Linear Runlength-Limited Subcodes of Reed-Muller Codes and Coding Schemes for Input-Constrained BMS Channels

V. Arvind Rameshwar and Navin Kashyap

2022 IEEE Information Theory Workshop (ITW) | 978-1-6654-8341-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/ITW54588.2022.9965911

***Abstract*—In this work, we address the question of the largest rate of linear subcodes of Reed-Muller (RM) codes, all of whose codewords respect a runlength-limited (RLL) constraint. Our interest is in the (d,∞)-RLL constraint, which mandates that every pair of successive 1s be separated by at least d 0s. Consider any sequence $\{C_m\}_{m\geq 1}$ of RM codes with increasing blocklength, whose rates approach R, in the limit as the blocklength goes to infinity. We show that for any linear (d,∞)-RLL subcode, $\hat{C}_m$, of the code C_m, it holds that the rate of $\hat{C}_m$ is at most $\frac{R}{d+1}$, in the limit as the blocklength goes to infinity. We also consider scenarios where the coordinates of the RM codes are not ordered according to the standard lexicographic ordering, and derive rate upper bounds for linear (d,∞)-RLL subcodes, in those cases as well. Next, for the setting of a (d,∞)-RLL input-constrained binary memoryless symmetric (BMS) channel, we devise a new coding scheme, based on cosets of RM codes. Again, in the limit of blocklength going to infinity, this code outperforms any linear subcode of an RM code, in terms of rate, for low noise regimes of the channel.**

I. Introduction

Constrained coding is a method of eliminating error-prone sequences, in magnetic recording and communication systems, by encoding arbitrary user data sequences into sequences that respect a constraint (see, for example, [1] or [2]). In this work, we investigate the sizes of linear subcodes of binary Reed-Muller (RM) codes, all of whose codewords obey a certain runlength-limited (RLL) constraint.

The specific RLL constraint of interest to us is the (d,∞)-RLL constraint, which admits only binary sequences with at least d 0s between every pair of successive 1s (see Figure 1). This constraint is a special case of the (d,k)-RLL constraint, which admits only binary sequences with at least d and at most k 0s between successive 1s.

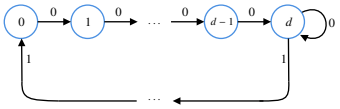

Fig. 1: The state transition graph for the (d,∞)-RLL constraint.

One of the motivations for studying this problem is the design of explicit coding schemes that achieve good rates over input-constrained discrete memoryless channels (DMCs). Figure 2 shows a generic binary memoryless symmetric (BMS) channel with input constraints. Input-constrained DMCs in general fall under the broad class of discrete finite-state channels (DFSCs, or FSCs).

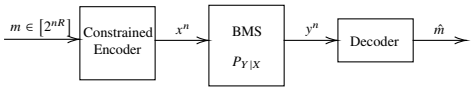

Fig. 2: System model of an input-constrained binary memoryless symmetric (BMS) channel without feedback.

While explicit codes achieving the capacities or whose rates are very close to the capacities of unconstrained DMCs have been derived in works such as [3]–[7], the problem of designing coding schemes for input-constrained DMCs has not received much attention in the literature. Moreover, an explicit expression for the capacity of an FSC is unknown, unlike the case of the unconstrained DMC, whose capacity is characterized by Shannon's formula, $C_{\mathrm{DMC}} = \sup_{P(x)} I(X;Y)$.

With the recent result of Reeves and Pfister [8] that Reed-Muller (RM) codes achieve the capacity of the unconstrained BMS channel under bit-MAP decoding, there opens the possibility of using such algebraic codes over input-constrained BMS channels as well. Suppose that C is the capacity of the unconstrained channel. In this paper, we prove that any linear RM subcode that respects the (d,∞)-RLL constraint, must have a rate of at most $\frac{C}{d+1}$, in the limit as the blocklength goes to infinity. In doing so, we show that one cannot do better, asymptotically, than the simple coding scheme using subcodes, in [9], if one requires that the subcodes be linear. We also consider the rates achieved using linear (d,∞)-RLL subcodes of permuted RM codes, and show that for codes of large enough blocklength, almost all permutations must respect an upper bound of $\frac{C}{d+1}+\delta$, for δ being as small as is required.

As an improvement over the rates achievable using linear (d,∞)-RLL subcodes of RM codes, we propose a new coding scheme that uses cosets of RM codes. The rate achieved by this scheme is $\frac{C_0 \cdot C^2 \cdot 2^{-\lceil \log_2(d+1) \rceil}}{C^2 \cdot 2^{-\lceil \log_2(d+1) \rceil} + 1 - C + \epsilon}$, where C_0 is the noiseless capacity of the input constraint, and $\epsilon > 0$ can be taken to be as small as is required. For example, when $d = 1$, the rates achieved using this cosets-based scheme are better than those achieved by any scheme that uses linear $(1,\infty)$-RLL subcodes of RM codes, when $C \gtrapprox 0.7613$. Moreover, as the capacity of the channel approaches 1, i.e., as the channel noise

The authors are with the Department of Electrical Communication Engineering, Indian Institute of Science, Bengaluru 560012. Email: {vrameshwar, nkashyap}@iisc.ac.in

The work of V. A. Rameshwar was supported by a Prime Minister's Research Fellowship, from the Ministry of Education, Govt. of India.

approaches 0, the rate achieved by our cosets-based scheme approaches a value arbitrarily close to C_0, which is the largest rate achievable, at zero noise, given the constraint.

