On the Worst-Case Complexity of Gibbs Decoding for Reed–Muller Codes

Xuzhe Xia
University of British Columbia
Vancouver, Canada
xia2019@student.ubc.ca

Nicholas Kwan
University of Toronto
Toronto, Canada
nick.kwan@mail.utoronto.ca

Lele Wang
University of British Columbia
Vancouver, Canada
lelewang@ece.ubc.edu

***Abstract*—THIS PAPER IS ELIGIBLE FOR THE STUDENT PAPER AWARD. Reed–Muller (RM) codes are known to achieve capacity on binary symmetric channels (BSC) under the Maximum a Posteriori (MAP) decoder. However, it remains an open problem to design a capacity achieving polynomial-time RM decoder. Due to a lemma by Liu, Cuff, and Verdú, it can be shown that decoding by sampling from the posterior distribution is also capacity-achieving for RM codes over BSC. The Gibbs decoder is one such Markov Chain Monte Carlo (MCMC) based method, which samples from the posterior distribution by flipping message bits according to the posterior, and can be modified to give other MCMC decoding methods. In this paper, we analyze the mixing time of the Gibbs decoder for RM codes. Our analysis reveals that the Gibbs decoder can exhibit slow mixing for certain carefully constructed sequences. This slow mixing implies that, in the worst-case scenario, the decoder requires super-polynomial time to converge to the desired posterior distribution.**

I. Introduction

Reed–Muller (RM) codes were invented in 1954 [1], with the first decoding algorithm invented and published by Reed in the same year [2]. Being among the first codes ever constructed, RM codes have been the subject of extensive study in both electrical engineering and theoretical computer science, especially since the introduction of the closely-related, widely-adopted polar codes. Despite this, many questions about the performance of RM codes have only been answered recently. In 2017, Kudekar et al. showed that the Maximum a Posteriori (MAP) Decoding of RM codes achieves capacity on the binary erasure channel (BEC) [3]. Then, in 2021, Reeves and Pfister showed that, under MAP decoding, RM codes have vanishing bit-error probability on binary memoryless symmetric (BMS) channels [4]. Finally, in 2023, Sandon and Abbe extended the aforementioned result to show that RM codes have vanishing block-error probability on BMS channels, providing a positive answer to the question of whether RM codes achieve capacity on BMS channels [5]. The natural next step in this line of inquiry is to try to establish capacity-achieving performance for a polynomial-time RM decoder. This would help bridge the gap between the existing theory developed by [3]–[5] and practical use. Indeed, no existing polynomial-time decoders, such as Reed's original decoder, for RM codes are known to achieve the same block-error rate performance as MAP decoding for sequences of RM codes that achieve capacity over the BSC. For sequences of RM codes of fixed order, where the number of codewords is polynomial in the blocklength–so that MAP decoding can be done in polynomial time, it is known that there are efficient algorithms, such as the Fast Hadamard Transform [6] for order-1 RM codes, that achieve the same block-error rate performance as MAP decoding. Other practical decoders, such as the Recursive Projection Aggregation algorithm [7], Dumer's recursive decoding algorithms [8], and Sakkour's algorithm [9], have focused on the short-to-medium blocklength regime and particular ranges of r, and either lack time-complexity guarantees in the asymptotic setting, or error-rate guarantees for longer blocklengths.

A. Decoding by Posterior Sampling

In this paper, we study a specific type of decoding algorithm based on posterior sampling. To state our problem setting precisely, we consider the channel coding problem, where some message $\mathbf{m} \in \mathcal{M}$ is encoded as $\mathbf{x}(\mathbf{m}) \in \mathcal{X}^n$ and then transmitted over some discrete memoryless channel $p(y|x)$. Assuming a uniform prior on the set of messages and given a received sequence $\mathbf{y} \in \mathcal{Y}^n$, the MAP decoder produces an $\hat{\mathbf{m}} \in \mathcal{M}$ according to the rule $\hat{\mathbf{m}} = \arg\max_{\mathbf{m}' \in \mathcal{M}} p(\mathbf{x}(\mathbf{m})|\mathbf{y})$. This is optimal in the sense that the probability of error $P_e = \Pr[\mathbf{m} \neq \hat{\mathbf{m}}]$ is minimized. By a Lemma of Liu, Cuff, and Verdú [10], however, it turns out that the asymptotic performance of MAP decoding can be matched by sampling from the posterior $p(\mathbf{m}|\mathbf{y})$, rather than selecting the message with the largest probability mass. This is formalized in the below statement.

Lemma 1 ([10])**.** *Let P_e^* be the optimal probability of error (ie. the one given by MAP decoding), and let $M' \sim p(\mathbf{m}|\mathbf{y})$. Then*

$$P_e^* \leq \Pr[M' \neq \mathbf{m}] \leq 2P_e^*.$$

Thus, if a sequence of codes (such as RM codes) is capacity-achieving under MAP decoding, then they are also capacity achieving under posterior sampling.

