On Codes Decoding a Constant Fraction of Errors on the BSC

Jan Hązła
EPFL
Lausanne, Switzerland
jan.hazla@epfl.ch

Alex Samorodnitsky*
The Hebrew University of Jerusalem
Jerusalem, Israel
salex@cs.huji.ac.il

Ori Sberlo†
Tel-Aviv University
Tel-Aviv, Israel
orisberlo@mail.tau.ac.il

ABSTRACT

We strengthen the results from a recent work by the second author, achieving bounds on the weight distribution of binary linear codes that are successful under block-MAP (as well as bit-MAP) decoding on the BEC. We conclude that a linear code that is successful on the BEC can also decode over a range of binary memoryless symmetric (BMS) channels. In particular, applying the result of Kudekar, Kumar, Mondelli, Pfister, Şaşoğlu and Urbanke from STOC 2016, we prove that a Reed–Muller code of positive rate R decodes errors on the $\mathrm{BSC}(p)$ with high probability if $p < 1/2 - \sqrt{2^{-R}(1-2^{-R})}$.

CCS CONCEPTS

• **Mathematics of computing → Coding theory**; • **Theory of computation → Error-correcting codes**.

KEYWORDS

capacity-achieving codes, weight enumerator, Reed–Muller codes

ACM Reference Format:

Jan Hązła, Alex Samorodnitsky, and Ori Sberlo. 2021. On Codes Decoding a Constant Fraction of Errors on the BSC. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC '21), June 21–25, 2021, Virtual, Italy.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3406325.3451015

1 INTRODUCTION

In this work we study binary linear codes over binary memoryless symmetric channels and their weight distribution.

A binary linear error correcting code is a linear subspace $V \subseteq \mathbb{F}_2^n$. The subspace V should have the property that given an erroneous version of $v \in V$ one can extract from it some information on v (v is usually referred to as a codeword). This is possible because while $v \in \mathbb{F}_2^n$ is an n-bit vector, it belongs to a subspace smaller than the entire space, or equivalently because v contains redundancies. The amount of redundancy in V is captured by the ratio $R(V) = \frac{\dim(V)}{n}$ which is a fundamental property of a code called the *rate*.

In order to make the above concrete one has to formally define the manner in which errors are induced and what information on $v \in V$ should be retrieved. One such simple model was proposed by Hamming [11] in which corruptions are adversarial and we seek to recover the entire original codeword. The behaviour in this setting is completely determined by the *minimum distance* which is the Hamming distance between the two closest codewords. Other interesting models include list decoding, deletion channels, locally decodable codes and more. In this work we shall focus on Shannon's model in which corruptions are induced randomly and we seek to recover the original codeword with high probability. Perhaps the simplest types of corruption are the binary erasure channel (BEC) and the binary symmetric channel (BSC). In the $\mathrm{BEC}(p)$, every bit is independently erased[1] with fixed probability $p \in [0,1]$ and in the $\mathrm{BSC}(p)$ every bit is flipped with probability $p \in [0, 1/2]$. These two channels belong to a larger family of *binary memoryless symmetric channels* (BMS channels). Memoryless means that the noise is independent for every coordinate and symmetric roughly means that the corruption of 1 and 0 is symmetric. For instance, had we flipped 0 with probability p_0 and flipped 1 with probability p_1 for $p_0 \neq p_1$ then this would not have been a symmetric channel. Another important example of a BMS channel is the binary additive Gaussian white noise channel $\mathrm{BAWGN}(\sigma)$ where we add a Gaussian noise $z \sim \mathcal{N}(0, \sigma^2)$ to every coordinate independently.

In his seminal work [25], Shannon provided an upper bound on the amount of noise a code can tolerate over any BMS channel. He proved that a code $V \subseteq \mathbb{F}_2^n$ can be decoded successfully over a channel $\mathcal{W}$ only if its rate $R(V)$ does not exceed the *channel capacity* $C(\mathcal{W})$. For instance, $C(\mathrm{BEC}(p)) = 1 - p$ and so codes with rate larger than $1 - p$ cannot recover from p fraction of random erasures. Codes that achieve this bound are called capacity achieving (with respect to a channel $\mathcal{W}$). One may wonder if such codes even exist and indeed, random codes as well as random linear codes are capacity achieving. There are also explicit constructions of capacity achieving codes, most notably *polar codes* introduced by Arıkan [5] that achieve capacity for any BMS channel with efficient encoding and decoding.

It is often the case that good error correcting codes come from algebraic structures. Examples include the Reed–Muller codes and the BCH codes which are among the oldest binary codes (we shall soon discuss the Reed–Muller codes). Yet, none of these were known to achieve capacity even for the BEC. This was resolved by Kudekar, Kumar, Mondelli, Pfister, Şaşoğlu and Urbanke [14] who proved that a binary linear code which is doubly transitive (see Definition 3.10)

*Research partially supported by ISF grant 1724/15.
†The research leading to these results has received funding from the Israel Science Foundation (grant number 552/16) and from the Len Blavatnik and the Blavatnik Family foundation.

[1] By erasing a coordinate we mean to replace it with '?'. Do not confuse it with deletion in which the codeword length varies.

STOC '21, June 21–25, 2021, Virtual, Italy

ACM ISBN 978-1-4503-8053-9/21/06…$15.00
https://doi.org/10.1145/3406325.3451015

achieves capacity for the BEC. In particular, the aforementioned codes satisfy this symmetry property and hence are capacity achieving for the BEC. Unfortunately, their technique seems less amenable to other channels such as the BSC or the BAWGN. Our result extends theirs to arbitrary BMS channels, albeit not to the capacity limit, in the following way.

Theorem 1.1 (Informal)**.** *Let $V \subseteq \mathbb{F}_2^n$ be a doubly transitive code with rate $R = R(V)$, and minimum distance $d(V) = \Omega(n^\alpha)$. Then, V decodes errors on $\mathrm{BSC}(p(R,\alpha))$, $\mathrm{BAWGN}(\sigma(R,\alpha))$ where $p(R,\alpha)$, $\sigma(R,\alpha)$ are some explicit functions depending on α, R.*

This can be extended, in an appropriate way, to arbitrary BMS channels.

For details see Section 3.2. This is achieved by going through the *weight distribution* of a code which is the sequence enumerating how many codewords $v \in V$ there are with a given number of ones. For linear codes, this determines how far apart the codewords are and hence serves as a good statistic to measure how much noise the code can tolerate. Indeed, sufficiently strong bounds on the weight distribution of a linear code imply that it can decode errors on a given BMS channel.

Recently, bounds on the weight distribution of codes that achieve capacity for the BEC were given in [22]. By combining the two approaches of [14, 22] we derive slightly stronger bounds on the weight distribution of such codes which in turn implies good performance for general BMS channels.

In fact, we establish a general method that applies to any linear code that is good enough at recovering from random erasures. We give exact definitions later, but in the theorem below $P_{\mathrm{B}}(\mathcal{W}, V)$ is the probability of failure in recovering a codeword using the optimal (so-called maximum a posteriori, or block-MAP) decoder for a code V on a channel $\mathcal{W}$. The Bhattacharyya parameter $Z(\mathcal{W}) \in [0,1]$ is a property of a channel, with a lower value of $Z(\mathcal{W})$ intuitively corresponding to a less noisy channel:

Theorem 1.2. *Let V be a binary linear code with dimension $\dim(V) = k$, minimum distance d and block-MAP error probability over $\mathrm{BEC}(\lambda)$ less than $1/k$. Then, for any BMS channel $\mathcal{W}$,*

$$P_{\mathrm{B}}(\mathcal{W}, V) < 2\left(\frac{Z(\mathcal{W})}{2^\lambda - 1}\right)^d .$$

Therefore, a linear code that recovers from random erasures also decodes errors on BMS channels with small enough Bhattacharayya parameter. As discussed in Remark 3.5, our assumptions on the minimum distance and error probability over rgw BEC are mild. While below we discuss a specific application to Reed–Muller codes, we note that our result can be applied more generally, e.g., for BCH codes, LDPC codes or polar codes for the BEC. On the other hand, we stress that we are only concerned with MAP decoding, and in general there is no reason to expect that it is efficient.

Our result applies to an important family of binary codes — the Reed–Muller codes. Reed–Muller (RM) codes were introduced by Muller [17] and rediscovered shortly after by Reed [19]. The RM code $\mathrm{RM}(m,r)$ is defined[2] by all the evaluation vectors of multilinear polynomials over $\mathbb{F}_2$ with m variables and degree at most r.

[2]It is possible to generalize RM codes to arbitrary fields. This is sometimes referred to as generalized RM codes [7].