Our results supplement the analysis in [10], on rates achievable by (d,k)-RLL subcodes of cosets of a linear block code. Specifically, Corollary 1 of [10] provides an existence result on cosets of capacity-achieving (over the unconstrained BMS channel) codes, whose constrained subcodes have rate at least $C_0 + C - 1$. The coding scheme in this paper achieves rates close to the lower bound in [10], for values of C close to 1.

The remainder of the paper is organized as follows: Section II introduces the notation and provides the necessary background. Section III states our main results. Section IV discusses upper bounds on the rate achievable over the BMS channel, using linear (d,∞)-RLL subcodes. Section V then discusses a construction that uses cosets of RM codes to achieve good rates. Finally, Section VI contains concluding remarks and a discussion on possible future work.

II. Notation and Preliminaries

A. Notation

Random variables will be denoted by capital letters, and their realizations by lower-case letters, e.g., X and x, respectively. Calligraphic letters, e.g., $\mathcal{X}$, denote sets. The notation $[n]$ denotes the set, $\{1,2,\ldots,n\}$, of integers, and the notation $[a:b]$, for $a<b$, denotes the set of integers $\{a,a+1,\ldots,b\}$. Moreover, for a real number x, we use $\lfloor x \rfloor$ to denote the largest integer smaller than or equal to x. For vectors $\mathbf{w}$ and $\mathbf{v}$ of length n and m, respectively, we denote their concatenation by the $(m+n)$-length vector, $\mathbf{wv}$. The notation x^N denotes the vector $(x_1,\ldots,x_N)$.

Throughout, we use the convenient notation $\binom{m}{\leq r}$ to denote the summation $\sum_{i=0}^{r}\binom{m}{i}$, and the notation $\binom{m}{\geq r}$ to denote $\sum_{i=r}^{m}\binom{m}{i}$.

B. Reed-Muller Codes

We recall the definition of the binary Reed-Muller (RM) family of codes. Codewords of binary RM codes consist of the evaluation vectors of multivariate polynomials over the binary field $\mathbb{F}_2$. Consider the polynomial ring $\mathbb{F}_2[x_1,x_2,\ldots,x_m]$ in m variables. Note that in the specification of a polynomial $f \in \mathbb{F}_2[x_1,x_2,\ldots,x_m]$, only monomials of the form $\prod_{j\in S} x_j$, for some $S \subseteq [m]$, need to be considered, since $x^2 = x$ over the field $\mathbb{F}_2$, for an indeterminate x. For a polynomial $f \in \mathbb{F}_2[x_1,x_2,\ldots,x_m]$ and a binary vector $\mathbf{z} = (z_1,\ldots,z_m) \in \mathbb{F}_2^m$, let $\mathrm{Eval}_{\mathbf{z}}(f) := f(z_1,\ldots,z_m)$. We let the evaluation points be ordered according to the standard lexicographic order on strings in $\mathbb{F}_2^m$, i.e., if $\mathbf{z} = (z_1,\ldots,z_m)$ and $\mathbf{z}' = (z_1',\ldots,z_m')$ are two distinct evaluation points, then, $\mathbf{z}$ occurs before $\mathbf{z}'$ in our ordering if and only if, for some $i \geq 1$, it holds that $z_j = z_j'$ for all $j < i$, and $z_i < z_i'$. Now, let $\mathrm{Eval}(f) := \left(\mathrm{Eval}_{\mathbf{z}}(f) : \mathbf{z} \in \mathbb{F}_2^m\right)$ be the evaluation vector of f, where the coordinates $\mathbf{z}$ are ordered according to the standard lexicographic order.

Definition II.1 (see [11], Chap. 13, or [12])**.** The r^{th} order binary Reed-Muller code $\mathrm{RM}(m,r)$ is defined as the set of binary vectors:

$$\mathrm{RM}(m,r) := \left\{\mathrm{Eval}(f) : f \in \mathbb{F}_2[x_1,x_2,\ldots,x_m],\ \deg(f) \leq r\right\},$$

where $\deg(f)$ is the degree of the largest monomial in f, and the degree of a monomial $\prod_{j\in S} x_j$ is simply $|S|$.

It is well-known that $\mathrm{RM}(m,r)$ has dimension $\binom{m}{\leq r}$ and minimum Hamming distance 2^{m-r}. The weight of a codeword $\mathbf{c} = \mathrm{Eval}(f)$ is the number of 1s in its evaluation vector, i.e,

$$\mathrm{wt}\left(\mathrm{Eval}(f)\right) := \left|\{\mathbf{z} \in \mathbb{F}_2^m : f(\mathbf{z}) = 1\}\right|.$$

In what follows, we let $G_{\mathrm{Lex}}(m,r)$ be the generator matrix of $\mathrm{RM}(m,r)$ consisting of rows that are the evaluations, in the lexicographic order, of monomials of degree less than or equal to r. The columns of $G_{\mathrm{Lex}}(m,r)$ will be indexed by m-tuples $\mathbf{b} = (b_1,\ldots,b_m)$ in the lexicographic order. We also interchangeably index the coordinates of any codeword in $\mathrm{RM}(m,r)$, by m-tuples in the lexicographic order, and by integers in $[0:2^m-1]$.