It still remains to sample from the posterior distribution, which is often computationally infeasible to do directly, since the distributions in question are often high-dimensional and very nonuniform. We can instead sample through Markov Chain Monte Carlo (MCMC) methods, which consider a Markov Chain whose stationary distribution is the desired posterior $p(\mathbf{m}|\mathbf{y})$. If such a Markov chain is ergodic, then it will converge to the desired posterior. The usual measure of time complexity for such a Markov chain is its mixing

time, which is the number of steps required to ensure that the resulting distribution is within some ϵ Total Variation (TV) distance of the stationary distribution, regardless of the initial distribution. If we take enough steps in a Markov chain to get within some $\epsilon \to 0$ TV distance of the posterior, then by the Liu-Cuff-Verdú lemma, posterior sampling through MCMC methods is capacity achieving if the underlying code is capacity achieving.

Using MCMC methods to decode was first proposed by Neal in 2001 [11]. The proposed algorithm considered Gibbs sampling on the codeword space, with some proposed modifications to decrease the mixing time of the decoder, and to ensure that the resulting Markov Chain was indeed ergodic. Huang explored, in his PhD Thesis [12], many MCMC decoders in the short blocklength regime, based on running a Markov Chain on the message space. Many of these schemes are based on a basic MCMC decoder, known as the Gibbs decoder or Glauber dynamics, where a message bit is chosen uniformly at random, and then resampled according to the marginal posterior obtained by fixing the other bits to their assigned values. In this paper we provide an analysis of the the mixing time of the Gibbs decoder.

B. Main Contribution

Our results demonstrate that for specific, carefully constructed channel output sequences, the Gibbs decoder exhibits slow mixing. This implies that, in the worst-case scenario, the Gibbs decoder may struggle to converge in polynomial time to the desired posterior distribution when decoding RM codes over the binary symmetric channel.

C. Organization

In the Preliminaries, we provide background knowledge and key results relevant to our main findings. Specifically, we introduce Reed–Muller codes in Section II-A, Markov Chains, Mixing Time, and related results in Section II-B, and define Gibbs Decoding in Section II-C. Our main result, which we present as Theorem 3, addresses the worst-case mixing time analysis. The proof of this theorem is detailed in Section IV.

II. Preliminaries

A. Reed–Muller (RM) codes

We begin by establishing the notation for RM codes. RM codes belong to the family of polynomial codes. We write $\mathbb{F}_2[z_1, \ldots, z_m]$ for the ring of m-variate polynomials over the finite field $\mathbb{F}_2$. Define the *evaluation vector* of the polynomial $f \in \mathbb{F}_2[z_1, \ldots, z_m]$, $\mathrm{Eval}(f)$, as a binary vector whose coordinates corresponding to the evaluations of f at all 2^m vectors in $\mathbb{F}_2^m$, usually in lexicographic ordering. The *RM code* $\mathrm{RM}(r, m)$ is defined by a $k \times n$ generator matrix G, where each row of G is the evaluation vector of a monomial $f \in \mathbb{F}_2[z_1, \ldots, z_m]$ of degree at most r. Hence, the RM code $\mathrm{RM}(r, m)$ is explicitly

$$\mathrm{RM}(r, m) = \{\mathrm{Eval}(f) : f \in \mathbb{F}_2\left[z_1, \ldots, z_m\right], \deg(f) \leq r\}.$$

For example, $\mathrm{RM}(2, 3)$ is defined by the following generator matrix:

$$G = \begin{bmatrix} \mathrm{Eval}\left(z_1 z_2\right) \\ \mathrm{Eval}\left(z_1 z_3\right) \\ \mathrm{Eval}\left(z_2 z_3\right) \\ \mathrm{Eval}\left(z_1\right) \\ \mathrm{Eval}\left(z_2\right) \\ \mathrm{Eval}\left(z_3\right) \\ \mathrm{Eval}(1) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

Since $z^2 = z$ for all $z \in \mathbb{F}_2$ it suffices to consider polynomials whose terms have degree 1 in each indeterminate. Consequently, the dimension k and length n satisfy

$$k = \binom{m}{\leq r} := \sum_{i=0}^{r} \binom{m}{i} \text{ and } n = 2^m.$$

The rate of the RM code $\mathrm{RM}(r, m)$ is given by $R = 2^{-m}\binom{m}{\leq r}$, and the minimum distance between distinct codewords is 2^{m-r}.

Consider transmission over a Binary Symmetric Channel (BSC) with crossover probability p. Let $\theta = \frac{p}{1-p}$ and assume that messages $\mathbf{m} \in \mathbb{F}_2^k$ are generated uniformly at random. Denote by $\mathbf{y} \in \mathbb{F}_2^n$ the received sequence and $d_H(\mathbf{x}, \mathbf{y})$ the Hamming distance between binary vectors $\mathbf{x}$ and $\mathbf{y}$. The posterior probability of a message $\mathbf{m}$ given the received sequence $\mathbf{y}$ is expressed as

$$\begin{aligned} p(\mathbf{m} \mid \mathbf{y}) &= \frac{p(\mathbf{y} \mid \mathbf{m}) p(\mathbf{m})}{p(\mathbf{y})} \\ &= \frac{p^{d_H(\mathbf{m}G, \mathbf{y})} \cdot (1-p)^{n - d_H(\mathbf{m}G, \mathbf{y})}}{2^k p(\mathbf{y})}, \\ &= \theta^{d_H(\mathbf{m}G, \mathbf{y})} \cdot \frac{(1-p)^n}{2^k p(\mathbf{y})}. \end{aligned}$$