Due to [14], we know that constant rate RM codes achieve capacity for the BEC. What about the BSC? BAWGN? It is considered plausible

that RM codes achieve capacity for those channels as well [4]. However, prior to this work it was unknown whether a constant rate RM code can correct even a tiny constant fraction of random errors. In the regime of non-constant rate there are some strong results. For rates approaching 0, Abbe, Shpilka and Wigderson [3] proved that $\mathrm{RM}(m,r)$ achieve capacity[3] for the BSC if $r = o(m)$ which was later improved to $r = m/70$ in [24]. The latter also provides some results on the BSC for any $r < (1/2 - o(\sqrt{\log m/m}))m$. For rates approaching 1 the best result is again due to [3] who proved that the Reed–Muller code $\mathrm{RM}(m,r)$ achieves capacity for the BSC if $r > m - O(\sqrt{m/\log m})$. For constant rates, [2] shows that subcodes of constant rate RM codes where an arbitrarily small constant fraction of basis elements was deleted correct a constant fraction of random errors.

In contrast, applying Theorem 1.2 to the result from [14] we obtain an unconditional result for constant rate RM codes, albeit falling short of the capacity threshold:

Theorem 1.3. *For any $0 < R < 1$, a family of RM codes with rates approaching R decodes errors on any BMS channel with $Z(\mathcal{W}) < 2^{1-R} - 1$.*

In particular, it decodes errors on channels $\mathrm{BSC}(p)$ with $p < 1/2 - \sqrt{2^{-R}(1-2^{-R})}$ and $\mathrm{BAWGN}(\sigma)$ with $\sigma^2 < -\frac{1}{2\ln(2^{1-R}-1)}$.

There is also a broader consequence of our result. Throughout the literature there are (somewhat varying) definitions of "asymptotically good" families of codes. In Hamming's model, a family of codes is usually considered good if it has both constant rate $R = \Omega(1)$ and linear minimum distance $d = \Omega(n)$. Therefore, being good is a property of the code. On the other hand, in Shannon's model, it is sometimes said [16] that a family of codes $\{V_n\}$ is good for a given channel $\mathcal{W}$ if it has constant rate $R = \Omega(1)$ and vanishing block-MAP error probability $P_{\mathrm{B}}(\mathcal{W}, V_n) = o_n(1)$. Our result shows that for linear codes with moderate minimum distance the notion of a good code is in fact independent of the channel. On the one hand, by a degradation argument (see Lemma 3.9), a code which is good for a BMS channel is also good for $\mathrm{BEC}(\lambda)$ for some $\lambda < 1$. On the other hand, by our results, a family of linear codes which is good for the BEC is also good for all BMS channels up to a certain Bhattacharyya parameter (and also above a certain capacity, see Remark 3.6 and Corollary 3.7).

2 PRELIMINARIES

In this section we give definitions that are most relevant to our results. Throughout we take $\log(\cdot)$ to denote the binary logarithm and $H(\cdot)$ the binary entropy. We also let $h_2(x) = -x\log x - (1-x)\log(1-x)$ for $0 \le x \le 1$ be the binary entropy function.

2.1 Error Correcting Codes

2.1.1 Basic Definitions. Let $V \subseteq \mathbb{F}_2^n$ be a binary linear code of block length n. Define the following:

[3]One should be careful about what "capacity-achieving" means in the regimes of rates approaching 0 or 1. See [3] for more details.

- Rate: $R(V) = \frac{\dim(V)}{n}$.
- Weight: For $v \in \mathbb{F}_2^n$ we define $\mathrm{wt}(v) = |\{i \mid v_i = 1\}|$.
- Minimum Distance: $d(V) = \min_{v \in V \setminus \{0^n\}} \mathrm{wt}(v)$.
- Weight Distribution: The sequence $(|\{v \in V \mid \mathrm{wt}(v) = k\}|)_{k=0}^n$ is the weight distribution of V.
- Dual: Let $V^\perp = \{u \in \mathbb{F}_2^n \mid \forall v \in V, u^T v = 0\}$ be the dual code of V which is a linear subspace of dimension $n - \dim(V)$.
- Restriction: Given $S \subseteq [n]$ define $V_S \subseteq \mathbb{F}_2^S$ as the restriction of V to the coordinates in S.

In the asymptotic setting, we consider families of codes $\{V_n\}$ where $n \in N \subseteq \mathbb{N}$ comes from an infinite set of block lengths. We define the rate of the family as $R = \lim_{n\to\infty} R(V_n)$ (if it exists). When $R = 0$ we say that the family $\{V_n\}$ has vanishing rate, and non-vanishing rate otherwise.

2.1.2 *Reed–Muller Codes.*

Definition 2.1. The Reed–Muller code $\mathrm{RM}(m, r) \subseteq \mathbb{F}_2^n$ is a linear code with block length $n = 2^m$ consisting of all evaluations of multilinear polynomials over $\mathbb{F}_2$ with m variables and degree at most r. That is, for every such polynomial f there is a codeword $(f(a))_{a \in \mathbb{F}_2^m}$.

It is known that $\mathrm{RM}(m, r)$ has minimum distance 2^{m-r} and rate $R(\mathrm{RM}(m, r)) = 2^{-m} \sum_{i=0}^{r} \binom{m}{i}$. Moreover, usually $r = r(m)$ is a function of m, and in this case the family $\{\mathrm{RM}(m, r(m))\}_m$ has non-vanishing rate smaller than 1 if and only if $r(m) = \frac{m}{2} \pm O(\sqrt{m})$. Therefore, a family of Reed–Muller codes with constant rate $R \in (0, 1)$ (i.e, $R \neq 0, 1$) has minimum distance $n^{1/2 \pm o_n(1)}$.

2.2 Shannon's Model

2.2.1 *Channels.* Abstractly, a binary channel is defined by two conditional probability distributions $p_{Y|X}$ where $X \in \{0, 1\}$ is binary and $Y \in \mathcal{Y}$ taken from some alphabet.

Definition 2.2 (Binary Erasure Channel)**.** Define the $\mathrm{BEC}(\lambda)$ over $\mathcal{Y} = \{0, 1, ?\}$ by $p_{Y|X}(? \mid 0) = p_{Y|X}(? \mid 1) = \lambda$, $p_{Y|X}(0 \mid 0) = p_{Y|X}(1 \mid 1) = 1 - \lambda$.

Definition 2.3 (Binary Symmetric Channel)**.** Define the $\mathrm{BSC}(p)$ over $\mathcal{Y} = \{0, 1\}$ by $p_{Y|X}(1 \mid 0) = p_{Y|X}(0 \mid 1) = p$, $p_{Y|X}(0 \mid 0) = p_{Y|X}(1 \mid 1) = 1 - p$.

Definition 2.4 (Binary Additive White Gaussian Noise Channel)**.** In this case we interpret the input as $X \in \{-1, +1\}$ and take $\mathrm{BAWGN}(\sigma)$ over $\mathcal{Y} = \mathbb{R}$ as

$$Y|X \sim \mathcal{N}(X, \sigma^2) \ .$$

In this work we focus on the class of binary memoryless symmetric channels:

Definition 2.5 (BMS Channel)**.** A BMS channel is a binary channel for which there exists an involution π on $\mathcal{Y}$ such that the distribution $p_{Y|X=0}$ is equal to $p_{\pi(Y)|X=1}$.

We remark that $\mathrm{BEC}(\lambda), \mathrm{BSC}(p), \mathrm{BAWGN}(\sigma)$ are all BMS channels.

When we say that a codeword $v \in \mathbb{F}_2^n$ is transmitted over a BMS channel $\mathcal{W}$, we always assume that the bits are transmitted over n independent instances of $\mathcal{W}$ (that is, in a memoryless fashion).

2.2.2 *MAP Decoding.* Let $\mathcal{W}$ be a BMS channel and $n \in \mathbb{N}$. Assume we transmit a uniformly random codeword $X \in V$ and denote $Y \in \mathcal{Y}^n$ the output of the channel. We define the *maximum a posteriori* (MAP) block decoding by

$$\hat{x}^{\mathrm{MAP}}(y) = \arg\max_{v \in V} p_{X|Y}(v \mid y) \ ,$$

breaking ties in an arbitrary way. For instance, for the BSC one can easily verify that $\hat{x}^{\mathrm{MAP}}(y)$ is simply the codeword $v \in V$ closest to y in the Hamming distance. Similarly, we can define the bit-MAP decoding

$$\hat{x_i}^{\mathrm{MAP}}(y) = \arg\max_{x_i \in \{0,1\}} p_{X_i|Y}(x_i \mid y).$$

The error of block/bit-MAP decoding is the probability that these two decoders are incorrect.

Definition 2.6 (Block Error)**.** The block-MAP error probability is defined by $P_{\mathrm{B}}(\mathcal{W}, V) = \mathbb{P}[\hat{x}^{\mathrm{MAP}}(Y) \neq X]$.