C. Codes for BMS Channels

The communication setting of an input-constrained binary memoryless symmetric (BMS) channel without feedback is shown in Figure 2. A message M is drawn uniformly from the set $\{1,2,\ldots,2^{nR}\}$, and is made available to the constrained encoder. The encoder produces a binary input sequence $x^n \in \{0,1\}^n = \mathcal{X}^n$, which is constrained to obey the (d,∞)-RLL input constraint, a state transition graph for which is shown in Figure 1.

The channel output alphabet is the extended real line, i.e., $\mathcal{Y} = \overline{\mathbb{R}}$. The channel is memoryless in that $P(y_i|x^i,y^{i-1}) = P(y_i|x_i)$, for all i. Further, the channel is symmetric, in that $P(y|1) = P(-y|0)$, for all $y \in \mathcal{Y}$. Common examples of BMS channels include the binary erasure channel (BEC(ϵ)), the binary symmetric channel (BSC) (see Figures 3a and 3b), and the binary additive white Gaussian noise (BI-AWGN) channel.

Fig. 3: (a) The BEC(ϵ) with erasure probability ϵ and output alphabet $\mathcal{Y} = \{-1,0,1\}$, with the output symbol 0 denoting an erasure. (b) The BSC(p) with crossover probability p and output alphabet $\mathcal{Y} = \{-1,1\}$.

Definition II.2. An $(n, 2^{nR}, (d,\infty))$ code for an input-constrained channel without feedback is defined by the encoding function: $f : \{1,\ldots,2^{nR}\} \to \mathcal{X}^n$, such that $(x_{i+1},\ldots,x_{\min\{i+d,n\}}) = (0,\ldots,0)$, if $x_i = 1$.

Given an output sequence y^n, the bit-MAP decoder $\Psi : \mathcal{Y}^n \to \mathcal{X}^n$ outputs $\hat{\mathbf{x}} := (\hat{x}_1,\ldots,\hat{x}_n)$, where, for each $i \in [n]$, the estimate $\hat{x}_i := \mathrm{argmax}_{x\in\{0,1\}} P(X_i = x|y^n)$. The error under bit-MAP decoding is defined as $P_b^{(n)} := 1 - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\max\{P(X_i=0|Y^n), P(X_i=1|Y^n)\}]$.

A rate R is said to be (d,∞)-achievable under bit-MAP decoding, if there exists a sequence of $(n, 2^{nR_n}, (d,\infty))$ codes, $\{C^{(n)}(R)\}_{n\geq 1}$, such that $\lim_{n\to\infty} P_b^{(n)} = 0$ and $\lim_{n\to\infty} R_n = R$. We then say that the sequence of codes $\{C^{(n)}(R)\}_{n\geq 1}$ *achieves* a rate R over the (d,∞)-RLL input-constrained

channel. The capacity, $C_{(d,\infty)}$, is defined to be the supremum over the respective (d,∞)-achievable rates, and is a function of the parameters of the noise process. Note that the definitions also hold when $d = 0$, which represents the unconstrained channel.

III. Main Results

Before we state our upper bound on the rates of linear RLL subcodes of RM codes, we recall the following result of Reeves and Pfister, in [8]. For a given $R \in (0,1)$, consider any sequence of RM codes $\{C_m(R) = \mathrm{RM}(m, r_m)\}_{m\geq 1}$, under the lexicographic ordering of coordinates, with $\mathrm{rate}(C_m(R)) \xrightarrow{m\to\infty} R$. The following theorem then holds true:

Theorem III.1 (Theorem 1 of [8])**.** *Consider an unconstrained BMS channel with capacity $C \in (0,1)$. Then, any rate $R \in (0, C)$ is achieved by the sequence of codes $\{C_m(R)\}_{m\geq 1}$, under bit-MAP decoding.*

We now discuss upper bounds on the largest rate achievable, using linear subcodes of RM codes, over a (d,∞)-RLL input-constrained BMS channel. Fix any sequence of codes $\{C_m(R) = \mathrm{RM}(m, r_m)\}_{m\geq 1}$, which achieves a rate R over the unconstrained BMS channel. Let $\overline{C}_d^{(m)}$ denote the largest *linear* subcode of $C_m(R)$, all of whose codewords respect the (d,∞)-RLL constraint. We then define $\mathsf{R}_{C,\mathrm{Lin}}^{(d,\infty)}(R) := \limsup_{m\to\infty} \frac{\log_2\left|\overline{C}_d^{(m)}\right|}{2^m}$, to be the largest rate achieved by linear (d,∞)-RLL subcodes of $\{C_m(R)\}_{m\geq 1}$, assuming that the ordering of the coordinates of the code is according to the lexicographic ordering. Then,

Theorem III.2. *For any sequence of codes $\{C_m(R) = RM(m, r_m)\}_{m\geq 1}$, with $rate(C_m(R)) \xrightarrow{m\to\infty} R$, it holds that $\mathsf{R}_{C,Lin}^{(d,\infty)}(R) \leq \frac{R}{d+1}$.*

Hence, from Theorem III.1, the largest rate achievable over a (d,∞)-RLL input-constrained BMS channel, under bit-MAP decoding, using linear (d,∞)-RLL subcodes of RM codes, is bounded above by $\frac{C}{d+1}$. Theorem III.2 is proved in Section IV.