This implies that the posterior probability of $\mathbf{m}$ given $\mathbf{y}$ is proportional to $\theta^{d_H(\mathbf{m}G, \mathbf{y})}$:

$$p(\mathbf{m} \mid \mathbf{y}) \propto \theta^{d_H(\mathbf{m}G, \mathbf{y})}.$$

B. Markov Chain and Mixing Time

Now, we introduce some basic properties of Markov Chains. Let P denote the transition matrix of a Markov chain on a finite state space Ω and P^t denote the t-fold matrix multiplication. Let $P(s_1 \to s_2)$ represent the transition probability from state s_1 to state s_2. We say P is *ergodic* if it satisfies the following conditions:

- (Irreducibility) For all $s_1, s_2 \in \Omega$, there exists a non-negative integer t such that $P^t(s_1 \to s_2) > 0$, and,
- (Aperiodicity) For all $s \in \Omega$, the greatest common divisor of the set $\{t \geq 1 \colon P^t(s \to s) > 0\}$ is equal to 1.

A distribution μ on Ω is said to be *stationary* with respect to P if $\mu P = \mu$. The following theorem gives a relationship between ergodicity and the uniqueness of, and convergence to, a stationary state:

Theorem 1 (Corollary 1.17, Theorem 4.9 [13])**.** *Let P be an ergodic Markov chain on a finite state space Ω. Then P*

possesses a unique stationary distribution μ on Ω, and for any initial distribution μ_0, the distribution $\mu_t = \mu_0 P^t$ of the state X_t converges pointwise to μ as $t \to \infty$.

Let $\epsilon > 0$. The *ϵ-mixing time of P* is defined as

$$T_{\text{mix}}(\epsilon) := \sup_{\mu_0} \min \left\{ t \geq 0 : \left\| \mu_0 P^t - \mu \right\|_{\text{TV}} \leq \epsilon \right\},$$

where the supremum is taken over all possible initial distributions μ_0. Here, $\| \cdot \|_{\text{TV}}$ represents the total variation distance between distributions.

We further define the *total variation mixing time*, or *mixing time* for short, of P as

$$T_{\text{mix}} := T_{\text{mix}}\left(\tfrac{1}{4}\right).$$

To analyze the slow mixing, it suffices to analyze the total variation mixing time, as this provides a lower bound for the mixing time for any $\epsilon \to 0$.

Next we define the *bottleneck ratio*, also known as *conductance*. This quantity provides a means of bounding the mixing time of a Markov Chain from below and is instrumental in analyzing slow mixing phenomena. Let P be an ergodic Markov chain with stationary distribution μ on a state space Ω. The *bottleneck ratio of a subset $\mathcal{S} \subseteq \Omega$* is defined as

$$\Phi(\mathcal{S}) := \frac{\sum_{s_1 \in \mathcal{S}, s_2 \in \Omega \setminus \mathcal{S}} \mu(s_1) P(s_1 \to s_2)}{\sum_{s_1 \in \mathcal{S}} \mu(s_1)},$$

where μ denotes the stationary distribution of P. The *bottleneck ratio of the Markov chain P* is then given by taking the infimum of the bottleneck ratio over all subsets of Ω with probability at most $\frac{1}{2}$:

$$\Phi(P) := \inf_{\substack{\mathcal{S} \in \Omega \\ \mu(S) \leq \frac{1}{2}}} \Phi(\mathcal{S}).$$

We can use the bottleneck ratio of a Markov chain P to bound its mixing time.

Theorem 2 (Theorem 7.4 [13])**.** *The mixing time T_{mix} of P satisfies*

$$T_{\text{mix}} \geq \frac{1}{4\Phi(P)}.$$

C. Gibbs Decoding

The Gibbs decoder is an MCMC method that recovers the message by sampling the message space $\mathbb{F}_2^k$ according to the posterior distribution $p(\mathbf{m}|\mathbf{y})$. For any message $\mathbf{m} \in \mathbb{F}_2^k$, write $\mathbf{m}^{\oplus i}$ for the message obtained by flipping the i-th coordinate of $\mathbf{m}$ (modulo 2), while keeping all other coordinates unchanged. Let μ be the posterior distribution on $\mathbb{F}_2^k$, that is $\mu(\mathbf{m}) = p(\mathbf{m} \mid \mathbf{y})$, where $\mathbf{y}$ is the received sequence. In each iteration, the Gibbs sampler performs the following:

1) Select a coordinate $i \in [k]$ uniformly at random.
2) Set m_i to $m_i \oplus 1$ with probability $\frac{\mu(\mathbf{m}^{\oplus i})}{\mu(\mathbf{m}) + \mu(\mathbf{m}^{\oplus i})}$.