Definition 2.7 (Bit Error)**.** For $i \in [n]$ define the error of bit i via $P_{\mathrm{b,i}}(\mathcal{W}, V) = \mathbb{P}[\hat{x_i}^{\mathrm{MAP}}(Y) \neq X_i]$. The bit-MAP error probability is defined by $P_{\mathrm{b}}(\mathcal{W}, V) = \frac{1}{n} \cdot \sum_{i=1}^{n} P_{\mathrm{b,i}}(\mathcal{W}, V)$.

Definition 2.8 (Decoding errors)**.** For a family of linear codes $\{V_n\}$, we say that V_n decodes errors on a BMS channel $\mathcal{W}$ under block-MAP decoding if $\lim_{n\to\infty} P_{\mathrm{B}}(\mathcal{W}, V_n) = 0$, and under bit-MAP decoding if $\lim_{n\to\infty} P_{\mathrm{b}}(\mathcal{W}, V_n) = 0$.

2.2.3 *Decomposition of BMS Channels.* We present a known characterization of BMS channels which is useful for our presentation. We stick to a concise treatment very similar to Appendix A in [1] with a more complete one, e.g., in Chapter 4 of [20].

Let X be uniform in $\{0, 1\}$ and consider a BMS channel $\mathcal{W} : \{0, 1\} \to [0, 1/2] \times \{0, 1\}$ that maps a bit X to a pair (P, X') satisfying two conditions: First, P is independent of X and second, conditioned on P, X' is distributed according to $\mathrm{BSC}(P)$. In other words, for every transmitted bit the decoder sees its noisy copy together with information that the bit was flipped with probability P.

It is known that any BMS channel is equivalent to a mixture of BSC channels as described above. Accordingly, a BMS channel is fully characterized by the distribution of P. For example, $\mathrm{BSC}(p)$ has deterministic $P = p$ and $\mathrm{BEC}(\lambda)$ has $P = 1/2$ with probability λ and $P = 0$ otherwise. With that characterization in mind, we define the following quantities:

Definition 2.9 (Channel properties)**.** Let $\mathcal{W}$ be a BMS channel. We let its:

- *Capacity* to be $C(\mathcal{W}) = 1 - \mathbb{E}\, h_2(P)$.
- *Bhattacharyya parameter* to be $Z(\mathcal{W}) = 2\, \mathbb{E} \sqrt{P(1-P)}$.
- *(Single bit) error probability* to be $P_e(\mathcal{W}) = \mathbb{E}\, P$.

Note that $P_e(\mathcal{W})$ is the probability of error of the MAP decoder given a transmission of one uniform bit over $\mathcal{W}$ (this is because such decoder decodes a pair (P, X) to X, which was flipped with probability $P \leq 1/2$). We also remark that our definition of capacity for BMS channels is equivalent to the standard definition from information theory.

2.2.4 EXIT Functions. Extrinsic Information Transfer (EXIT) functions were originally introduced by ten Brink [26]. Based on his work, it was later observed that these EXIT functions are intimately related to the question of achieving capacity. Indeed, they played a key role in the proof that doubly transitive codes (see Definition 3.10) achieve capacity over the BEC under bit-MAP decoding [14].

Definition 2.10 (EXIT Function). Let $V \subseteq \mathbb{F}_2^n$ be a binary linear code. Define the EXIT function of V by

$$h(\lambda) = \frac{1}{n} \cdot \sum_{i=1}^{n} H(X_i | (Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_n)),$$

where $X = (X_1, \dots, X_n)$ is a uniformly random codeword in V and $Y = (Y_1, \dots, Y_n)$ is the result of transmitting X over $\mathrm{BEC}(\lambda)$.

We now list several key properties of the EXIT function.

Lemma 2.11. *Let $V \subseteq \mathbb{F}_2^n$ be a binary code. Then,*

- *Monotonicity: h is increasing from 0 to 1.*
- *Area Theorem: $\int_0^1 h(\epsilon) \mathrm{d}\epsilon = R(V)$.*
- *Duality: Denote by $h^{\perp}(\epsilon)$ the EXIT function of $V^{\perp}$ then*
$$h^{\perp}(\epsilon) = 1 - h(1 - \epsilon).$$
- *$n \cdot \int_0^{\lambda} h(\epsilon) = H(X|Y)$ where X is a random uniform codeword in V and Y is the result of transmitting X over the $\mathrm{BEC}(\lambda)$ channel.*
- *For every $\lambda \in (0, 1)$*
$$\begin{aligned} n \cdot \int_0^{\lambda} h(\epsilon) &= \lambda \cdot n - \mathop{\mathbb{E}}_{S \sim \lambda}[\dim(V_S^{\perp})] \\ &= \dim(V) - \mathop{\mathbb{E}}_{S \sim \lambda}[\dim(V_{S^c})], \end{aligned}$$
where $S \sim \lambda$ means that $i \in S$ with probability λ independently for every $i \in [n]$.

For proofs and further information please see chapter 3.14 in [20].

2.3 Boolean Analysis

We introduce basic notions from boolean analysis that we use in our proofs. Let $f : \{0, 1\}^n \to \mathbb{R}$ be a function from the boolean hypercube to the reals. We will use the Walsh–Fourier decomposition $f(x) = \sum_S \widehat{f}(S) \chi_S(x)$, where $\chi_S(x) = \prod_{i \in S} (-1)^{x_i}$ and $\widehat{f}(S) = \mathbb{E}_x f(x) \chi_S(x)$.

Definition 2.12. $\|f\|_2 = \sqrt{\mathbb{E}_x f(x)^2}$.

Lemma 2.13 (Parseval's Identity). *For any $f, g : \{0, 1\}^n \to \mathbb{R}$ we have $\mathbb{E}_x f(x) g(x) = \sum_S \widehat{f}(S) \widehat{g}(S)$. In particular, $\|f\|_2 = \sum_S \hat{f}(S)^2$.*

2.3.1 Noise Operator. For $x \in \{0, 1\}^n$ and $-1 \le \rho \le 1$, let $y \sim N_{\rho}(x)$ be a random element of $\{0, 1\}^n$ with each coordinate y_i being i.i.d equal to x_i with probability $(1 + \rho)/2$ and flipped with probability $(1 - \rho)/2$.

Definition 2.14 (Noise Operator). Let $f : \{0, 1\}^n \to \mathbb{R}$ and $\rho \in [-1, 1]$. Define the function $T_{\rho} f : \{0, 1\}^n \to \mathbb{R}$ by

$$T_{\rho} f(x) = \mathop{\mathbb{E}}_{y \sim N_{\rho}(x)} f(y).$$

Lemma 2.15. $\widehat{T_{\rho} f}(S) = \rho^{|S|} \cdot \hat{f}(S)$.

2.3.2 An Inequality on Noisy Functions. The following inequality is the main technical tool on which our results are based [23]. In order to state it, we need the notion of the conditional expectation of a function.

Definition 2.16. Let $f : \{0, 1\}^n \to \mathbb{R}$ and $S \subseteq [n]$ be a subset of coordinates. We define another function $\mathbb{E}(f|S)(x) = \mathbb{E}_{y : y_S = x_S} f(y)$.

We shall only state the theorem for the ℓ_2 norm as that is all we are going to use (For the general statement see Theorem 1.1 in [23]).

Theorem 2.17. *Let $f : \{0, 1\}^n \to \mathbb{R}^{\ge 0}$ be a non-negative function, and $\rho \in (0, 1)$. Then,*

$$\log ||T_{\rho} f||_2 \le \mathop{\mathbb{E}}_{S \sim \lambda(\rho)} \log \mathbb{E}(f \mid S),$$

where $\lambda(\rho) = \log(1 + \rho^2)$ and $S \sim \lambda$ is a random subset S of $[n]$ in which each element is included independently with probability λ.

Theorem 2.17 can be compared to the classical hypercontractive inequality $\|T_{\rho} f\|_2 \le \|f\|_{1+\rho^2}$, see also [22] for a more extensive discussion.

3 OUR RESULTS

Our main contribution is in realizing that one can combine [14] with [23] as well as other techniques from coding theory to obtain a rather general understanding about the performance of a given binary linear code on various BMS channels.

3.1 Coding on BMS Channels

Our main technical result relates the weight distribution of a linear code to its decoding properties on the BEC:

Theorem 3.1. *Let V be a linear code, $0 \le \lambda \le 1$ and $(a_0, \dots, a_n)$ be the weight distribution of V. Then,*

$$\log \sum_{i=0}^{n} a_i \cdot (2^{\lambda} - 1)^i \le H(X|Y) , \tag{1}$$

Here X and Y are random variables such that X is a random uniform codeword in V and Y is the result of transmitting X over the channel $\mathrm{BEC}(\lambda)$.