Now, consider the sequence of RM codes $\{\hat{C}_m(R) = \mathrm{RM}(m, v_m)\}_{m\geq 1}$, with $v_m = \max\left\{\left\lfloor \frac{m}{2} + \frac{\sqrt{m}}{2} Q^{-1}(1-R)\right\rfloor, 0\right\}$, where $Q(\cdot)$ is the complementary cumulative distribution function (c.c.d.f.) of the standard normal distribution. Then,

Theorem III.3 (Theorem III.2 in [9])**.** *For any $R \in (0, C)$, there exists a sequence of linear codes, $\{C_m^{(d,\infty)}(R)\}_{m\geq 1}$, where $C_m^{(d,\infty)}(R) \subset \hat{C}_m(R)$, which achieve a rate of $\frac{R}{2^{\lceil \log_2(d+1)\rceil}}$, over a (d,∞)-RLL input-constrained BMS channel, under bit-MAP decoding.*

Thus, Theorem III.2 shows that the sequence of linear subcodes $\{C_m^{(d,\infty)}(R)\}_{m\geq 1}$, in Theorem III.3, achieves the rate upper bound of $R/(d+1)$, when $d+1$ is a power of 2. We remark here that the results in [13] show that the largest linear code within the set of (d,∞)-RLL constrained sequences of length m, has rate no larger than $\frac{1}{d+1}$, as $m \to \infty$. However, such a result offers no insight into rates achievable over BMS channels.

We then consider situations where the coordinates of the RM codes follow orderings different from the standard lexicographic ordering. We consider arbitrary orderings of coordinates, defined by the sequence of permutations $(\pi_m)_{m\geq 1}$, with $\pi_m : [0 : 2^m - 1] \to [0 : 2^m - 1]$. We define the sequence of π-ordered RM codes $\{C_m^{\pi}(R)\}_{m\geq 1}$, with $C_m^{\pi}(R) := \left\{(c_{\pi_m(0)}, c_{\pi_m(2)}, \ldots, c_{\pi_m(N_m-1)}) \;:\; (c_0, c_1, \ldots, c_{N_m-1}) \in C_m(R)\right\}$. We also define $\overline{C}_{d,\pi}^{(m)}$ be the largest *linear* (d,∞)-RLL subcode of $C_m^{\pi}(R)$. The theorem below is then shown to hold:

Theorem III.4. *For large m and for all but a vanishing fraction of coordinate permutations, $\pi_m : [0 : 2^m - 1] \to [0 : 2^m - 1]$, the rate upper bound, $\frac{\log_2\left|\overline{C}_{d,\pi}^{(m)}\right|}{2^m} \leq \frac{R}{d+1} + \delta_m$, holds, where $\delta_m \xrightarrow{m\to\infty} 0$.*

We refer the reader to the full version of the paper [14] for the proof of Theorem III.4.

Next, we turn our attention to the design of non-linear (d,∞)-RLL codes, whose rates improve on those in Theorem III.3. Our next theorem, stated below informally, uses cosets of RM codes, for this purpose. We denote by $C_0^{(d)}$, the noiseless capacity of the (d,∞)-RLL constraint.

Theorem III.5 (Informal)**.** *For any BMS channel of capacity C, there exists a sequence of (d,∞)-RLL constrained codes $\{C_m^{cos}\}_{m\geq 1}$, using cosets of RM codes, such that*

$$\liminf_{m\to\infty} rate(C_m^{cos}) \geq \frac{C_0^{(d)} \cdot C^2 \cdot 2^{-\lceil \log_2(d+1)\rceil}}{C^2 \cdot 2^{-\lceil \log_2(d+1)\rceil} + 1 - C + 2^{-\tau}},$$

with the above bound being achievable over any (d,∞)-RLL input-constrained BMS channel. Here, τ is an arbitrarily large, but fixed, positive integer.

It can be checked that the rates achieved using Theorem III.5 are better than those achieved using any sequence of linear (d,∞)-RLL subcodes of RM codes (see Theorem III.2), for low noise regimes of the BMS channel. For example, when $d = 1$, the rates achieved using the codes in Theorem III.5 are better than those achieved using linear subcodes, for certain values of $C \gtrapprox 0.7613$. Figures 4a and 4b show comparisons between the lower bounds (achievable rates) in Theorems III.3 and III.5, with the coset-averaging bound of [10], for $d = 1$ and $d = 2$, respectively. Our construction is more explicit than that in [10], although the rates calculated in [10] are better than those in Theorem III.5 in the low noise regimes of the BMS channel. A sketch of the construction leading to Theorem III.5 is provided in Section V.