Thus, the transition probabilities are given by

$$P\left(\mathbf{m} \to \mathbf{m}^{\oplus i}\right) = \frac{1}{k} \cdot \frac{\mu\left(\mathbf{m}^{\oplus i}\right)}{\mu(\mathbf{m}) + \mu\left(\mathbf{m}^{\oplus i}\right)}, \quad \forall i \in [k],$$

$$P(\mathbf{m} \to \mathbf{m}) = 1 - \sum_{i=1}^{k} P\left(\mathbf{m} \to \mathbf{m}^{\oplus i}\right),$$

and for all the other $\mathbf{m}'$, we have $P(\mathbf{m} \to \mathbf{m}') = 0$. Since

$$\begin{aligned} \mu(\mathbf{m}) \cdot P\left(\mathbf{m} \to \mathbf{m}^{\oplus i}\right) &= \frac{1}{k} \cdot \frac{\mu(\mathbf{m}) \cdot \mu\left(\mathbf{m}^{\oplus i}\right)}{\mu(\mathbf{m}) + \mu\left(\mathbf{m}^{\oplus i}\right)} \\ &= \mu\left(\mathbf{m}^{\oplus i}\right) \cdot P\left(\mathbf{m}^{\oplus i} \to \mathbf{m}\right), \end{aligned}$$

we have $\mu P = \mu$, which implies that μ is a stationary distribution of P. To show the uniqueness of μ and convergence to μ, it suffices to show, by Theorem 1, that P is ergodic.

Lemma 2. *The Gibbs Sampler P defined through the posterior $p(\mathbf{m}|\mathbf{y})$ above is ergodic, for any received sequence $\mathbf{y} \in \mathbb{F}_2^n$.*

Proof: First, note that for any $\mathbf{y} \in \mathbb{F}_2^n$, $\mu(\mathbf{m}) > 0$ for all $\mathbf{m} \in \mathbb{F}_2^k$, so that $0 < P(\mathbf{m} \to \mathbf{m}^{\oplus i}) < 1/k$ for all $\mathbf{m} \in \mathbb{F}_2^k$ and $i \in [k]$. To see that P is aperiodic, note that $P(\mathbf{m} \to \mathbf{m}^{\oplus i}) < 1/k$ for all $i \in [k]$ and $\mathbf{m} \in \mathbb{F}_2^k$. Hence, $\sum_{i=1}^{k} P(\mathbf{m} \to \mathbf{m}^{\oplus i}) < 1$ and $P(\mathbf{m} \to \mathbf{m}) > 0$ for all $\mathbf{m} \in \mathbb{F}_2^k$. Thus, for each $\mathbf{m} \in \mathbb{F}_2^k$, $1 \in \{t \geq 1 : P^t(\mathbf{m} \to \mathbf{m}) > 0\}$ so that $\gcd\{t \geq 1 : P^t(\mathbf{m} \to \mathbf{m}) > 0\} = 1$. To see that P is irreducible, note that $P(\mathbf{m} \to \mathbf{m}^{\oplus i}) > 0$ for all $i \in [k]$ and $\mathbf{m} \in \mathbb{F}_2^k$, so that for any $\mathbf{m}, \mathbf{m}' \in \mathbb{F}_2^k$, and any $j \geq d_H(\mathbf{m}, \mathbf{m}')$, $P^j(\mathbf{m} \to \mathbf{m}') > 0$. ∎

III. Main Results

Recall that the set of conditionally ε-typical length n sequences for $\text{BSC}(p)$ with input $\mathbf{x}$ is defined as

$$\mathcal{T}_{\varepsilon}^{(n)}(Y|\mathbf{x}) = \{\mathbf{y} : np(1-\varepsilon) \leq d_H(\mathbf{x}, \mathbf{y}) \leq np(1+\varepsilon)\}.$$

We prove the following theorem for the worst-case performance of the Gibbs Decoder.

Theorem 3. *Given any crossover probability $p < \frac{1}{2}$ and rate $R < 1 - H(p)$, for any sequence of RM codes $\text{RM}(r_j, m_j)$ satisfying $\lim_{j\to\infty} m_j = \infty$ and $\lim_{j\to\infty} R_j = R$, there exists a sequence of pairs $\{(\mathbf{m}_j G_j, \mathbf{y}_j)\}$ satisfying*

1) *$\mathbf{y}_j \in \mathcal{T}_{1/2}^{(n_j)}(Y|\mathbf{m}_j G_j)$ and*
2) *the mixing time for running the Gibbs decoder on $\mathbf{y}_j$ using $\text{RM}(r_j, m_j)$ is*

$$T_{\text{mix}} = \Omega(\exp(\sqrt{n} \cdot \exp(-\sqrt{\log(n)}))).$$

Remark 1. This result suggests that the worst-case decoding complexity is superpolynomial. Although the $\mathbf{y}$ we construct is a typical received sequence for some codeword, the probability of receiving such $\mathbf{y}$ might be vanishing. Thus, it may be the case that the average-case decoding complexity for the Gibbs decoder is polynomial. Numerical evidence, however, suggests that the average-case convergence rate of the Gibbs decoder is also slow [12, Section 3.3.3].

