in *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*. ACM, 2016, pp. 658–669.

[9] E. Abbe, A. Shpilka, and A. Wigderson, "Reed–Muller codes for random erasures and errors," *IEEE Transactions on Information Theory*, vol. 61, no. 10, pp. 5229–5252, 2015.

[10] O. Sberlo and A. Shpilka, "On the performance of Reed-Muller codes with respect to random errors and erasures," *arXiv:1811.12447*, 2018.

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

[12] E. Arıkan and E. Telatar, "On the rate of channel polarization," in *2009 IEEE International Symposium on Information Theory*. IEEE, 2009, pp. 1493–1495.

[13] I. Tal and A. Vardy, "How to construct polar codes," *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 6562–6582, 2013.

[14] S. H. Hassani, S. B. Korada, and R. Urbanke, "The compound capacity of polar codes," in *47th Annual Allerton Conference on Communication, Control, and Computing*. IEEE, 2009, pp. 16–21.

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

[16] H. Hassani, S. Kudekar, O. Ordentlich, Y. Polyanskiy, and R. Urbanke, "Almost optimal scaling of Reed-Muller codes on BEC and BSC channels," in *2018 IEEE International Symposium on Information Theory (ISIT)*, June 2018, pp. 311–315.

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

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

[19] J. Błasiok, V. Guruswami, P. Nakkiran, A. Rudra, and M. Sudan, "General strong polarization," in *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, 2018, pp. 485–492.

[20] I. Tal and A. Vardy, "List decoding of polar codes," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2213–2226, 2015.

[21] "Final report of 3GPP TSG RAN WG1 #87 v1.0.0," http://www.3gpp.org/ftp/tsg_ran/WG1_RL1/TSGR1_87/Report/.

[22] I. Dumer, "Recursive decoding and its performance for low-rate Reed-Muller codes," *IEEE Transactions on Information Theory*, vol. 50, no. 5, pp. 811–823, 2004.

[23] R. Saptharishi, A. Shpilka, and B. L. Volk, "Efficiently decoding Reed–Muller codes from random errors," *IEEE Transactions on Information Theory*, vol. 63, no. 4, pp. 1954–1960, 2017.

[24] E. Santi, C. Häger, and H. D. Pfister, "Decoding Reed-Muller codes using minimum-weight parity checks," 2018, arXiv:1804.10319.

[25] M. Ye and E. Abbe, "Recursive projection-aggregation decoding of Reed-Muller codes," 2019, arXiv:1902.01470.

[26] J. Kahn, G. Kalai, and N. Linial, "The influence of variables on boolean functions," in *Proceedings of the 29th Annual Symposium on Foundations of Computer Science*. IEEE Computer Society, 1988, pp. 68–80.

[27] M. Talagrand, "On Russo's approximate zero-one law," *The Annals of Probability*, vol. 22, no. 3, pp. 1576–1587, 1994.

[28] E. Friedgut and G. Kalai, "Every monotone graph property has a sharp threshold," *Proceedings of the American mathematical Society*, vol. 124, no. 10, pp. 2993–3002, 1996.

[29] J. Bourgain, J. Kahn, G. Kalai, Y. Katznelson, and N. Linial, "The influence of variables in product spaces," *Israel Journal of Mathematics*, vol. 77, no. 1-2, pp. 55–64, 1992.

[30] Y. Wigderson, "Algebraic properties of tensor product matrices, with applications to coding," *Senior thesis, Princeton University*, 2016.

[31] E. Abbe and Y. Wigderson, "High-girth matrices and polarization," in *2015 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2015, pp. 2461–2465.

[32] M. Bardet, V. Dragoi, A. Otmani, and J. P. Tillich, "Algebraic properties of polar codes from a new polynomial formalism," in *2016 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2016, pp. 230–234.

[33] F. J. MacWilliams and N. J. A. Sloane, *The theory of error-correcting codes*. Elsevier, 1977.

[34] S. B. Korada, "Polar codes for channel and source coding," Ph.D. dissertation, ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE, 2009.