Theorem 3.1 is a reformulation of Proposition 1.3 in [22] using a well-known identity. Using the original formulation, it was established in [22] that the right-hand side in Theorem 3.1 becomes $o(n)$ for duals of linear codes that achieve capacity on the BEC. The difference in this work is to rewrite the bound in terms of conditional entropy $H(X|Y)$. While the difference is slight, it allows for easier bounds in terms of both block-MAP and bit-MAP error probabilities and gives new applications to coding. The proof of Theorem 3.1 appears in Appendix A.

Corollary 3.2. *Let V be a binary linear code with dimension k and weight distribution $(a_0, \dots, a_n)$ then*

$$\sum_{i=0}^{n} a_i (2^{\lambda} - 1)^i \le 2^{k \cdot P_{\mathrm{B}}(\mathrm{BEC}(\lambda), V)} ,$$
$$\sum_{i=0}^{n} a_i (2^{\lambda} - 1)^i \le 2^{n \cdot P_{\mathrm{b}}(\mathrm{BEC}(\lambda), V)} .$$

In particular, if $k \cdot P_{\mathrm{B}}(\mathrm{BEC}(\lambda), V) < 1$ *or* $n \cdot P_{\mathrm{b}}(\mathrm{BEC}(\lambda), V) < 1$, *then* $a_i < 2(2^\lambda - 1)^{-i}$ *for every* i.

Proof. Since on the BEC we have $H(X|Y = y) = 0$ if and only if the block-MAP decoder succeeds on the received pattern y and $H(X|Y = y) \leq k = \dim V$ in any event, the right-hand side of (1) can be bounded by

$$H(X|Y) \leq k \cdot P_{\mathrm{B}}(\mathrm{BEC}(\lambda), V) \ . \tag{2}$$

On the other hand, by the chain rule there also holds a bound in terms of the bit-error probability

$$\begin{aligned} H(X|Y) &\leq \sum_{i=1}^{n} H(X_i|Y) \\ &= \sum_{i=1}^{n} P_{\mathrm{b,i}}(\mathrm{BEC}(\lambda), V) = n \cdot P_{\mathrm{b}}(\mathrm{BEC}(\lambda), V). \end{aligned} \tag{3}$$

□

The proof of Theorem 1.2 is based on the following well-known bound (see, e.g., Lemma 4.67 in [20] and also [10] for the tighter version):

Theorem 3.3 (Bhattacharyya Bound)**.** *Let* V *be a linear code and* $(a_0, a_1 \dots)$ *be its weight distribution. Then for any BMS channel* $\mathcal{W}$,

$$P_{\mathrm{B}}(\mathcal{W}, V) \leq \sum_{i=1}^{n} a_i \cdot Z(\mathcal{W})^i.$$

In order to make our argument self-contained, we provide a proof for Theorem 3.3 in Appendix B.

Proof of Theorem 1.2. We use the Bhattacharyya bound, the minimum distance and Corollary 3.2 to conclude that

$$\begin{aligned} P_{\mathrm{B}}(\mathcal{W}, V) &\leq \sum_{i=1}^{n} a_i Z(\mathcal{W})^i = \sum_{i=d}^{n} a_i Z(\mathcal{W})^i \\ &\leq \left(\frac{Z(\mathcal{W})}{2^\lambda - 1}\right)^d \sum_{i=d}^{n} a_i (2^\lambda - 1)^i < 2\left(\frac{Z(\mathcal{W})}{2^\lambda - 1}\right)^d . \end{aligned}$$

□

Recall that we say that a family of codes $\{V_n\}$ decodes errors on channel $\mathcal{W}$ if $\lim_{n\to\infty} P_{\mathrm{B}}(\mathcal{W}, V_n) = 0$. We will apply Theorem 1.2 to the canonical cases of $\mathrm{BSC}(p)$ and $\mathrm{BAWGN}(\sigma)$, obtaining:

Corollary 3.4. *Let* $\{V_n\}$ *be a family of linear codes with dimensions* $\dim(V_n) = k_n \to \infty$ *that satisfies* $P_{\mathrm{B}}(\mathrm{BEC}(\lambda), V_n) < 1/k_n$ *for large* n. *Then:*

(1) $\{V_n\}$ *decodes errors on any BMS channel* $\mathcal{W}$ *with* $Z(\mathcal{W}) < 2^\lambda - 1$.
(2) $\{V_n\}$ *decodes errors on* $\mathrm{BSC}(p)$ *as long as* $p < \frac{1}{2} - \sqrt{2^{\lambda-1}(1 - 2^{\lambda-1})}$.
(3) $\{V_n\}$ *decodes errors on* $\mathrm{BAWGN}(\sigma)$ *as long as* $\sigma^2 < -\frac{1}{2\ln(2^\lambda - 1)}$.

Proof. The first point is immediate from Theorem 1.2, with the additional observation that $k_n \to \infty$ together with

$$P_{\mathrm{B}}(\mathrm{BEC}(\lambda), V_n) < 1/k_n$$

imply that the minimum distance d_n grows to infinity with n and therefore indeed $P_B < 2(Z(\mathcal{W})/\theta)^{d_n}$ vanishes. The other points now follow substituting known formulas $Z(\mathrm{BSC}(p)) = 2\sqrt{p(1-p)}$ and $Z(\mathrm{BAWGN}(\sigma)) = \exp(-1/2\sigma^2)$ (see, e.g., Examples 4.128–4.130 in [20]). □

A graphical illustration of the functions from Corollary 3.4 is provided in Figure 1. Theorem 1.3 follows by Corollary 3.4 and the fact that constant rate RM codes achieve capacity on the BEC [14].

Remark 3.5. We note that the requirement $P_{\mathrm{B}}(\mathrm{BEC}(\lambda), V_n) < 1/k_n$ is not much stronger than $P_{\mathrm{B}}(\mathrm{BEC}(\lambda), V_n) = o(1)$. In particular, by Theorem 5.2 in [27], if the minimum distance of a linear code satisfies $d_n = \omega(\log n)$, then $P_{\mathrm{B}}(\mathrm{BEC}(\lambda), V_n) = o(1)$ implies $P_{\mathrm{B}}(\mathrm{BEC}(\lambda'), V_n) = o(n^{-c})$ for every $\lambda' < \lambda$ and $c > 0$.

Remark 3.6. Since among BMS channels with fixed capacity $C(\mathcal{W})$ the Bhattacharyya coefficient is maximized by the BSC (see Problem 4.60 in [20]), a family of linear codes that decodes errors on $\mathrm{BEC}(\lambda)$ also decodes errors on all BMS channels with capacity $C(\mathcal{W}) > C(\mathrm{BSC}(p)) = 1 - h(p)$ where $p = p(\lambda)$ is given in Corollary 3.4. Similarly, a BMS channel $\mathcal{W}$ can be degraded to $\mathrm{BSC}(P_e(\mathcal{W}))$, where $P_e(\mathcal{W})$ is the (one-bit) error probability of $\mathcal{W}$. Therefore, a family of linear codes that decodes errors on $\mathrm{BEC}(\lambda)$ also decodes errors on all BMS channels with $P_e(\mathcal{W}) < P_e(\mathrm{BSC}(p)) = p$ with $p = p(\lambda)$ as above.

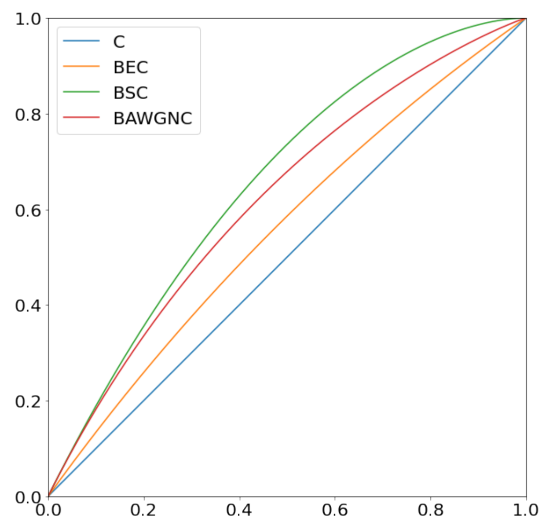

Figure 1: An illustration of the results in Corollary 3.4. Assume we are given a family of linear codes that decodes errors on the channel $\mathrm{BEC}(\lambda)$ with capacity $C(\mathrm{BEC}(\lambda)) = 1 - \lambda$. Then, by Corollary 3.4 it is also good for the BSC and the BAWGN exceeding certain capacities. These capacities are plotted here as functions of $C(\mathrm{BEC}(\lambda))$.
For reference, we also plot the identity function. The difference between identity and the BSC and the BAWGN curves represents the "loss of capacity" resulting when Corollary 3.4 is applied.
As another reference point, the graph labeled "BEC" shows this loss of capacity if our argument is applied to the BEC channel.

