Codes over integers, and the singularity of random matrices with large entries

Sankeerth Rao Karingula*   Shachar Lovett†
Department of Computer Science
University of California, San Diego

October 22, 2020

Abstract

The prototypical construction of error correcting codes is based on linear codes over finite fields. In this work, we make first steps in the study of codes defined over integers. We focus on Maximum Distance Separable (MDS) codes, and show that MDS codes with linear rate and distance can be realized over the integers with a constant alphabet size. This is in contrast to the situation over finite fields, where a linear size finite field is needed.

The core of this paper is a new result on the singularity probability of random matrices. We show that for a random $n \times n$ matrix with entries chosen independently from the range $\{-m, \ldots, m\}$, the probability that it is singular is at most m^{-cn} for some absolute constant $c > 0$.

1 Introduction

Coding theory is the study of error correction codes. Codes are widely used in many applications, such as data compression, cryptography, error detection and correction, data transmission and data storage. Algorithms needed to implement codes perform arithmetic operations over an underlying alphabet, and hence their computational complexity is constrained by this alphabet size. Thus, understanding the alphabet size needed to support a given code structure is a central question in coding theory. By far, the most common approach to design codes is to use linear codes over finite fields. The main focus of this paper is to investigate the possibility of designing codes over integers. In particular, we study the alphabet size needed to support basic code structures, and focus on the most basic and well-studied family of codes - Maximum Distance Separable (MDS) codes.

1.1 Alphabet size for MDS codes

An MDS code is a code with the best possible tradeoff between the message length, codeword length and minimal distance. Concretely, an (n, k, d)-code is a code with message length k, codeword length n and minimal distance d. The Singleton bound [28] gives that $d \leq n - k + 1$. MDS codes are codes achieving this bound, namely (n, k, d)-codes with $d = n - k + 1$. If we consider linear codes, then it is well-known that MDS codes are generated by the row span of *MDS matrices*.

Definition 1.1 (MDS matrix). *Let $n \geq k$. A $k \times n$ matrix is called an* MDS matrix *if any k columns in it are linearly independent. Equivalently, if any $k \times k$ minor of it is nonsingular.*

Note that MDS matrices can be defined over finite fields or over the integers. If we define them over a finite field $\mathbb{F}_q$, then it is well-known that a linear field size is needed to support MDS matrices. Concretely, if we assume $n \geq k + 2$, then it is known that $q \geq \max(k, n - k + 1)$ (see for example the introduction of [2] for a proof). In particular, this implies that $q \geq n/2$. Reed-Solomon codes can be constructed over fields of size $q \geq n - 1$, which is tight up to a factor of two. The MDS conjecture of Segre [27] speculates that this is indeed the best possible (except for a few special cases), and Ball [2] proved this over prime finite fields. In summary, over finite fields a linear field size $q = \Theta(n)$ is both necessary and sufficient.

We show that over the integers, MDS matrices exist over much smaller alphabet sizes.

Theorem 1.2 (MDS matrices over integers). *Let $n \geq k$. There exist $k \times n$ MDS matrices over integers whose entries are in $\{-m, \ldots, m\}$, where $m \leq (cn/k)^c$ for some absolute constant $c > 0$.*

The typical regime in coding theory is that of linear rate and linear distance; namely, where $k = \alpha n$ for some constant $\alpha \in (0, 1)$. Note that in this regime Theorem 1.2 shows that MDS codes over the integers exist with a *constant* alphabet size, which is in stark contrast with the case over finite fields. It is easy to see that Theorem 1.2 is best possible, up to the unspecified constant c.

Claim 1.3. *Let $n \geq k \geq 2$. Let M be a $k \times n$ MDS matrix whose entries are in an alphabet Σ. Then $|\Sigma| \geq \sqrt{n/k}$.*

Proof. Let $P_i = (M_{1,i}, M_{2,i}) \in \Sigma^2$ denote the first two elements in the i-th column of M. If $n > |\Sigma|^2 k$, then there must be k distinct columns $i_1, \ldots, i_k \in [n]$ such that $P_{i_1} = \ldots = P_{i_k}$. But then M cannot be an MDS matrix, as the $k \times k$ minor formed by taking these columns has the first two rows being a scalar multiple of each other, and hence cannot be nonsingular. □

We prove Theorem 1.2 by choosing the matrix M randomly, and showing that with high probability it will be an MDS matrix. This is another aspect in which codes over integers seem to be different from codes over finite fields. Constructing MDS matrices over finite fields seems to require algebraic constructions (such as Reed-Solomon codes), unless the field size is exponential in n; whereas over the integers, random matrices work well even for very small entries.