IV. Upper Bounds for Linear Subcodes

In this section, we derive upper bounds on the rates achieved by linear (d,∞)-RLL subcodes of any sequence of RM codes of rate R. We fix a sequence of codes $\{C_m(R) = \mathrm{RM}(m, r_m)\}_{m\geq 1}$ that achieves a rate R over the unconstrained BMS channel.

We first state a fairly general proposition, whose proof can be found in [14], on the rates of linear (d,∞)-RLL subcodes of linear codes. Recall that for a linear code $\overline{C}$ over $\mathbb{F}_2$, of blocklength N and dimension K, an information set is a

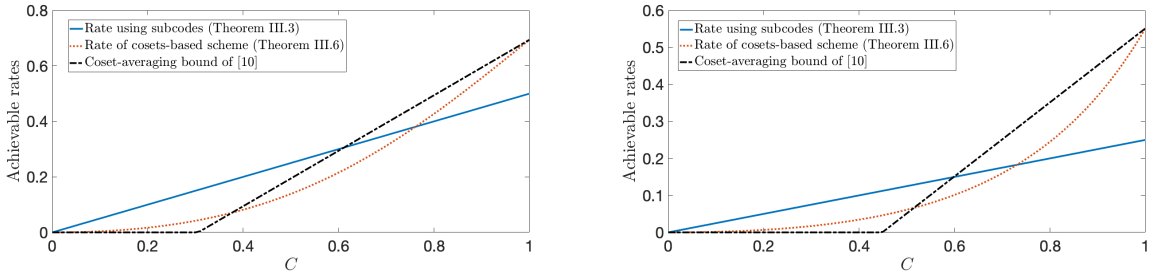

Fig. 4: Plots (a) and (b) compare, for $d = 1$ and $d = 2$, respectively, the rate lower bound achieved using subcodes, from Theorem III.3, the rate lower bound achieved using Theorem III.5, with $\tau = 50$, and the lower bound of $\max(0, C_0^{(d)} + C - 1)$, of [10].

collection of K coordinates in which all possible K-tuples over $\mathbb{F}_2$ can appear. Equivalently, if G is any generator matrix for $\overline{C}$, an information set is a set of K column indices such that G restricted to those columns is a full-rank matrix.

Proposition IV.1. *Let $\overline{C}$ be an $[N, K]$ binary linear code. If $\mathcal{I}$ is an information set of $\overline{C}$ that contains t disjoint $(d+1)$-tuples of consecutive coordinates $(i_1, i_1+1, \ldots, i_1+d), (i_2, i_2+1, \ldots, i_2+d), \ldots, (i_t, i_t+1, \ldots, i_t+d)$, with $i_1 \geq 1$, $i_j > i_{j-1}+d$, for all $j \in [2:t]$, and $i_t \leq n-d$, then the dimension of any linear (d, ∞)-RLL subcode of $\overline{C}$ is at most $K - dt$.*

In order to obtain an upper bound, as in Theorem III.2, on the rate of linear (d, ∞)-RLL subcodes of the sequence of codes $\{C_m(R)\}_{m \geq 1}$, we shall first identify an information set $\mathcal{I}_{m,r_m}$ of $C_m(R) = \mathrm{RM}(m, r_m)$. We then compute the number of disjoint $(d+1)$-tuples of consecutive coordinates in $\mathcal{I}_{m,r_m}$, and then apply Proposition IV.1.

We introduce some notation for ease of reading: given a matrix $M_{p \times q}$, we use the notation $M[\mathcal{U}, \mathcal{V}]$ to denote the submatrix of M consisting of the rows in $\mathcal{U} \subseteq [p]$ and the columns in $\mathcal{V} \subseteq [q]$. We also recall the definition of the generator matrix $G_{\mathrm{Lex}}(m, r)$, of $\mathrm{RM}(m, r)$, and the indexing of columns of the matrix, from Section II-B. Further, the notation $\mathbf{e}_{\mathbf{b}}^{(2^m)}$ denotes the standard basis vector with a 1 in the coordinate indexed by $\mathbf{b} = (b_1, \ldots, b_m)$, in the lexicographic order. The superscript '(2^m)' will be dropped henceforth.

Now, given the code $\mathrm{RM}(m, r)$, consider the binary linear code (a subspace of $\mathbb{F}_2^{2^m}$), $\tilde{C}(m, r)$, spanned by the codewords in the set $\mathcal{B}_{m,r} := \{\mathrm{Eval}\left(\prod_{i \in S} x_i\right) : S \subseteq [m] \text{ with } |S| \geq r+1\}$. It can be checked that $\mathcal{B}_{m,r}$ forms a basis for $\tilde{C}(m, r)$, with $\dim\left(\tilde{C}(m, r)\right) = \binom{m}{\geq r+1}$.

The following lemma, whose proof is provided in [14], identifies an alternative basis for $\tilde{C}(m, r)$.