Finally, we have that a family of linear codes that decodes errors well enough on any BMS channel $\mathcal{W}$ also decodes errors on a range of other BMS channels:

Corollary 3.7. *Let $\{V_n\}$ be a family of linear codes with dimensions $k_n \to \infty$ satisfying $P_{\mathrm{B}}(\mathcal{W}, V_n) < 1/k_n$ for large n on some BMS channel $\mathcal{W}$. Then, $\{V_n\}$ decodes errors on any BMS channel $\mathcal{W}'$ satisfying*

$$Z(\mathcal{W}') < 4^{P_e(\mathcal{W})} - 1\ .$$

To prove Corollary 3.7, we need the notion of channel degradation:

Definition 3.8. Let $\mathcal{W} : \{0,1\} \to \mathcal{Y}$ and $\mathcal{W}' : \{0,1\} \to \mathcal{Y}'$ be two BMS channels. We say that $\mathcal{W}$ can be degraded to $\mathcal{W}'$ if there exists a channel $\mathcal{V} : \mathcal{Y} \to \mathcal{Y}'$ such that $\mathcal{W}'$ is the composition of $\mathcal{W}$ and $\mathcal{V}$.

We now use the following folklore fact:

Lemma 3.9. *If $\mathcal{W}$ is a BMS channel, then* $\mathrm{BEC}(2 \cdot P_e(\mathcal{W}))$ *can be degraded to $\mathcal{W}$.*

The proof of Lemma 3.9 can be found, e.g., as Lemma 4.80 in [20]. In short, it follows from the decomposition of BMS channels into BSC that we described in Section 2, an easily checked fact that $\mathrm{BEC}(2p)$ can be degraded to $\mathrm{BSC}(p)$, and the fact that a convex combination of BEC channels is itself a BEC channel.

Proof of Corollary 3.7. Let $\{V_n\}$ be the family of linear codes from the statement and $\mathcal{W}$ a BMS channel. By Lemma 3.9, $\mathrm{BEC}(2 \cdot P_e(\mathcal{W}))$ can be degraded to $\mathcal{W}$. Clearly, that gives

$$P_{\mathrm{B}}(\mathrm{BEC}(2 \cdot P_e(\mathcal{W})), V_n) \le P_{\mathrm{B}}(\mathcal{W}, V_n) < 1/k_n$$

(since a decoder for V_n on $\mathrm{BEC}(2 \cdot P_e(\mathcal{W}))$ can use the channel $\mathcal{V}$ from the definition of degradation to simulate $\mathcal{W}$ and invoke the MAP decoder for $\mathcal{W}$). The proof is concluded by an invocation of Corollary 3.4. □

3.2 Doubly Transitive Codes

In [14] it was proved that any *doubly transitive* binary linear code (see definition below) achieves capacity for the BEC by proving that its corresponding EXIT function has a *sharp threshold* (i.e., rises quickly from nearly 0 to nearly 1). Recall that $h(\lambda)$ is a monotone function increasing from 0 to 1 (see Lemma 2.11). Thus, if $h(\lambda)$ has a sharp threshold then this transition has to occur around $\lambda = 1-R(C)$ by area considerations. In fact, $h(\lambda)$ has a sharp threshold if and only if V achieves capacity for the BEC. In order to prove that $h(\lambda)$ has a sharp threshold they use tools from boolean analysis which we shall soon cover.

Definition 3.10. The permutation group of a binary linear code $V \subseteq \mathbb{F}_2^n$ is the group of all permutations that the code is invariant under. Namely,

$$G = \{\pi \in S_n \mid \forall v \in V,\ \pi(v) \in V\}.$$

We say that V is doubly transitive if G is, i.e for all i, j, k distinct there exists $\pi \in G$ such that $\pi(i) = i$, $\pi(j) = k$.

Throughout this section, $V \subseteq \mathbb{F}_2^n$ is a doubly transitive code and $h(p)$ denotes its EXIT function. The main technical tool in proving that $h(p)$ has a sharp threshold is the following general theorem on monotone sets which are *sufficiently symmetric*.

Definition 3.11. Let $\Omega \subseteq \{0,1\}^n$ then:

- We say Ω is monotone if $x \in \Omega$ and $x \le y$ (i.e, $\forall i,\ x_i \le y_i$) implies $y \in \Omega$.
- We denote $\mu_p(\Omega) = \sum_{x \in \Omega} p^{|\{i:x_i=1\}|}(1-p)^{|\{i:x_i=0\}|}$.

Theorem 3.12 ([9] (Informal))**.** *Let $\Omega \subseteq \{0,1\}^n$ be a monotone set which is sufficiently symmetric then,*

$$\frac{\mathrm{d}\mu_p(\Omega)}{\mathrm{d}p} \ge (c(p) - o_n(1)) \cdot \ln(n) \cdot \mu_p(\Omega)(1 - \mu_p(\Omega)),$$

where $c(p) = \frac{1-2p}{p(1-p)\ln\left(\frac{1-p}{p}\right)}$[4].

Remark 3.13. In [9] the above was proved with a different constant. The bound above with $c(p)$ was obtained in [21].

In the terminology of boolean analysis, the EXIT function of doubly transitive code $h(p)$ is the μ_p-measure of some monotone set $\Omega \subseteq \{0,1\}^n$, and the *sufficiently symmetric* requirement in the above theorem is fulfilled since V is doubly transitive. Thus, by applying Theorem 3.12 we obtain

$$h'(p) \ge c(p) \cdot \ln(n) \cdot h(p)(1 - h(p)). \tag{4}$$

The following implications of Theorem 3.12 appeared in [14].

Lemma 3.14. *Let $h(\lambda)$ be the EXIT function of a doubly transitive linear code $V \subseteq \mathbb{F}_2^n$, and $p_c = h^{-1}(1/2)$ then*

$$h(p_c - \epsilon) \le n^{-k(p_c)\cdot\epsilon}, \tag{5}$$

where $k(p) = \begin{cases} c(p) - o_n(1) & p < 1/2 \\ c(1/2) - o_n(1) & p \ge 1/2 \end{cases}$.

Proof. The constant $c(p)$ in Equation (4) is decreasing on the interval $(0, 1/2)$ and attains its global minimum on $[0,1]$ at $p = 1/2$. Thus, by Equation (4) we get

$$\forall p \le h^{-1}(1/2),\ h'(p) \ge k(p)\ln(n) \cdot h(p)(1 - h(p)).$$

It is known that for any monotone function $h : [0,1] \to [0,1]$ increasing from 0 to 1 the above inequality implies that $h(h^{-1}(1/2) - \epsilon) \le \exp(-k(p)\ln(n)\epsilon)$ (for more details see Lemma 34 in [14]). □

As a consequence of Lemma 3.14 and of the first bullet in Lemma 2.11 one gets the following (Proposition 11 in [14]).

Lemma 3.15. *Let $h_n(\lambda)$ be the EXIT function of a family of doubly transitive linear codes $\{V_n\}$ of rate R then* $\lim_{n\to\infty} h_n^{-1}(1/2) = 1 - R$.

For more details see section III in [14].

Now we present another approach for bounding $h(p)$.

Lemma 3.16. *Let $h(\lambda)$ be the EXIT function of a doubly transitive linear code $V \subseteq \mathbb{F}_2^n$. Then for every $p \in [0,1]$ and $t \ge 1$ we have $h(\lambda^t) \le h(\lambda)^t$.*

[4] At $p = 1/2$ the function $c(p)$ has a removable discontinuity and so we define $c(1/2) = \lim_{p\to 1/2} c(p) = 2$.

Proof. For a doubly transitive code we have $h(p) = \mu_p(\Omega)$ for some monotone set $\Omega \subseteq \{0,1\}^n$. By Lemma 2.7 in [8], for a monotone set Ω and $t \geq 1$ it holds that $\mu_{p^t}(\Omega) \leq \mu_p(\Omega)^t$. □

Theorem 3.17. *Let $\{V_n\}$ be a family of doubly transitive binary linear codes of rate $R \in (0,1)$, and $\mathcal{W}$ be a BMS channel. Denote by $\alpha = \liminf_{n\to\infty} \frac{\log d(V_n)}{\log n}$. Then, V can decode errors on $\mathcal{W}$ if $Z(\mathcal{W}) < 2^{\lambda(R,\alpha)} - 1$ where*

$$\lambda(R,\alpha) = \begin{cases} (1-R) - \frac{1-\alpha}{k(1-R)} & \text{if } \psi(R,\alpha) \geq 0 \\ \left(\frac{(1-R)/e}{-W_{-1}(-(1-R)/e)}\right)^{\frac{1-\alpha}{k(1-R)\cdot(1-R)}} & \text{otherwise.} \end{cases},$$

where $\psi(R,\alpha) = (1-R) + \frac{1-R}{W_{-1}(-(1-R)/e)} - \frac{1-\alpha}{k(1-R)}$ and $W_{-1}(z)$ is the inverse of the function $y = xe^x$ at the interval $x \in (-1/e, 0)$ for $y \leq -1$ (a.k.a the negative branch of the Lambert function).