Remark 2. We also note that there are several variations of the Gibbs decoder that show empirically faster convergence to the posterior. For instance, numerical simulations have shown that block updates (i.e. updating multiple bits in each iteration) decreases the number of iterations required to converge (although

sampling from the conditional posterior for multiple bits is harder than for a single bit). Another method of improving the convergence is to consider an annealing schedule, so that at the t-th iteration, we consider the Gibbs update given by the distribution $p_{\alpha(t)}(\mathbf{m}|\mathbf{y}) \propto p(\mathbf{m}|\mathbf{y})^{\alpha(t)}$ for some $\alpha(t) \in [0,1]$. These methods are covered in greater detail in [14], where the authors run multiple annealed block Gibbs decoders on RM codes, on different subsets of $\mathbb{F}_2^k$ in parallel, to improve the mixing time. While our mixing time analysis seems to rule out initializing Gibbs decoding at an arbitrary distribution for RM codes, [15] considers using the Gibbs decoder after performing belief propagation on LDPC codes. As post-processing, this is able to overcome the error-floor phenomenon associated with belief propagation for some codes.

IV. Proof of Theorem 3

The proof of Theorem 3 is divided into two key steps. In the first step, Lemma 3 provides a sufficient condition for the Gibbs decoder to exhibit slow mixing. If there exists a message $\mathbf{m}$ with posterior probability $\mu(\mathbf{m}) \leq \frac{1}{2}$, and flipping any bit of $\mathbf{m}$ results in a significant increase in the Hamming distance to the received sequence $\mathbf{y}$, then the decoder struggles to move away from $\mathbf{m}$. The rationale is that messages around $\mathbf{m}$ have much smaller posterior probabilities, effectively "trapping" the decoder near $\mathbf{m}$. As a result, the decoder cannot explore the space of messages with significant posterior probability, leading to slow mixing.

In the second step, we explicitly construct such a typical received sequence $\mathbf{y}$ for any RM code $\mathrm{RM}(r,m)$ with sufficiently large m in Lemma 4. Specifically, when the current message is the all-zero message $\mathbf{0}$, we construct a received sequence $\mathbf{y}$ that ensures that flipping any bit of the all-zero message results in a significant increase in the Hamming distance to $\mathbf{y}$.

Lemma 3. *Consider the Reed–Muller code $\mathrm{RM}(r,m)$ with generator matrix G. Let $\mathbf{y} \in \mathbb{F}_2^n$. If there exists a message $\mathbf{m} \in \mathbb{F}_2^k$ such that $\mu(\mathbf{m}) \leq \frac{1}{2}$, and*

$$\delta_i(\mathbf{m}) := d_H\left(\mathbf{m}^{\oplus i}G, \mathbf{y}\right) - d_H(\mathbf{m}G, \mathbf{y}) = \Omega(f(n)), \ \ \forall i \in [k],$$

then the mixing time of the corresponding Gibbs decoder on $\mathbf{y}$ is $T_{\mathrm{mix}} = \Omega\left(e^{f(n)}\right)$.

Lemma 4. *Given a crossover probability $p < \frac{1}{2}$, let $q = \left\lceil \log_2 \frac{1}{p} \right\rceil$. For the RM code $\mathrm{RM}(r,m)$ with $m \geq q$, there exists a non-zero vector $\mathbf{y} \in \mathbb{F}_2^m$ such that:*

1) *There exists a codeword $\mathbf{c}$ such that $d_H(\mathbf{c}, \mathbf{y}) = 2^{m-q}$;*
2) *The posterior distribution for the all-zero message given received sequence $\mathbf{y}$ satisfies $\mu(\mathbf{0}) \leq 0.5$; and*
3) *Adding any row $\mathbf{g}_i$ of the generator matrix[1] of $\mathrm{RM}(r,m)$ to $\mathbf{y}$ increases its weight by at least $2^{m-r-q+1}$, i.e.,*

$$\mathrm{wt}(\mathbf{y} + \mathbf{g}_i) - \mathrm{wt}(\mathbf{y}) \geq \frac{d}{2^{q-1}} = 2^{m-r-q+1}, \quad \forall i \in [k].$$

[1]We consider the standard generator matrix where each row corresponds to a monomial of degree at most r as introduced in Section II-A.

The key idea in establishing Lemma 3 is to identify a singleton message set $\mathcal{S}$, whose bottleneck ratio gives a lower bound of the bottleneck ratio of P.