1.2 Singularity of random matrices

Our main result is a bound on the singularity probability of random $n \times n$ matrices with uniform integer entries in $\{-m, \ldots, m\}$. Note that the probability that such a matrix is singular is at least $(2m + 1)^{-n}$, which is the probability that its first two rows are the same. We show that this bound is tight, up to polynomial factors.

Theorem 1.4 (Singularity of random matrices). *Let $n, m \geq 1$. Let M be an $n \times n$ random matrix with random integer entries chosen uniformly in $\{-m, \ldots, m\}$. Then for some absolute constant $c > 0$,*

$$\Pr[M \text{ is singular}] \leq m^{-cn}.$$

Previous works studied this question in two regimes: fixed m and large n, or fixed n and large m. Ours is the first work that can achieve good dependence on both n and m. Before discussing the connection of our result to previous works, we first show how Theorem 1.2 follows directly from Theorem 1.4.

Proof of Theorem 1.2. Let M be a random $k \times n$ matrix with entries chosen uniformly from $\{-m, \ldots, m\}$. The number of $k \times k$ minors for M is $\binom{n}{k}$, and the probability that each one is singular is at most m^{-ck} by Theorem 1.4. Thus

$$\Pr[M \text{ is not MDS}] \leq \binom{n}{k} m^{-ck} \leq \left(\frac{en}{k}\right)^k m^{-ck} = \left(\frac{en}{km^c}\right)^k.$$

In particular, this probability is at most 2^{-k} (say) whenever $m \geq (2en/k)^{1/c}$. $\qquad\square$

Comparison to previous works in random matrix theory. Most of the previous works in random matrix theory focused on the regime of fixed m and large n. Specifically, on $n \times n$ random matrices whose entries are sampled independently from distributions with bounded tail behaviour. The most studied case is that of random sign matrices, namely $\{-1, 1\}$ entries. Komlós [15] proved that the probability that a random $n \times n$ sign matrix is singular is $o(1)$ as $n \to \infty$, which already is a nontrivial result. It took nearly 30 years until Kahn, Komlós and Szemerédi [12] improved the bound to c^n for some constant $c \in (0, 1)$. A sequence of works [3, 29, 30] improved the value of the constant c, and recently Tikhomirov [31] proved that $c = 1/2 + o(1)$, which is best possible, as the probability that the first two rows of the matrix are equal is 2^{-n}. For more general distributions, Rudelson and Vershynin [22] proved that if the entries of an $n \times n$ matrix are sampled from a sub-Gaussian distribution, then the probability it is singular is at most c^n for some $c \in (0, 1)$. See also [23] for a recent survey on these results.

The other regime, of large m and constant n, was less explored. The only work we are aware of is by Katznelson [14] which gave a bound of the form $c_n m^{-n}$ for some constant c_n depending on n. While having optimal dependence on m for constant n, it has a caveat - it only applies in the regime where m is much larger than n (more precisely, for every fixed n, in the limit of large m).

Random matrices over integers vs over finite fields. Note that if instead we chose M to be a random $n \times n$ matrix over a finite field $\mathbb{F}_q$, then the probability that M is singular would be about $1/q$, independent of how large n is. This is the main point of difference between random matrices over integers and over finite fields - the singularity probability over integers decreases as the matrix becomes larger, whereas over finite fields it stabilizes.

1.3 Discussion

We view Theorem 1.2 as a first step towards the study of codes over integers. There are many intriguing questions that arise in coding theory, once we can show that random integer matrices are MDS with high probability. There are also interesting conjectures on the singularity probability of matrices with entries sampled from general distributions. We discuss both briefly below.

Explicit constructions. A natural question is to give an explicit construction of MDS matrices over integers with small integer values. Concretely, when $k = \alpha n$ for some constant $\alpha \in (0, 1)$, to give an explicit construction of a $k \times n$ MDS matrix with a constant alphabet size (namely, independent of n).