Lemma IV.1. *Consider the code $\tilde{C}(m, r) = \mathrm{span}\left(\mathcal{B}_{m,r}\right)$. It holds that $\tilde{C}(m, r) = \mathrm{span}\left(\{\mathbf{e}_{\mathbf{b}} : wt(\mathbf{b}) \geq r+1\}\right)$.*

The following result then holds true:

Lemma IV.2. *An information set of $RM(m, r)$ is the set of coordinates $\mathcal{I}_{m,r} := \{\mathbf{b} = (b_1, \ldots, b_m) \in \mathbb{F}_2^m : wt(\mathbf{b}) \leq r\}$.*

Proof. We wish to show that $G_{\mathrm{Lex}}(m, r)$ restricted to the columns in $\mathcal{I}_{m,r}$ is of full rank.

Now, consider the generator matrix $\tilde{G}(m, r)$, of $\tilde{C}(m, r)$, consisting of rows that are vectors in $\mathcal{B}_{m,r}$. We build the $2^m \times 2^m$ matrix

$$\mathsf{H} := \left[\begin{array}{c} \tilde{G}(m, r) \\ \hline \\ G_{\mathrm{Lex}}(m, r) \end{array}\right],$$

with H being full rank. Note that, from Lemma IV.1, any standard basis vector $\mathbf{e}_{\mathbf{b}}$, with $\mathbf{b} \in \mathcal{I}_{m,r}^c$, belongs to rowspace$(\tilde{G}(m, r))$. From Lemma IV.1 and from the fact that H is full rank, it holds that $\mathsf{H}\left[\left[\binom{m}{\geq r+1}\right) + 1 : 2^m\right], \mathcal{I}_{m,r}\right]$ is full rank, or, $G_{\mathrm{Lex}}(m, r)$, restricted to columns in $\mathcal{I}_{m,r}$, is full rank. □

Now that we have identified an information set $\mathcal{I}_{m,r_m}$ of $C_m(R)$, we need only calculate the number of disjoint $(d+1)$-tuples of consecutive coordinates in $\mathcal{I}_{m,r_m}$. We introduce the notation $\mathbf{B}(i)$ to denote the length-m binary representation of i, for $0 \leq i \leq 2^m - 1$.

We shall first compute the number of runs of consecutive coordinates, in the lexicographic ordering, which belong to the information set $\mathcal{I}_{m,r_m}$. Formally, if we define

$$\Gamma_{m,r_m} := \{s : \mathbf{B}(s+1) \notin \mathcal{I}_{m,r_m}, \text{ and } \mathbf{B}(s-p), \ldots, \mathbf{B}(s) \in \mathcal{I}_{m,r_m}, \text{ for some } p \geq 0\}, \quad (1)$$

to be the set of starting coordinates of runs that belong to $\mathcal{I}_{m,r_m}$, then the required number of runs is $\left|\Gamma_{m,r_m}\right|$.

Lemma IV.3. *Under the lexicographic ordering, it holds that $\left|\Gamma_{m,r}\right| = \binom{m-1}{r}$, for $0 \leq r \leq m-1$.*

We refer the reader to [14] for the proof of Lemma IV.3. With the ingredients in place, we are now in a position to prove Theorem III.2.

Proof of Theorem III.2. We work with the sequence of codes $\{C_m(R)\}_{m \geq 1}$, with $r_m \leq m-1$, for all m. We use the notation $K_m := \binom{m}{\leq r_m}$ to denote the dimension of $C_m(R)$.

For a given m, we know from Lemma IV.3 that the number of runs under the lexicographic ordering, $\left|\Gamma_{m,r_m}\right|$, of coordinates that lie in the information set $\mathcal{I}_{m,r_m}$, is exactly $\binom{m-1}{r_m}$. Now, note that the i^{th} run $(s_i, \ldots, s_i + \ell_i)$, of length ℓ_i,

with $s_i \in \Gamma_{m,r_m}$ and $i \in \left[\left|\Gamma_{m,r_m}\right|\right]$, contributes $\left\lfloor \frac{\ell_i}{d+1} \right\rfloor$ disjoint $(d+1)$-tuples of consecutive coordinates in $\mathcal{I}_{m,r}$. It then holds that the overall number of disjoint $(d+1)$-tuples of consecutive coordinates in $\mathcal{I}_{m,r}$ is t_m, where

$$t_m \geq \sum_{i=1}^{|\Gamma_{m,r_m}|} \left(\frac{\ell_i}{d+1} - 1 \right) = \frac{K_m}{d+1} - \left|\Gamma_{m,r_m}\right| = \frac{K_m}{d+1} - \binom{m-1}{r_m},$$

where the last equality follows from Lemma IV.3.

Now, from Proposition IV.1, it holds that

$$\begin{aligned} \mathsf{R}_{C,\mathrm{Lin}}^{(d,\infty)}(R) &\leq \limsup_{m\to\infty} \frac{K_m - \frac{dK_m}{d+1} + d \cdot \binom{m-1}{r_m}}{2^m} \\ &\leq \lim_{m\to\infty} \frac{\frac{K_m}{d+1} + d \cdot \binom{m-1}{\lfloor \frac{m-1}{2} \rfloor}}{2^m} = \frac{R}{d+1}, \end{aligned}$$

where the last equality holds from the fact that $\binom{m-1}{\lfloor \frac{m-1}{2} \rfloor} \sim c \cdot \frac{2^m}{\sqrt{m-1}}$ (see. for example, equation (5.28) in [15], where '$\sim$' is used to mean "grows as"), and $\lim_{m\to\infty} \frac{K_m}{2^m} = R$. □

V. Achievable Rates Using Cosets Of RM Codes

The results of the previous sections provide bounds on achievable rates by using subcodes of RM codes. In this section, we provide a sketch of another construction, which uses cosets of RM codes. The rates achieved by this construction, under bit-MAP decoding, are better than those in Theorem III.3, for low noise regimes of the BMS channel.