Remark 3.18. It is not clear why the two cases of $\lambda(R,\alpha)$ coincide at $\psi(R,\alpha) = 0$, but one can verify this easily using simple identities of the Lambert function.

For instance, plugging $\alpha = 0.5$ and $R = 0.8$ into Theorem 3.17 we get $\lambda(R,\alpha) \approx 0.0331$. This implies that any family $\{V_n\}$ of doubly transitive codes with rate $R = 0.8$ and minimum distance $d(V_n) = \Omega(\sqrt{n})$ can decode errors from BEC(0.023) under block-MAP decoding.

Proof. Let us denote $(a_0, a_1, \ldots, a_n)$ the weight distribution of V_n, $d = d(V_n)$, and $\lambda = \lambda(R,\alpha)$. Note that for any $\delta \in (0,1)$ we have

$$\sum_{i=d}^{n} a_i \cdot \left(\frac{2^\lambda - 1}{1+\delta}\right)^i \leq (1+\delta)^{-d} \cdot \sum_{i=d}^{n} a_i \cdot (2^\lambda - 1)^i,$$

and using Theorem 3.1 we have

$$\sum_{i=d}^{n} a_i \cdot \left(\frac{2^\lambda - 1}{1+\delta}\right)^i \leq (1+\delta)^{-d} \cdot 2^{H(X|Y)}, \tag{6}$$

where X and Y are random variables such that X is a random uniform codeword in V and Y is the result of transmitting X over the channel BEC(λ). By Lemma 2.11 we get

$$\sum_{i=d}^{n} a_i \cdot \left(\frac{2^\lambda - 1}{1+\delta}\right)^i \leq (1+\delta)^{-d} \cdot 2^{n\cdot h(\lambda)}.$$

Recall that by assumption $Z(\mathcal{W}) < 2\lambda - 1$ and hence for small enough constant $\delta \in (0,1)$ we get $\frac{2^\lambda - 1}{1+\delta} \geq Z(\mathcal{W})$. Then by Theorem 3.3 we get

$$P_{\mathrm{B}}(\mathcal{W}, V_n) \leq \sum_{i=d}^{n} a_i \cdot Z(\mathcal{W})^i \leq (1+\delta)^{-d} \cdot 2^{n\cdot h(\lambda)}.$$

We conclude that it suffices to show that $n \cdot h(\lambda) = o(d(V_n))$. Moreover, in our notation $d(V_n) = \Omega(n^{\alpha \pm o_n(1)})$, and so a bound of $n \cdot h(\lambda) = O(n^\beta)$ with constant $\beta < \alpha$ suffices. This means that we want λ so that $h(\lambda) < n^{-\beta}$ with $\beta > 1 - \alpha$. It remains to verify that our choice of $\lambda = \lambda(R,\alpha)$ satisfies this.

We have two ways in which we can bound $h(\lambda)$: The additive bound of Lemma 3.14, and the multiplicative bound of Lemma 3.16. It is easy to see that the additive bound is stronger for values close to the critical value $p_c = h^{-1}(1/2)$, and on the other hand the multiplicative bound is better for values that are further away from p_c. Moreover, the additive bound is limited to $\epsilon < p_c$. Applying first the additive bound with parameter ϵ and then the multiplicative bound with parameter $t \geq 1$ we get

$$h((p_c - \epsilon)^t) \leq h(p_c - \epsilon)^t \leq n^{-t\epsilon k(p_c)}.$$

We are going to choose λ of the form $\lambda = (p_c - \epsilon)^t$ for some valid choice of ϵ, t so that $t \cdot \epsilon \cdot k(p_c) > 1 - \alpha$. Under this constraint, we want λ to be as large as possible. We are now going to cheat slightly. Instead of solving the optimization problem for λ under the constraint $t \cdot \epsilon \cdot k(p_c) > 1 - \alpha$ we will solve it for $t \cdot \epsilon \cdot k(p_c) = 1 - \alpha$. This suffices to prove the theorem, e.g by repeat the argument for λ' such that $Z(\mathcal{W}) < \lambda' < \lambda(R,\alpha)$ and applying Lemma 3.16. The optimal value of λ under this constraint will be shown to be $\lambda(R,\alpha) - o_n(1)$.

The above yields the following optimization problem.

$$\max_{t\geq 1, \epsilon \leq p_c} \{(p_c - \epsilon)^t\}, \quad tk(p_c)\epsilon = (1-\alpha). \tag{7}$$

The solution to this optimization problem is given by

$$p_{opt} = \begin{cases} p_c - \frac{1-\alpha}{k(p_c)} & \frac{1-\alpha}{k(p_c)} \leq p_c + \frac{p_c}{W_{-1}(-p_c/e)} \\ \left(\frac{p_c/e}{-W_{-1}(-p_c/e)}\right)^{\frac{1-\alpha}{k(p_c)\cdot p_c}} & otherwise. \end{cases}$$

See Appendix D for proof. Recall that by Lemma 3.15 we have $p_c = 1 - R - o_n(1)$ and hence $p_{opt} = \lambda(R,\alpha) - o_n(1)$ as claimed. □

Remark 3.19. It is was shown in [14] that any family of binary linear doubly transitive codes $\{V_n\}$ such that $\liminf_{n\to\infty} \frac{\log d(V)}{\log n} = 1$ achieves capacity for the BEC under block-MAP decoding (Theorem 21 in [14]). Combining this with Theorem 1.2 implies Theorem 3.17 for the special case $\alpha = 1$.

3.2.1 Discussion on Reed–Muller Codes. It is interesting to see that Theorem 3.17 is inferior to Theorem 1.3 in the case of Reed–Muller codes. In fact, recall that Reed–Muller codes of constant rate have minimum distance of roughly $n^{1/2}$, are doubly transitive, and yet plugging $\alpha = 1/2$ in Theorem 3.17 does not yield the parameters of Theorem 1.3. Rather, the parameters of Theorem 1.3 correspond to the case $\alpha = 1$ in Theorem 3.17. This means that RM codes perform better on BMS channels than might be expected from their minimal distance as expected from Theorem 3.17. Let us try to explain this phenomenon. We give two explanations. First, in [15] Kudekar, Kumar, Mondelli, Pfister, and Urbanke proved that for any constants $z, \beta \in (0,1)$ it holds that

$$\sum_{i=1}^{n^{1-\beta}} a_i \cdot z^i = o_n(1),$$

where $(a_0, a_1, \ldots, a_n)$ is the weight distribution of the Reed–Muller code (Lemma 4 in [15]). Hence, we can omit the first $n^{1-o_n(1)}$ terms in Equation (6) and continue the argument as in Theorem 3.17 but with $\alpha = 1$. This means, in a well-defined sense, that the Reed–Muller codes *effectively* have distance $n^{1-o(1)}$. An alternative explanation, which in fact follows the original approach in [14], is to use stronger results on sharp thresholds since RM codes are more than just doubly transitive. Specifically, the bound of Lemma 3.15 can be improved to $n^{-\Omega(\log\log n)}$ in the case of the Reed–Muller codes [6]. Using this improved bound and following the same approach as in Theorem 3.17 also leads to Theorem 1.3.

3.3 Weight Distribution of Codes

Coming back to weight distributions, and using the argument from [22], Theorem 3.1 can be applied to the dual code $V^{\perp}$, resulting in a different bound on the weight distribution:

Proposition 3.20. *Let V be linear code of dimension k, $0 \leq q \leq 1$ and $(b_0, \ldots, b_n)$ be the weight distribution of the dual code $V^{\perp}$. For $0 \leq i \leq n$, let $i^* = \min\{i, n-i\}$. Then,*

$$b_i \;\leq\; 2^{H(X|Y)} \cdot \begin{cases} \frac{|V^{\perp}|}{(1-\theta)^{i^*}(1+\theta)^{n-i^*}} & 0 \leq i^* \leq \frac{1-\theta}{2} \cdot n \\ \frac{|V^{\perp}|}{2^n} \cdot 2^{h_2(i/n)\cdot n} & \text{otherwise} \end{cases}$$

where X is a random uniform codeword in V, Y is the result of transmitting X over $\mathrm{BEC}(\lambda)$ and $\theta = 2^{\lambda} - 1$.

The proof of Proposition 3.20 uses simple Fourier analysis and is very similar to the proof of Proposition 1.6 in [22]. We include a sketch to make the argument self-contained in Appendix C.

Remark 3.21. Since $2^{h_2(i/n)n} \leq O(\sqrt{n})\binom{n}{i}$, whenever $H(X|Y) = o(n)$ for a code V, the weight distribution of the dual code $V^{\perp}$ in a band of weights of width θ around $\frac{n}{2}$ is essentially upper-bounded by that of a random code of the same rate. This occurs even if V does not achieve capacity: It is enough that V decodes errors on $\mathrm{BEC}(\lambda)$ for some constant $\lambda < 1$. (Cf. [13], where similar behavior was inferred for codes with large dual distance.)