Algorithms. The next natural question, once there are explicit constructions, is to design efficient decoding algorithms for such codes. In particular, it would be intriguing to see if the smaller alphabet size can be utilized to obtain improved runtime (even by logarithmic factors).

General code designs. In this paper, we focus on MDS codes and the alphabet size needed to realize them over integers. Many other code designs have been studied, many of which have the following common form. Let P be a *pattern matrix* whose entries are $\{0, *\}$. A matrix M (over a finite field, or over the integers) of the same dimensions as P, is said to *realize* P if it satisfies the following two conditions:

(i) If $P_{i,j} = 0$ then $M_{i,j} = 0$.

(ii) For any maximal minor in P, if it can be realized by some nonsingular matrix, then the corresponding minor in M is nonsingular.

Questions of this form, for various patterns P, have been studied in coding theory. For example, MDS matrices correspond to patterns P which are all $*$. In some applications, condition (ii) is replaced with the following stronger condition (in which case we say that M *strongly* realizes P):

(ii)' For any (maximal or not) minor in P, if it can be realized by some nonsingular matrix, then the corresponding minor in M is nonsingular.

Some areas where these questions arise are: MDS codes with sparse generating matrices (also known as GM-MDS) [7, 10, 11, 17, 32]; tree codes, used in coding for interactive communication [4, 6, 19, 24, 25]; and maximally recoverable codes, used in coding for distributed storage (there are too many results to list here, we refer to [1] for a recent survey).

Given a pattern P, it is not known when it can be realized (or strongly realized) over small finite fields. Some works show that an exponential field size is needed in some cases [9, 13], whereas other works show that in other cases, a polynomial field size is sufficient, using an algebraic construction [17, 32]. However, a general understanding is currently lacking. In contrast, we speculate that every pattern (except maybe some pathological cases) can be realized over integers with small entries.

To pose a concrete conjecture, let P_n be the $n \times n$ pattern with $*$ on and below the diagonal, and 0 above the diagonal. Such patterns underlie optimal tree codes. For example for $n = 4$:

$$P_4 = \begin{pmatrix} * & 0 & 0 & 0 \\ * & * & 0 & 0 \\ * & * & * & 0 \\ * & * & * & * \end{pmatrix}$$

The best known construction (see [6]) of a matrix M realizing P_n is the binomial coefficients matrix, namely $M_{i,j} = \binom{i}{j}$, whose entries are integers of magnitude about 2^n. We conjecture that this cannot be improved much over finite fields, but can be reduced to $\text{poly}(n)$ over the integers.

Conjecture 1.5. *The following holds for the pattern P_n:*

1. *Any matrix M strongly realizing P_n over a finite field $\mathbb{F}_q$ requires exponential field size, namely $q \geq \exp(\Omega(n))$.*

2. *There exist matrices M strongly realizing P_n over the integers, with the nonzero entries in $\{-m, \ldots, m\}$ for $m = poly(n)$. In fact, random matrices of this form should work with high probability.*

Singularity of matrices over general distributions. As we discussed above, most works on the singularity of random matrices give a bound on the singularity of c^n for some absolute constant $c \in (0, 1)$. Theorem 1.4 shows that if the entries are uniformly sampled from $\{-m, \ldots, m\}$, we can take $c = 1/\text{poly}(m)$. We speculate that this is an instance of a much more general phenomena - the singularity probability is determined by the anti-concentration of the underlying entries distribution. Given a distribution $\mathcal{D}$ over $\mathbb{R}$, define its max-probability as $\|\mathcal{D}\|_\infty = \max_x \mathcal{D}(x)$. For example, if $\mathcal{D}$ is the uniform distribution over $\{-m, \ldots, m\}$, then $\|\mathcal{D}\|_\infty = 1/(2m+1)$.

Conjecture 1.6. *Let $\mathcal{D}$ be a distribution over $\mathbb{R}$ and set $p = \|D\|_\infty$. Let M be a random $n \times n$ matrix with independent entries from $\mathcal{D}$. Then for some absolute constant $c > 0$,*

$$\Pr[M \text{ is singular}] \leq p^{cn}.$$

One can even speculate a more general conjecture, where each entry comes from a different underlying distribution, as long as they all have bounded max-probability.