Proof of Lemma 3: For a subset of messages $\mathcal{S} \subseteq \mathbb{F}_2^k$ with $\mu(\mathcal{S}) \leq \frac{1}{2}$, the bottleneck ratio can be expressed as

$$\Phi(\mathcal{S}) = \frac{\sum_{\mathbf{m}\in\mathcal{S}} \mu(\mathbf{m})[1 - P(\mathbf{m} \to \mathbf{m})]}{\mu(\mathcal{S})}.$$

Consider the special case where $\mathcal{S} = \{\mathbf{m}\}$ is a singleton set containing only one state $\mathbf{m}$, such that $\mu(\mathbf{m}) \leq \frac{1}{2}$. In this scenario, the bottleneck ratio simplifies to

$$\Phi(\mathcal{S}) = 1 - P(\mathbf{m} \to \mathbf{m}).$$

Define $\delta_i(\mathbf{m}) := d_H\left(\mathbf{m}^{\oplus i}G, \mathbf{y}\right) - d_H(\mathbf{m}G, \mathbf{y})$, where d_H denotes the Hamming distance. Then, the probability $1 - P(\mathbf{m} \to \mathbf{m})$ can be written as

$$1 - P(\mathbf{m} \to \mathbf{m}) = \frac{1}{k}\sum_{i=1}^{k} \frac{\mu\left(\mathbf{m}^{\oplus i}\right)}{\mu(\mathbf{m}) + \mu\left(\mathbf{m}^{\oplus i}\right)} = \frac{1}{k}\sum_{i=1}^{k} \frac{\frac{\mu\left(\mathbf{m}^{\oplus i}\right)}{\mu(\mathbf{m})}}{1 + \frac{\mu\left(\mathbf{m}^{\oplus i}\right)}{\mu(\mathbf{m})}}.$$

Using the inequality $\frac{x}{1+x} \leq x$ for $x \geq 0$, we obtain

$$1 - P(\mathbf{m} \to \mathbf{m}) \leq \frac{1}{k}\sum_{i=1}^{k} \frac{\mu\left(\mathbf{m}^{\oplus i}\right)}{\mu(\mathbf{m})}.$$

From the proportionality relation on the posterior, we may write the ratio of posteriors as $\theta^{\delta_i(\mathbf{m})} = \frac{\mu\left(\mathbf{m}^{\oplus i}\right)}{\mu(\mathbf{m})}$, where $\theta = \frac{p}{1-p} < 1$ is a parameter given by the channel. Substituting this, we get

$$1 - P(\mathbf{m} \to \mathbf{m}) \leq \frac{1}{k}\sum_{i=1}^{k} \theta^{\delta_i(\mathbf{m})}.$$

From the lower bound Theorem 2 on mixing time, we know

$$T_{\mathrm{mix}} \geq \frac{1}{4\Phi(P)} = \frac{1}{4(1 - P(\mathbf{m} \to \mathbf{m}))}.$$

Therefore, if $\delta_i(\mathbf{m}) = \Omega(f(n))$ for all i, then

$$1 - P(\mathbf{m} \to \mathbf{m}) = O(\exp(-f(n)))$$

and consequently $T_{\mathrm{mix}} = \Omega(\exp(f(n)))$. ∎

Proof of Lemma 4: Recall that in an RM Code, each codeword is the evaluation vector of an m-variate polynomial $f \in \mathbb{F}_2[z_1, z_2, \ldots, z_m]$ of total degree at most r over all 2^m vectors $\mathbf{z} = (z_1, z_2, \ldots, z_m) \in \mathbb{F}_2^m$. Define a m-variate polynomial

$$f_{\mathbf{y}}(z_1, \ldots, z_m) = \left(\prod_{i=1}^{q-1} z_i + 1\right)(z_m + 1). \tag{1}$$

Let $\mathbf{y} = \mathrm{Eval}(f_{\mathbf{y}})$ denote the evaluation vector of $f_{\mathbf{y}}$. Since $f_{\mathbf{y}}(\mathbf{0}) = 1$, it follows that $\mathbf{y} \neq \mathbf{0}$.

Next, consider the evaluation vector of the m-variate polynomial $z_m + 1$, denoted as $\mathbf{c} \triangleq \mathrm{Eval}(z_m + 1)$. Since this is a degree-1 polynomial, its evaluation vector $\mathbf{c}$ is a codeword for any RM code $\mathrm{RM}(r,m)$ with $r \geq 1$. Let $\mathbf{u}$ be the corresponding message, i.e., $\mathbf{c} = \mathbf{u}G$. The vector $\mathbf{c}$ differs from $\mathbf{y}$ only at the coordinates $\mathbf{z}$ with $z_1 = z_2 = \cdots =$

$z_{q-1} = 1$ and $z_m = 0$. With q variables fixed, there are 2^{m-q} such coordinates. Therefore, the Hamming distance $d_H(\mathbf{c}, \mathbf{y}) = 2^{m-q}$, establishing statement 1) in Lemma 4.

To establish statement 2), we analyze the posterior probabilities of the all-zero message and message $\mathbf{u}$. The posterior probability of the all-zero message $\mathbf{0}$ is proportional to $\theta^{\mathrm{wt}(\mathbf{y})}$. To compute $\mathrm{wt}(\mathbf{y})$, note that $f_{\mathbf{y}}(\mathbf{z})$ is 1 only when $z_m = 0$ and $\prod_{i=1}^{q-1} z_i = 0$, which implies $\mathrm{wt}(\mathbf{y}) = 2^{m-1} - 2^{m-q}$. Thus, the posterior probability of the all-zero message is proportional to $\theta^{2^{m-1}-2^{m-q}}$. Since $d_H(\mathbf{c}, \mathbf{y}) = 2^{m-q}$, the posterior probability of message $\mathbf{u}$ is proportional to $\theta^{2^{m-q}}$. For $p < \frac{1}{2}$, we have $q \geq 2$, which leads to $\theta^{2^{m-1}-2^{m-q}} \leq \theta^{2^{m-q}}$. This implies that the posterior probability $\mu(\mathbf{u}) \geq \mu(\mathbf{0})$. If $\mu(\mathbf{0}) > \frac{1}{2}$, it would result in a contradiction, as the posterior probabilities of both messages would sum to more than 1. Therefore, we conclude that $\mu(\mathbf{0}) \leq \frac{1}{2}$, establishing statement 2).