In particular, consider a family of doubly transitive binary linear codes of constant rate R. Since by [14] such a family achieves capacity under bit-MAP decoding, due to (3) it will have the bound from Proposition 3.20 holding with $H(X|Y) = o(n)$. Such bound holds for both primal and dual codes, since the dual of a doubly transitive code is also doubly transitive.

Similarly, again building on [22], we improve the bounds on the weight distribution of doubly transitive codes with distance $\Omega(n^{\alpha})$.

Proposition 3.22. *Let $\{V_n\}$ be a family of doubly transitive binary linear codes of rate R and set $\alpha = \liminf_{n\to\infty} \frac{\log d(V)}{\log n}$. Also, let $(a_0, a_1, \ldots)$ denote the weight distribution of V_n. Then*

$$a_i \leq (2^{\lambda(R,\alpha)-o_n(1)} - 1)^i,$$

where $\lambda(R, \alpha)$ is as in Theorem 3.17.

One should compare the above with Proposition 1.6 in [22]. The main difference is that proposition in [22] applies only to codes that achieve capacity for the BEC under block-MAP decoding, and it is not known whether a doubly transitive code with minimum distance $\Omega(n^{\alpha})$ for $\alpha < 1$ achieves capacity for the BEC under block-MAP decoding. Yet, Proposition 3.22 holds for any doubly transitive code with minimum distance $n^{\Omega(1)}$. Moreover, Proposition 1.6 in [22] had an error term that dominated the estimate for weights $i = o(n)$.

For the weight distribution of Reed–Muller codes we obtain the following bound.

Proposition 3.23. *Let $(a_0, a_1, \ldots,)$ denote the weight distribution of Reed–Muller codes with rate $R \in (0, 1)$. Then*

$$a_i \leq O\left((2^{1-R-o_n(1)} - 1)^i\right).$$

Again, the above estimate improves over [22] by losing the error term that dominated the estimate for weights $i = o(n)$. However, in the range of $i = o(n)$ for Reed–Muller codes there are stronger bounds for weights $i = n^{1-\delta}$ (e.g, see [3, 12, 15, 24]) for every fixed constant $\delta \in (0, 1)$. Hence, the improvement lies in the narrow region of weights $n^{1-o_n(1)}$ for some $o_n(1)$.

The proofs of Proposition 3.23 and Proposition 3.22 are omitted as those can be easily derived by the arguments of Theorem 3.17.

ACKNOWLEDGEMENTS

We are grateful to Or Ordentlich, Nathan Keller, Emmanuel Abbe, Ido Nachum, and Amir Shpilka for many very helpful conversations and valuable remarks. Additionally, we thank Emmanuel Abbe and a referee for pointing out the connection to the Bhattacharyya bound.

REFERENCES

[1] Emmanuel Abbe, Elisabetta Cornacchia, Yuzhou Gu, and Yury Polyanskiy. 2021. Stochastic block model entropy and broadcasting on trees with survey. (2021).

[2] Emmanuel Abbe, Jan Hązła, and Ido Nachum. 2020. Almost-Reed–Muller codes achieve constant rates for random errors. (2020). arXiv:2004.09590.

[3] Emmanuel Abbe, Amir Shpilka, and Avi Wigderson. 2015. Reed–Muller codes for random erasures and errors. *IEEE Transactions on Information Theory* 61, 10 (2015), 5229–5252.

[4] Emmanuel Abbe, Amir Shpilka, and Min Ye. 2020. Reed–Muller codes: Theory and algorithms. *IEEE Transactions on Information Theory* (2020).

[5] Erdal Arıkan. 2009. Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels. *IEEE Transactions on information Theory* 55, 7 (2009), 3051–3073.

[6] Jean Bourgain and Gil Kalai. 1997. Influences of variables and threshold intervals under group symmetries. *Geometric & Functional Analysis GAFA* 7, 3 (1997), 438–461.

[7] Philippe Delsarte, Jean-Marie Goethals, and F Jessie Mac Williams. 1970. On generalized Reed–Muller codes and their relatives. *Information and Control* 16, 5 (1970), 403–442.

[8] David Ellis, Nathan Keller, and Noam Lifshitz. 2019. Stability versions of Erdős–Ko–Rado type theorems via isoperimetry. *Journal of the European Mathematical Society* 21, 12 (2019), 3857–3902.

[9] Ehud Friedgut and Gil Kalai. 1996. Every monotone graph property has a sharp threshold. *Proc. Amer. Math. Soc.* 124, 10 (1996), 2993–3002.

[10] Miguel Griot, Wen-Yen Weng, and Richard D Wesel. 2007. A tighter Bhattacharyya bound for decoding error probability. *IEEE Communications Letters* 11, 4 (2007), 346–347.

[11] Richard W Hamming. 1950. Error detecting and error correcting codes. *Bell System technical journal* 29, 2 (1950), 147–160.

[12] Tali Kaufman, Shachar Lovett, and Ely Porat. 2012. Weight distribution and list-decoding size of Reed–Muller codes. *IEEE Transactions on Information Theory* 58, 5 (2012), 2689–2696.

[13] Ilia Krasikov and Simon Litsyn. 1997. Estimates for the range of binomiality in codes' spectra. *IEEE Transactions on Information Theory* 43, 3 (1997), 987–991.

[14] Shrinivas Kudekar, Santhosh Kumar, Marco Mondelli, Henry D Pfister, Eren Şaşoğlu, and Rüdiger L Urbanke. 2017. Reed–Muller codes achieve capacity on erasure channels. *IEEE Transactions on information theory* 63, 7 (2017), 4298–4316.

[15] Shrinivas Kudekar, Santhosh Kumar, Marco Mondelli, Henry D Pfister, and Rüdiger Urbankez. 2016. Comparing the bit-MAP and block-MAP decoding thresholds of Reed-Muller codes on BMS channels. In *2016 IEEE International Symposium on Information Theory (ISIT)*. Ieee, 1755–1759.

[16] David JC MacKay. 1999. Good error-correcting codes based on very sparse matrices. *IEEE Transactions on Information Theory* 45, 2 (1999), 399–431.

[17] David E Muller. 1954. Application of Boolean algebra to switching circuit design and to error detection. *Transactions of the IRE Professional Group on Electronic Computers* EC-3, 3 (1954), 6–12.

[18] James G Oxley. 2006. *Matroid theory*. Vol. 3. Oxford University Press, USA.

[19] I. Reed. 1954. A class of multiple-error-correcting codes and the decoding scheme. *Transactions of the IRE Professional Group on Information Theory* 4, 4 (1954), 38–49.

[20] Tom Richardson and Rüdiger Urbanke. 2008. *Modern Coding Theory*. Cambridge University Press.

[21] Raphaël Rossignol. 2005. Threshold for monotone symmetric properties through a logarithmic Sobolev inequality. *The Annals of Probability* 34, 5 (2005), 1707–1725.

[22] Alex Samorodnitsky. 2019. An upper bound on ℓ_q norms of noisy functions. *IEEE Transactions on Information Theory* 66, 2 (2019), 742–748.
[23] Alex Samorodnitsky. 2020. An improved bound on ℓ_q norms of noisy functions. (2020). arXiv:2010.02721.
[24] Ori Sberlo and Amir Shpilka. 2020. On the performance of Reed-Muller codes with respect to random errors and erasures. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 1357–1376.
[25] Claude Elwood Shannon. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal* 27, 3 (1948), 379–423.
[26] Stephan Ten Brink. 1999. Convergence of iterative decoding. *Electronics letters* 35, 10 (1999), 806–808.
[27] Jean-Pierre Tillich and Gilles Zémor. 2000. Discrete isoperimetric inequalities and the probability of a decoding error. *Combinatorics Probability and Computing* 9, 5 (2000), 465–479.

A NORMS OF NOISY FUNCTIONS

In this section we prove Theorem 3.1. We start with the following lemma which is simply the formula for the rank of the dual matroid, but we state it as a separate claim because of its importance to us.

Lemma A.1 (Proposition 2.1.9 in [18]). *Let V be a linear code of dimension k. Then,* $\dim V_S^{\perp} = |S| - \left(k - \dim V_{S^c}\right)$.

We now prove Theorem 3.1 by following the argument in [22].