Paper Outline. We prove Theorem 1.4 in the remainder of the paper. We start with a high-level overview of our framework in Section 2. We compute some preliminary estimates in Section 3, define and study incompressible vectors in Section 4, define the LCD condition in Section 5, where we also prove some properties of it, and bound the LCD of random vectors in Section 6. We put all the ingredients together and complete the proof in Section 7.

2 General approach

We will follow the general approach of Rudelson [21] with several modifications needed to obtain effective bounds for large m.

Notation. It will be convenient to scale the entries to be in $[-1, 1]$; we denote by $\mathcal{D}$ the uniform distribution over $\{a/m : a \in \{-m, \ldots, m\}\}$. We denote by $\mathcal{D}^n$ the distribution over n-dimensional vectors with independent entries from $\mathcal{D}$, and by $\mathcal{D}^{n \times n}$ the distribution over $n \times n$ matrices with independent entries from $\mathcal{D}$. We denote by S^{n-1} the unit sphere in $\mathbb{R}^n$, namely $S^{n-1} = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$. We will use the c, c', c_0, etc, to denote unspecified positive constants. Note that the same letter (e.g. c) can mean different unspecified constants in different lemmas.

We may assume that n, m are large enough. We will assume throughout the proof that n, m are large enough; concretely, for any absolute constants n_0, m_0, we may assume that $n \geq n_0, m \geq m_0$, and this would only effect the value of the constant c in Theorem 1.4.

To see why, consider first the regime of constant m and large n. The distribution $\mathcal{D}$ is symmetric and bounded in $[-1, 1]$. The results of [22] show that in such a case,

$$\Pr[M \text{ is singular}] \leq c^n$$

for some absolute constant $c \in (0, 1)$. This proves Theorem 1.4 for any constant m.

The other regime is that of constant n and large m. While we may appeal to the result of Katznelson [14] in this regime, which gives a sharp bound of $c_n m^{-n}$, there is a much simpler argument that gives a bound of the order of $1/m$ which is good enough to establish Theorem 1.4 in this regime. As the determinant of an $n \times n$ matrix is a polynomial of degree n, the Schwartz-Zippel lemma [26, 33] gives

$$\Pr[M \text{ is singular}] = \Pr[\det(M) = 0] \leq \frac{n}{m}.$$

In particular, for constant n and large m, this probability is of the order of $1/m$, which is consistent with the claimed bound of Theorem 1.4 (taking $c < 1/n$).

General approach. Let $M \sim \mathcal{D}^{n \times n}$, and let $X_1, \ldots, X_n$ denote its rows. If M is singular, then one of the rows belongs to the span of the other rows. By symmetry we have

$$\Pr[M \text{ is singular}] \leq n \cdot \Pr[X_n \in \mathrm{Span}(X_1, \ldots, X_{n-1})].$$

Let X^* be any unit vector orthogonal to $X_1, \ldots, X_{n-1}$ (if there are multiple ones, choose one in some deterministic way). We call it a *random normal vector*. We will shorthand $X = X_n$. Observe that X, X^* are independent. Thus we can bound

$$\Pr[M \text{ is singular}] \leq n \cdot \Pr[\langle X^*, X \rangle = 0].$$

To do so, we will show that unless X^* belongs to a set of "bad" vectors, then the above probability is at most m^{-cn}, and that the probability that X^* is bad is also at most m^{-cn}.

3 Preliminary estimates

We establish some preliminary estimates in this section, which will be needed later in the proof.

Maximal eigenvalues of random matrices. The first ingredient is bounding the spectral norm of M. In fact, we would need this bound for rectangular matrices. Given an $n \times k$ matrix R we denote its spectral norm as $\|R\| = \max\{\|Rx\|_2 : x \in S^{k-1}\}$. Note that $\|R\| = \|R^T\|$ since $\|R\| = \max_{x \in S^{k-1}, y \in S^{n-1}} y^T R x$.

The following claim is a special case of [21, proposition 4.4], who showed that it holds for any symmetric distribution $\mathcal{D}$ supported in $[-1, 1]$.