Finally, to prove statement 3), we define $\mathcal{S}(f) := \{\mathbf{z} \in \mathbb{F}_2^m : f(\mathbf{z}) = 1\}$. Consider an arbitrary row $\mathbf{g}_i$ of G, which corresponds to the evaluation vector of some monomial $f \in \mathbb{F}_2[z_1, z_2, \ldots, z_m]$ with $\deg(f) \leq r$. The weight of $\mathbf{g}_i$ is equal to $|\mathcal{S}(f)|$. Denote $f = z_{j_1} z_{j_2} \cdots z_{j_t}$ where $\{j_1, \ldots, j_t\}$ is some subset of $[m]$ with size $t \leq r$. There are two cases.

Case 1: $m \in \{j_1, \ldots, j_t\}$. In this case, for any evaluation coordinate $\mathbf{z} \in \mathbb{F}_2^m$ such that $f(\mathbf{z}) = 1$, we must have $z_m = 1$. This forces $f_{\mathbf{y}}(\mathbf{z}) = 0$ since $f_{\mathbf{y}} = (\prod_{i=1}^{q-1} z_i + 1)(z_m + 1)$. Thus, $\mathcal{S}(f) \cap \mathcal{S}(f_{\mathbf{y}}) = \emptyset$ and we have

$$\begin{aligned}\mathrm{wt}(\mathbf{y} + \mathbf{g}_i) - \mathrm{wt}(\mathbf{y}) &= |\mathcal{S}(f + f_{\mathbf{y}})| - |\mathcal{S}(f_{\mathbf{y}})| \\ &= |\mathcal{S}(f)| - 2|\mathcal{S}(f) \cap \mathcal{S}(f_{\mathbf{y}})| \\ &= |\mathcal{S}(f)| \geq 2^{m-t} \geq 2^{m-r}.\end{aligned}$$

Case 2: $m \notin \{j_1, \ldots, j_t\}$. Let $\mathcal{I} := \{j_1, \ldots, j_t\} \cap \{1, 2, \ldots, q-1\}$. We have $0 \leq |\mathcal{I}| \leq \min\{t, q-1\}$. Then,

$$\begin{aligned}\mathcal{S}(f) \cap \mathcal{S}(f_{\mathbf{y}}) = \{\mathbf{z} \in \mathbb{F}_2^m : z_{j_1} = \cdots = z_{j_t} = 1, z_m = 0, \\ \text{and at least one of } z_1, \ldots, z_{q-1} \text{ is } 0\}.\end{aligned}$$

Partition the indices into three disjoint sets:

$$\begin{aligned}\mathcal{A} &= \{j_1, \ldots, j_t, m\}, \\ \mathcal{B} &= \{1, 2, \ldots, q-1\} \backslash \mathcal{I}, \\ \mathcal{C} &= [m] \backslash (\mathcal{A} \cup \mathcal{B}).\end{aligned}$$

The number of $\mathbf{z} \in \mathbb{F}_2^m$ where all variables in $\mathcal{A}$ are fixed and at least one variable in $\mathcal{B}$ is 0 is:

$$\begin{aligned}|\mathcal{S}(f) \cap \mathcal{S}(f_{\mathbf{y}})| &= (1)\left(2^{q-1-|\mathcal{I}|} - 1\right)\left(2^{m-t-q+|\mathcal{I}|}\right) \\ &= 2^{m-t-1} - 2^{m-t-q+|\mathcal{I}|}.\end{aligned}$$

Thus, we have

$$\begin{aligned}\mathrm{wt}(\mathbf{y} + \mathbf{g}_i) - \mathrm{wt}(\mathbf{y}) &= |\mathcal{S}(f + f_{\mathbf{y}})| - |\mathcal{S}(f_{\mathbf{y}})| \\ &= |\mathcal{S}(f)| - 2|\mathcal{S}(f) \cap \mathcal{S}(f_{\mathbf{y}})| \\ &= 2^{m-t} - 2\left(2^{m-t-1} - 2^{m-t-q+|\mathcal{I}|}\right) \\ &= 2^{m-t-(q-1-|\mathcal{I}|)} \geq 2^{m-r-q+1}.\end{aligned}$$

■

With the two lemmas established, the final step of the proof involves analyzing the asymptotic expression of $2^{m-r-q+1}$ in terms of the codeword length $n = 2^m$.