Proof of Theorem 3.1. Let V be a linear code of dimension k and $0 \le \lambda \le 1$. We shall prove that

$$\log \sum_{i=0}^{n} a_i (2^{\lambda} - 1)^i \le \lambda n - \mathop{\mathbb{E}}_{S\sim\lambda} \dim V_S^{\perp} \tag{8}$$
$$= k - \mathop{\mathbb{E}}_{S\sim 1-\lambda} \dim V_S = H(X|Y) ,$$

where X is a random uniform codeword in V and Y is the result of transmitting X over the channel BEC(λ). Define $f = 2^k 1_{C^{\perp}}$ and $\rho = \sqrt{2^{\lambda} - 1}$ and recall that the Walsh–Fourier transform of f satisfies $\widehat{f} = 1_C$. Since $\widehat{T_\rho f}(y) = \rho^{|y|}\widehat{f}(y)$, using Parseval's identity this means $\log \|T_\rho f\|_2^2 = \log \sum_{i=0}^{n} a_i (2^{\lambda} - 1)^i$. On the other hand, by properties of linear codes it can be checked that

$$\mathbb{E}(f|S)(x) = \begin{cases} \cdot 2^{|S| - \dim V_S^{\perp}} & \exists y \in V^{\perp} \text{ s.t. } x_S = y_S \\ 0 & \textit{otherwise} \end{cases} ,$$

and that $\Pr_x[\exists y \in C^{\perp} \text{ s.t. } x_S = y_S] = 2^{\dim V_S^{\perp} - |S|}$. Accordingly, $\log \|\mathbb{E}(f|S)\|_2^2 = |S| - \dim V_S^{\perp}$. The inequality in (8) follows by substituting on both sides of Theorem 2.17. We still need to justify the two equalities in (8). The first one follows immediately from Lemma A.1. For the second equality, observe that $S \sim 1 - \lambda$ has the same distribution as the set of non-erased coordinates over BEC(λ). Then, note that given $Y = y$ with non-erased coordinates in S, there are $2^{k - \dim V_S}$ equally likely possibilities for decoding, and therefore $H(X|Y = y) = k - \dim V_S$. □

B THE BHATTACHARYYA BOUND

In this section we prove the well-known Bhattacharyya bound.

Proof of Theorem 3.3. We are analyzing the error probability of the block-MAP decoder for code V on the BMS channel $\mathcal{W}$. Since the code is linear and the channel symmetric, this is equal to the probability that the MAP decoder fails conditioned on the transmitted all-zeros codeword 0^n. Let Y be the output of the channel assuming the all-zeros codeword was transmitted.

We analyze the likelihood ratio between 0^n and another fixed codeword $x \in C$. Without loss of generality assume that $x = 1^i 0^{n-i}$ for some $0 < i \le n$. Assuming the decoder observes $y \in \mathcal{Y}^n$, the respective likelihood ratio is then[5]

$$\frac{\mathcal{W}(x|y)}{\mathcal{W}(0^n|y)} = \frac{\mathcal{W}(y|x)}{\mathcal{W}(y|0^n)} = \prod_{j=1}^{i} \frac{\mathcal{W}(y_j|1)}{\mathcal{W}(y_j|0)} .$$

Let $1 \le j \le i$ and define a random variable L_j equal to the likelihood ratio $\mathcal{W}(Y_j|1)/\mathcal{W}(Y_j|0)$ conditioned on all-zeros codeword. Recall the characterization of $\mathcal{W}$ as a mixture of BSC channels from Section 2, and let $Y_j = (P_j, X_j')$. Observe that, conditioned on P_j, the likelihood ratio L_j is equal to $(1 - P_j)/P_j$ with probability P_j and $P_j/(1 - P_j)$ with probability $1 - P_j$. Accordingly,

$$\mathbb{E}\left[\sqrt{L_j} \mid P_j\right] = 2\sqrt{P_j(1 - P_j)} , \qquad \mathbb{E}\sqrt{L_j} = Z(\mathcal{W}) .$$

Let $P_B(x)$ denote the probability that the MAP decoder concludes that $x \in C$ was more likely to be transmitted than 0^n. By the considerations above and independence of L_j,

$$P_B(x) \le \Pr\left[\prod_{j=1}^{i} L_j \ge 1\right] = \Pr\left[\prod_{j=1}^{i} \sqrt{L_j} \ge 1\right]$$
$$\le \mathbb{E}\left[\sqrt{L_j}\right]^i = Z(\mathcal{W})^{|x|} .$$

Finally, applying the union bound,

$$P_{\mathrm{B}}(\mathcal{W}, V) \le \sum_{x \in C, x \ne 0^n} P_B(x) \le \sum_{i=1}^{n} a_i Z(\mathcal{W})^i . \qquad \square$$

C AN UPPER BOUND ON THE WEIGHT DISTRIBUTION

Proof of Proposition 3.20. For this proof, let $f = 1_V$. In that case one checks that $\widehat{f} = 2^{k-n} \cdot 1_{V^{\perp}}$. Furthermore, let $g(x) = \theta^{|x|}$ and verify that $\widehat{g}(y) = \frac{1}{2^n}(1 - \theta)^{|y|}(1 + \theta)^{n - |y|}$.

Let $(a_0, \ldots, a_n)$ be the weight distribution of V. We calculate,

$$\frac{1}{2^n} \sum_{i=0}^{n} a_i \theta^i = \mathbb{E}_x f(x) g(x) = \sum_y \widehat{f}(y)\widehat{g}(y)$$
$$= \frac{1}{2^{n-k}} \frac{1}{2^n} \sum_{i=0}^{n} b_i (1 - \theta)^i (1 + \theta)^{n-i}$$

and consequently

$$\sum_{i=0}^{n} a_i \theta^i = \frac{1}{|V^{\perp}|} \sum_{i=0}^{n} b_i (1 - \theta)^i (1 + \theta)^{n-i} . \tag{9}$$

Substituting (9) into Theorem 3.1, we have for every $0 \le i \le n$

$$b_i \le 2^{H(X|Y)} \frac{|V^{\perp}|}{(1 - \theta)^i (1 + \theta)^{n-i}} .$$

This already establishes the result for $0 \le i < \frac{1-\theta}{2} n$. In fact, since the left-hand side of (1) is monotone in θ, we also have

$$b_i \le 2^{H(X|Y)} \frac{|V^{\perp}|}{(1 - \alpha)^i (1 + \alpha)^{n-i}}$$

[5] We are abusing notation here, but the meaning of $\mathcal{W}(y|x)/\mathcal{W}(y|0^n)$ should be clear, at least for discrete channels and channels with distributions that have densities.

for any $0 \le \alpha \le \theta$. If $\frac{1-\theta}{2}n \le i \le \frac{n}{2}$, then we take $\alpha = 1 - \frac{2i}{n} \le \theta$ and check that

$$\frac{1}{(1-\alpha)^i(1+\alpha)^{n-i}} = \frac{2^{h_2(i/n)n}}{2^n} ,$$

therefore we have proved our statement for $0 \le i \le \frac{n}{2}$. To deal with the case $\frac{n}{2} < i \le n$, we invoke the calculation at the end of the proof of Proposition 1.6 in [22] to see that

$$\sum_{i=0}^{n} b_{n-i}(1-\theta)^i(1+\theta)^{n-i} \le \sum_{i=0}^{n} b_i(1-\theta)^i(1+\theta)^{n-i}$$

and apply the argument above to b_{n-i} with $i \le \frac{n}{2}$. □

D AN OPTIMIZATION PROBLEM

We are interested in the following simple optimization problem for fixed $p, \beta \in (0,1)$,

$$\max_{1 \le t \le \frac{\beta}{p}} (p - \beta/t)^t$$

Let $f(t) = (p - \beta/t)^t = e^{\ln(p-\beta/t)t}$. Then we are looking for the global maximum of $f(t)$ in $[1, \infty)$. Differentiating $f(t)$ we get

$$f'(t) = e^{t\ln(p-\beta/t)} \cdot \left(\ln(p-\beta/t) + \frac{\beta/t}{p-\beta/t}\right) \tag{10}$$

The first term is positive so we focus on the term inside the parenthesis. Setting $s = p - \beta/t$ and equating the left term to zero we get the equation $\ln(s)s + p - s = 0$. The solution to this equation is $s_{opt} = e^{W_{-1}(-p/e)+1} = -\frac{p}{W_{-1}(-p/e)}$ where $W_{-1}(z)$ is the Lambert function. Note that the requirement $s \ge 0$, or equivalently $t \le \frac{\beta}{p}$, is implicit if we take the real solution if this equation. Thus, $f(t)$ obtains its unique maximum, i.e $t_{opt} = \frac{\beta}{p - s_{opt}}$. Note that $t\ln(s) = -\beta/s$ and so in fact $f(t) = e^{-\beta/s}$ hence

$$f(t_{opt}) = e^{-\beta \cdot e^{-W_{-1}(-p/e)-1}} = \left(\frac{p/e}{-W_{-1}(p/e)}\right)^{\beta/p}$$

It remains to check when $t_{opt} \ge 1$ which happens precisely when $\beta \ge p + \frac{p}{W_{-1}(-p/e)}$. If $t_{opt} < 1$ then the solution to our optimization problem is simply attained at $t = 1$.