Claim 3.1. *Let $R \sim \mathcal{D}^{n \times k}$ for $n \geq k$. Then for any $\lambda > 0$,*

$$\Pr[\|R\| \geq \lambda \sqrt{n}] \leq 2^{-c\lambda^2 k}.$$

Anti-concentration of projections. Next, we need anti-concentration results for projections of $\mathcal{D}^n$. To begin with we consider projections of the uniform distribution over the solid cube $[-1,1]^n$.

Claim 3.2. *Let $U \sim [-1,1]^n$ be uniformly distributed. Then for every $x \in S^{n-1}$ and $\varepsilon > 0$,*

$$\Pr_u [|\langle U, x \rangle| \leq \varepsilon] \leq c\varepsilon.$$

Proof. The uniform distribution $U \sim [-1,1]^n$ is a log-concave distribution. Let $S = \langle U, x \rangle$ and note that S is a projection of U along the direction x. The Prékopa–Leindler inequality [16, 20] states that projections of log-concave distributions are log-concave, and so S is a log-concave distribution. Carbery and Wright [5, Theorem 8] show that the required anti-concentration bound holds for any log-concave distribution. □

We extend this anti-concentration to the discrete case using a coupling argument. Here and throughout, we denote by $\log(\cdot)$ logarithm in base 2.

Claim 3.3. *Let $X \sim \mathcal{D}^n$ and set $\varepsilon_0 = \frac{\sqrt{\log m}}{m}$. Then for every $x \in S^{n-1}$ and $\varepsilon \geq \varepsilon_0$,*

$$\Pr [|\langle X, x \rangle| \leq \varepsilon] \leq c\varepsilon.$$

Proof. We apply a coupling argument between the uniform distribution in $[-1,1]^n$ and $\mathcal{D}^n$. Sample $X \sim \mathcal{D}^n$, $Y \sim [-1,1]^n$ and set $Z = X + Y/2m$. Observe that Z is uniform in the solid cube $[-1 - 1/2m, 1 + 1/2m]^n$. Next, fix $\varepsilon > 0$ and observe that $\langle X, x \rangle = \langle Z, x \rangle - \langle Y, x \rangle/2m$. Thus we can bound

$$\Pr[|\langle X, x \rangle| \leq \varepsilon] \leq \Pr[|\langle Z, x \rangle| \leq 2\varepsilon] + \Pr[|\langle Y, x \rangle| \geq 2\varepsilon m].$$

For the first term, Claim 3.2 bounds its probability by $c_1 \varepsilon$. For the second term, the Chernoff bound bounds its probability for $\varepsilon \geq \varepsilon_0$ by $1/m$. As we have $1/m \leq \varepsilon$, the claim follows. □

Tensorization lemma. We also need the following "tensorization lemma" [21, Lemma 6.5].

Claim 3.4. *Let $Y_1, \ldots, Y_n$ be independent real-valued random variables. Assume for some $K, \varepsilon_0 > 0$ that*

$$\Pr[|Y_i| \leq \varepsilon] \leq K\varepsilon \qquad \text{for all } \varepsilon \geq \varepsilon_0.$$

Then

$$\Pr \left[\sum_{i=1}^n Y_i^2 \leq \varepsilon^2 n \right] \leq (cK\varepsilon)^n \qquad \text{for all } \varepsilon \geq \varepsilon_0.$$

Nets. A set of unit vectors $\mathcal{N} \subset S^{n-1}$ is called an ε-net, for $\varepsilon > 0$, if it satisfies:

$$\forall x \in S^{n-1} \, \exists y \in \mathcal{N} \, \|x - y\|_2 \leq \varepsilon.$$

The following claim bounds the size of such a net. For a proof see for example [18, Lemma 2.6].

Claim 3.5. *For any $\varepsilon > 0$, there exists a ε-net $\mathcal{N} \subset S^{n-1}$ of size $|\mathcal{N}| \leq (3/\varepsilon)^n$.*

Integer points in ball. We need a bound on the number of integer vectors in a ball of a given radius. Let $B_n(r) = \{x \in \mathbb{R}^n : \|x\|_2 \leq r\}$ denote the ball of radius r in $\mathbb{R}^n$. The following bound is well known.