Proof of Theorem 3: Let $p < \frac{1}{2}$ and $q = \left\lceil \log_2 \frac{1}{p} \right\rceil$. For any $m \geq q$ and the corresponding RM code $\mathrm{RM}(r, m)$, use Lemma 4 to construct $\mathbf{y} \neq 0$. There exists a codeword $\mathbf{c} \in \mathbb{F}_2^n$ such that $d_H(\mathbf{c}, \mathbf{y}) = 2^{m-q}$. Moreover, since

$$\frac{1}{2}p \leq 2^{-\lceil \log_2(1/p) \rceil} \leq \frac{3}{2}p,$$

there exists a message $\mathbf{m}$ such that $\mathbf{m}G = \mathbf{c}$ and $(\mathbf{m}G, \mathbf{y}) \in \mathcal{T}_{1/2}^{(n)}(Y|\mathbf{m}G)$.

For all $i \in [k]$, consider the i-th row $\mathbf{g}_i$ of G. From Lemma 4,

$$\mathrm{wt}(\mathbf{y} + \mathbf{g}_i) - \mathrm{wt}(\mathbf{y}) \geq 2^{m-r-q+1}.$$

This implies $\delta_i \geq 2^{m-r-q+1}$, where $\mathbf{m}_0 = \mathbf{0} \in \mathbb{F}_2^k$.

To analyze the asymptotic behavior of $2^{m-r-q+1}$, note that the binomial distribution satisfies:

$$\Pr[\mathrm{Bin}(m, 0.5) = r] = \frac{\binom{m}{r}}{2^m}.$$

Let $k = \binom{m}{\leq r}$. For large m, the rate converges to R, i.e., $\frac{k}{2^m} \to R$. Approximating via the normal distribution:

$$\Pr\left[\mathcal{N}\left(\frac{m}{2}, \frac{m}{4}\right) \leq r\right] = \Phi\left(\frac{r - \frac{m}{2}}{\frac{\sqrt{m}}{2}}\right) = R,$$

where Φ is the cumulative distribution function of the standard normal distribution. Let $c := \Phi^{-1}(R)$. Then,

$$r \leq \frac{m}{2} + c \cdot \frac{\sqrt{m}}{2}.$$

Thus,

$$\delta_i \geq 2^{m-r-q} \geq 2^{(m-c\cdot\sqrt{m})/2-q}.$$

Since $n = 2^m$, it follows from Lemma 3 that:

$$T_{\mathrm{mix}} = \Omega(\exp(\sqrt{n} \cdot \exp(-\sqrt{\log(n)}))).$$

■

V. Conclusion

In this paper, we have analyzed the worst-case complexity of the Gibbs decoder on RM codes. We have shown that, in the case of certain conditionally typical received sequences, Gibbs decoding has superpolynomial mixing time. For the posterior given by these received sequences, all-zeros message presents a bottleneck, preventing the Gibbs decoder from exploring the message space. Despite this, the question of whether other MCMC decoders can have polynomial-time complexity is left open.

Acknowledgment

The authors are grateful to Wei Yu for encouraging them to work on this problem and Ryan Song for helpful discussions.

REFERENCES

[1] D. E. Muller, "Application of boolean algebra to switching circuit design and to error detection," *Transactions of the I.R.E. Professional Group on Electronic Computers*, vol. EC-3, no. 3, pp. 6–12, 1954.

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

[3] 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.

[4] G. Reeves and H. D. Pfister, "Reed–muller codes on bms channels achieve vanishing bit-error probability for all rates below capacity," *IEEE Transactions on Information Theory*, vol. 70, no. 2, pp. 920–949, 2024.

[5] E. Abbe and C. Sandon, "A proof that reed-muller codes achieve shannon capacity on symmetric channels," 2023. [Online]. Available: https://arxiv.org/abs/2304.02509

[6] R. R. Green, "A serial orthogonal decoder," in *JPL Space Programs Summary*, vol. 37, 1966, pp. 247–253.

[7] M. Ye and E. Abbe, "Recursive projection-aggregation decoding of reed-muller codes," in *2019 IEEE International Symposium on Information Theory (ISIT)*, 2019, pp. 2064–2068.

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

[9] B. Sakkour, "Decoding of second order reed-muller codes with a large number of errors," in *IEEE Information Theory Workshop, 2005.*, 2005, pp. 3 pp.–.

[10] J. Liu, P. Cuff, and S. Verdú, "On α-decodability and α-likelihood decoder," in *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2017, pp. 118–124.

[11] R. M. Neal, "Monte carlo decoding of LDPC codes," 2001, talk given at ICTP Workshop on Statistical Physics and Capacity-Approaching Codes.

[12] J.-T. Huang, "Markov chain monte carlo methods for detection and decoding," PhD thesis, University of California San Diego, 2023.

[13] D. A. Levin and Y. Peres, *Markov chains and mixing times*. American Mathematical Soc., 2017, vol. 107.

[14] J.-T. Huang and Y.-H. Kim, "Parallel monte carlo markov chain decoding of linear codes," in *2023 IEEE International Symposium on Information Theory (ISIT)*, 2023, pp. 2051–2056.

[15] ——, "MCMC decoding of LDPC codes with BP preprocessing," in *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, 2020, pp. 1–5.