Fix a rate $R \in (0, C)$ and any sequence $\{C_m(R) = \mathrm{RM}(m, r_m)\}_{m\geq 1}$ that achieves a rate R over the unconstrained BMS channel, under bit-MAP decoding. Recall, from Lemma IV.2, that the set $\mathcal{I}_{m,r_m} := \{\mathbf{b} = (b_1, \ldots, b_m) \in \mathbb{F}_2^m : \mathrm{wt}(\mathbf{b}) \leq r_m\}$ is an information set of $C_m(R)$. For the remainder of this section, we let m be a large positive integer.

We set $K_m = \dim(C_m(R)) = \binom{m}{\leq r_m}$. Also, let $R_m^{(d,\infty)}$ be the rate of the code $C_m^{(d,\infty)}$ (see Theorem III.3, which is Theorem III.2 of [9]).

Consider any permutation $\pi_m : [0 : N_m - 1] \to [0 : N_m - 1]$ with the property that $\pi_m([0 : K_m - 1]) = \mathcal{I}_{m,v_m}$, where, for a permutation σ, and a set $\mathcal{A} \subseteq [0 : N_m - 1]$, we define the notation $\sigma(\mathcal{A}) := \{\sigma(i) : i \in \mathcal{A}\}$. As in Section III, we define the permuted code $C_m^\pi(R)$ as $C_m^\pi(R) = \left\{(c_{\pi_m(0)}, c_{\pi_m(1)}, \ldots, c_{\pi_m(N_m-1)}) : (c_0, c_1, \ldots, c_{N_m-1}) \in C_m(R)\right\}$.

Thus, $C_m^\pi(R)$ is the code obtained by permuting the coordinates of codewords in $C_m(R)$, such that the coordinates in the information set $\mathcal{I}_{m,r_m}$ occur in the first block of K_m positions. Note that the permuted code $C_m^\pi(R)$ is systematic, and hence all K_m-tuples that respect that (d,∞)-RLL constraint, occur in the first K_m coordinates. We let G_m^π be a *systematic* generator matrix for $C_m^\pi(R)$. For the lemma that follows, whose proof can be found in [14], we shall use the notation $\tilde{C}_m^\pi := \left\{(\tilde{c}_{\pi_m(0)}, \tilde{c}_{\pi_m(1)}, \ldots, \tilde{c}_{\pi_m(N_m-1)} : (\tilde{c}_0, \tilde{c}_1, \ldots, \tilde{c}_{N_m-1}) \in \tilde{C}(m, r_m)\right\}$, where $\tilde{C}(m, r_m) = \mathrm{span}\left(\mathcal{B}_{m,r_m}\right)$ (see Section IV).

Lemma V.1. *For every codeword* $\mathbf{w} \in C_m^\pi(R)$*, there exists a vector* $\mathbf{v} \in \tilde{C}_m^\pi$*, such that* $\mathbf{w} + \mathbf{v}$ *(over* $\mathbb{F}_2$*) equals the concatenation* $w_1^{K_m}\mathbf{0}$.

Remark. Note that words $\mathbf{v} \in \tilde{C}_m^\pi$, which are of the form $\mathbf{v} = 0^{K_m} v_{K_m+1}^{N_m}$, for some $v_{K_m+1}, \ldots, v_{N_m} \in \{0, 1\}$, are in one-to-one correspondence with the cosets of $C_m^\pi(R)$. In other words, each word in $\tilde{C}_m^\pi$ uniquely identifies a coset of $C_m^\pi(R)$. In what follows, we consider $\tilde{C}_m^\pi$ to be the collection of coset leaders for the code $C_m^\pi(R)$.

We now describe a simple coding scheme to transmit (d,∞)-RLL input-constrained words over the BMS channel:

1) Pick a (d,∞)-RLL constrained K_m-tuple, $w_1^{K_m}$. Encode $w_1^{K_m}$ into a codeword $\mathbf{c} \in C_m^\pi(R)$, using the systematic generator matrix, G_m^π, with $\mathbf{c} = w_1^{K_m} G_m^\pi$. Note that $c_1^{K_m} = w_1^{K_m}$.
2) Choose a coset leader $\mathbf{v} \in \tilde{C}_m^\pi$ such that the word, $\mathbf{c} + \mathbf{v} = w_1^{K_m}\mathbf{0}$, is also (d,∞)-RLL constrained.
3) Transmit the first K_m bits, $w_1^{K_m}$, of $\mathbf{c} + \mathbf{v}$.
4) Transmit the identity of the coset leader as follows:
   a) Divide $c_{K_m+1}^{N_m}$ into L equal parts, $\mathbf{c}_1, \ldots, \mathbf{c}_L$, where L is a suitably chosen, large positive integer.
   b) Encode each part $\mathbf{c}_i$, for $i \in [L]$, into a codeword of the code $C_n^{(d,\infty)}(R)$, of a carefully chosen blocklength $N_{\mathrm{part}} \geq \frac{N_m - K_m}{L \cdot R_m^{(d,\infty)}}$, where $n = \log_2 N_{\mathrm{part}}$.