Claim 3.6. *The number of integer vectors in $B_n(r)$ is at most $\left(1 + \frac{cr}{\sqrt{n}}\right)^n$.*

4 Compressible vectors

The first set of "bad" vectors that we want to rule out are vectors which are close to sparse. A vector $u \in \mathbb{R}^n$ is k-sparse if it has at most k nonzero coordinates.

Definition 4.1 (Compressible vectors). *Let $\alpha, \beta \in (0, 1)$. A unit vector $x \in S^{n-1}$ is called (α, β)-compressible if it can be expressed as $x = u + v$, where u is (αn)-sparse and $\|v\|_2 \leq \beta$. Otherwise, we say that x is (α, β)-incompressible.*

We will later choose α, β, but we note here that α will be a small enough absolute constant and β a small polynomial in $1/m$. Concrete values that work are $\alpha = 1/50, \beta = 1/\sqrt{m}$. We will implicitly assume that both n, m are large enough; concretely, at various places we assume that $\alpha n \geq 2$.

The main lemma we prove in this section is the following.

Lemma 4.2. *Let $\alpha \in (0, 1/8), \beta \in (\varepsilon_0, 1/2)$ where $\varepsilon_0 = \frac{\sqrt{\log m}}{m}$. Then*

$$\Pr\left[X^* \text{ is } (\alpha, \beta)\text{-compressible}\right] \leq (c\beta)^{n/8}.$$

We need a bound on the smallest singular value of a rectangular matrix.

Claim 4.3. *Let $R \sim \mathcal{D}^{n \times k}$ for $n \geq k$. Then for every $x \in S^{k-1}$ and $\varepsilon \geq \varepsilon_0$,*

$$\Pr\left[\|Rx\|_2 \leq \varepsilon\sqrt{n}\right] \leq (c\varepsilon)^{n/2}.$$

Proof. Assume $\|Rx\|_2 < \varepsilon\sqrt{n}$. This implies that $|(Rx)_i| \leq 2\varepsilon$ for at least $n/2$ coordinates $i \in [n]$. Note that for each fixed i, the value $(Rx)_i$ is distributed as $\langle X, x \rangle$ for some $X \sim \mathcal{D}^k$. Applying Claim 3.3 and the union bound over the choice of the $n/2$ coordinates gives

$$\Pr\left[\|Rx\|_2 \leq \varepsilon\sqrt{n}\right] \leq 2^n (c_1\varepsilon)^{n/2} = (c\varepsilon)^{n/2}.$$

$\square$

Claim 4.4. *Let $R \sim \mathcal{D}^{n \times k}$ for $n \geq 8k$. Then for every $\varepsilon \geq \varepsilon_0$,*

$$\Pr\left[\min_{x \in S^{k-1}} \|Rx\|_2 \leq \varepsilon\sqrt{n}\right] \leq (c\varepsilon)^{n/4}.$$

Proof. We may assume that $\varepsilon \leq 1$ by taking $c \geq 1$. Let $\mathcal{N}$ be an (ε^2)-net in S^{k-1} of size $|\mathcal{N}| \leq (3/\varepsilon^2)^k$, as given by Claim 3.5. Let E_1 denote the event that there exists $y \in \mathcal{N}$ for which $\|Ry\|_2 \leq 2\varepsilon\sqrt{n}$. Applying Claim 4.3 and a union bound gives

$$\Pr\left[E_1\right] \leq (3/\varepsilon^2)^k \cdot (c_1\varepsilon)^{n/2} \leq (c_2\varepsilon)^{n/4},$$

8

where we used the assumption $n \geq 8k$. Let E_2 denote the event that $\|R\| \geq \lambda\sqrt{n}$ for $\lambda = \sqrt{\log(1/\varepsilon)}$. Claim 3.1 shows that $\Pr[E_2] \leq (c_3\varepsilon)^n$. We next show that if E_1, E_2 don't hold then the condition of the claim also doesn't hold, namely that $\|Rx\|_2 > \varepsilon\sqrt{n}$ for all $x \in S^{k-1}$.

Let $x \in S^{k-1}$ be arbitrary and let $y \in \mathcal{N}$ be such that $\|x - y\|_2 \leq \varepsilon^2$. Then

$$\|Rx\|_2 \geq \|Ry\|_2 - \|R\| \cdot \|x - y\|_2 \geq (2\varepsilon - \varepsilon^2\lambda)\sqrt{n}.$$

It can be verified that for $\varepsilon \leq 1$ we have $\varepsilon\lambda \leq 1$, which implies that $\|Rx\|_2 \geq \varepsilon\sqrt{n}$. $\qquad\square$

We will now use these two claims to prove Lemma 4.2.

Proof of Lemma 4.2. Let M' be the $(n-1) \times n$ matrix with rows $X_1, \ldots, X_{n-1}$. Assume that there exists an (α, β)-compressible vector $x \in S^{n-1}$ in the kernel of M'. By definition, $x = u + v$ where u is (αn)-sparse and $\|v\|_2 \leq \beta$. In particular, $M'(u + v) = 0$ and hence $\|M'u\|_2 = \|M'v\|_2$. In addition, $\|u\|_2 \geq \|x\|_2 - \|v\|_2 \geq 1/2$ since x is a unit vector and $\|v\|_2 \leq \beta \leq 1/2$.

Let E denote the event that $\|M'\| \geq \lambda\sqrt{n}$ for $\lambda = c_1\sqrt{\log(1/\beta)}$, where we choose $c_1 \geq 1$ large enough so that by Claim 3.1, $\Pr[E] \leq \beta^n$. Note that as we assume $\beta \leq 1/2$ we have $\lambda \geq c_1 \geq 1$. Assuming that E doesn't hold, we have

$$\|M'u\|_2 = \|M'v\|_2 \leq \|M'\| \cdot \|v\|_2 \leq \lambda\beta\sqrt{n}.$$

In particular, $y = u/\|u\|_2$ is an (αn)-sparse unit vector that satisfies $\|M'y\|_2 \leq 2\lambda\beta\sqrt{n}$. We next bound the probability that such a vector exists.

Let $\varepsilon = 2\lambda\beta$, and note that $\varepsilon \geq \varepsilon_0$ since $\beta \geq \varepsilon_0$ and $\lambda \geq 1$. There are $\binom{n}{\alpha n}$ options for the support of y. Let $I = \{i : y_i \neq 0\}$ denote a possible support, set $k = |I|$ and let R be an $(n-1) \times k$ matrix with columns $(Y_i : i \in I)$. As $\alpha < 1/8$ we have $n - 1 \geq 8k$. Thus we can apply Claim 4.4 and obtain that

$$\Pr\left[\neg E \quad \wedge \quad \exists y \in S^{k-1}, \; \|Ry\|_2 \leq \varepsilon\sqrt{n}\right] \leq (c_2\varepsilon)^{n/4} = \left(c_3\beta\sqrt{\log 1/\beta}\right)^{n/4}.$$

Note that for $\beta \leq 1$ we have $\beta\log(1/\beta) \leq 1$ and hence the above bound is at most $(c_4\beta)^{n/8}$.

To conclude, we union bound over the choices for I, the number of which is $\binom{n}{\alpha n} \leq 2^n$. Thus we can bound the total probability by $2^n(c_4\beta)^{n/8} = (c_5\beta)^{n/8}$. $\qquad\square$

5   The LCD condition

In this section we will introduce the notion of the Lowest Common Denominator (LCD) of a vector, which is a variant of the LCD definition in [21]. Informally, the LCD of a vector is a robust notion of "almost periodicity"; concretely, it is the least multiple where most of its entries are close to integers.

Given $x \in \mathbb{R}^n$ let $x = [x] + \{x\}$ be its decomposition into integer and fractional parts, where $[x] \in \mathbb{Z}^n$ and $\{x\} \in [-1/2, 1/2]^n$.