Choosing an RLL constrained word in Step 1 above can be accomplished using well-known constrained encoders (see, for example, [16] and Chapters 4 and 5 of [1]), of rates arbitrarily close to the noiseless capacity, $C_0^{(d)}$, of the (d,∞)-RLL constraint. Further, Lemma V.1 shows that Step 2 can also be achieved. We refer the reader to [14] for a detailed explanation of Step 4, and, in particular, the choice of the integers L and N_{part}. At the decoder end, the coset leader $\mathbf{v}$ is recovered first, and this information is used to decode the original codeword, $\mathbf{c} \in C_m^\pi(R)$.

The coding scheme described above obeys the rate lower bound given in Theorem III.5, in the limit as $m \to \infty$. We refer the reader to [14] for the proof of the theorem.

VI. Conclusion

In this paper, we derived upper bounds on the rates of linear (d,∞)-RLL subcodes of Reed-Muller (RM) codes. We showed that if C is the capacity of an unconstrained BMS channel, then the rate of any linear (d,∞)-RLL subcode of an RM code, is bounded above by $\frac{C}{d+1}$, in the limit as the blocklength of the code goes to infinity. Besides, we showed that for large enough blocklength, for nearly all coordinate orderings, a rate upper bound of $\frac{C}{d+1} + \delta$ holds, where δ can be taken to be as small as required. Further, we devised a constrained coding scheme based on cosets of RM codes that, for low noise regimes, outperforms any linear coding scheme, in terms of rate.

For future work, as regards the cosets-based coding scheme proposed in this paper, other sequential decoding algorithms (such as those in [17]), adapted to RM codes, can be explored to check if the extra channel uses in our coding scheme can be eliminated altogether.

VII. Acknowledgements

The authors thank Prof. H. D. Pfister for stimulating discussions.

REFERENCES

[1] B. H. Marcus, R. M. Roth, and P. H. Siegel, "An introduction to coding for constrained systems," *Lecture notes*, 2001.

[2] K. A. S. Immink, P. H. Siegel, and J. K. Wolf, "Codes for digital recorders," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2260–2299, Oct. 1998.

[3] E. Arikan, "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.

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

[5] M. Luby, M. Mitzenmacher, M. Shokrollahi, and D. Spielman, "Efficient erasure correcting codes," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 569–584, 2001.

[6] T. Richardson, M. Shokrollahi, and R. Urbanke, "Design of capacity-approaching irregular low-density parity-check codes," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 619–637, 2001.

[7] S. Kudekar, T. Richardson, and R. L. Urbanke, "Spatially coupled ensembles universally achieve capacity under belief propagation," *IEEE Transactions on Information Theory*, vol. 59, no. 12, pp. 7761–7813, 2013.

[8] G. Reeves and H. D. Pfister, "Reed-Muller codes achieve capacity on BMS channels," *arXiv e-prints*, p. arXiv:2110.14631, Oct. 2021.

[9] V. A. Rameshwar and N. Kashyap, "On the Performance of Reed-Muller Codes Over (d, ∞)-RLL Input-Constrained BMS Channels," *accepted to the International Symposium on Information Theory (ISIT) 2022*, available [Online] at https://arxiv.org/abs/2201.02035.

[10] A. Patapoutian and P. Kumar, "The (d, k) subcode of a linear block code," *IEEE Transactions on Information Theory*, vol. 38, no. 4, pp. 1375–1382, 1992.

[11] F. MacWilliams and N. Sloane, *The Theory of Error-Correcting Codes*, 2nd ed. North-holland Publishing Company, 1978.

[12] E. Abbe, A. Shpilka, and M. Ye, "Reed–Muller codes: Theory and algorithms," *IEEE Transactions on Information Theory*, vol. 67, no. 6, pp. 3251–3277, 2021.

[13] G. Lechner, I. Land, and A. Grant, "Linear and non-linear run length limited codes," *IEEE Communications Letters*, vol. 19, no. 7, pp. 1085–1088, Jul. 2015.

[14] V. A. Rameshwar and N. Kashyap, "Linear Runlength-Limited Subcodes of Reed-Muller Codes and Coding Schemes for Input-Constrained BMS Channels," *arXiv e-prints*, p. arXiv:2205.04153, May 2022.

[15] J. Spencer, *Asymptopia*. American Mathematical Society, 2014.

[16] R. Adler, D. Coppersmith, and M. Hassner, "Algorithms for sliding block codes - an application of symbolic dynamics to information theory," *IEEE Transactions on Information Theory*, vol. 29, no. 1, pp. 5–22, 1983.

[17] J. Honda and H. Yamamoto, "Polar coding without alphabet extension for asymmetric models," *IEEE Transactions on Information Theory*, vol. 59, no. 12, pp. 7829–7838, 2013.