Definition 5.1 (Least common denominator (LCD)). *Let $\alpha, \beta \in (0,1)$. Given a unit vector $x \in S^{n-1}$, its least common denominator (LCD), denoted $LCD_{\alpha,\beta}(x)$, is the infimum of $D > 0$ such that we can decompose $\{Dx\} = u + v$, where u is (αn)-sparse and $\|v\|_2 \leq \beta\min(D, \sqrt{n})$.*

Claim 5.2. *Assume $x \in S^{n-1}$ is $(5\alpha, \beta)$-incompressible. Then $LCD_{\alpha,\beta}(x) > \sqrt{\alpha n}$.*

Proof. Let $D = \text{LCD}_{\alpha,\beta}(x)$ and assume towards a contradiction that $D \leq \sqrt{\alpha n}$. Let $y = Dx$. As $\|y\|_2^2 \leq \alpha n$ there are at most $4\alpha n$ coordinates $i \in [n]$ where $|y_i| \geq 1/2$. In all other coordinates $\{y_i\} = y_i$, and hence $y - \{y\}$ is $(4\alpha n)$-sparse. By assumption we can decompose $\{y\} = u + v$ where u is (αn)-sparse and $\|v\|_2 \leq \beta D$. This implies that we can decompose $y = u' + v$ where u' is $(5\alpha n)$-sparse. Thus, we can decompose $x = y/D$ as $x = u'' + v''$, where $u'' = u/D$ is $(5\alpha n)$-sparse and $v'' = v/D$ satisfies $\|v''\|_2 \leq \beta$. This violates the assumption that x is $(5\alpha, \beta)$-incompressible. $\qquad\square$

Our main goal in this section is to prove the following lemma, which extends Claim 3.3 assuming x has large LCD. To get intuition, we note that the lemma below is useful as long as $\beta \ll \gamma \ll 1$. We will later set $\gamma = \sqrt{\beta}$ to be such a choice. In particular, if we set $\beta = m^{-1/2}$ then we have $\gamma = m^{-1/4}$.

Lemma 5.3. *Let $X \sim \mathcal{D}^n$. Let $\alpha, \beta, \gamma \in (0, 1/2)$, $x \in S^{n-1}$ be (α, γ)-incompressible and set $D = \text{LCD}_{\alpha,\beta}(x)$. Then for every $\varepsilon \geq 1/2\pi mD$, it holds that*

$$\Pr\left[|\langle X, x \rangle| \leq \varepsilon\right] \leq c \left(\frac{\varepsilon}{\gamma} + \frac{1}{(\alpha\beta m)^{\alpha n}}\right).$$

The proof of Lemma 5.3 relies on Esseen's lemma [8].

Lemma 5.4 (Esseen's Lemma). *Let Y be a real-valued random variable. Let $\phi_Y(t) = \mathbb{E}[e^{itY}]$ denote the characteristic function of Y. Then for any $\varepsilon > 0$, it holds that*

$$\Pr[|Y| \leq \varepsilon] \leq c\varepsilon \int_{-1/\varepsilon}^{1/\varepsilon} |\phi_Y(t)|dt.$$

Before proving Lemma 5.3, we need some auxiliary claims. Fix some $x \in S^{n-1}$, let $X \sim \mathcal{D}^n$ and let $Y = \langle X, x \rangle$. In order to apply Lemma 5.4, we need to compute the characteristic function of Y.

Claim 5.5. *Let $X \sim \mathcal{D}^n$, $x \in S^{n-1}$ and set $Y = \langle X, x \rangle$. For $t \in \mathbb{R}$ it holds that*

$$|\phi_Y(t)| = \prod_{k=1}^{n} F\left(\frac{x_k t}{2\pi m}\right)$$

where $F : \mathbb{R} \to \mathbb{R}$ is defined as follows:

$$F(y) = \left|\frac{\sin\left((2m+1)\pi y\right)}{(2m+1)\sin(\pi y)}\right|.$$

Proof. We have $Y = \sum x_i \xi_i$ where $\xi_1, \ldots, \xi_n \sim \mathcal{D}$ are independent. Hence

$$\phi_Y(t) = \prod_{k=1}^{n} \mathbb{E}[e^{ix_k \xi_k t}].$$

Next we compute

$$\mathbb{E}[e^{ix_k \xi_k t}] = \frac{1}{2m+1} \sum_{\ell=-m}^{m} e^{ix_k(\ell/m)t} = \frac{1}{2m+1} \cdot \frac{\sin(\frac{2m+1}{2m} x_k t)}{\sin(\frac{1}{2m} x_k t)}.$$

10

Hence

$$\left|\mathbb{E}[e^{itx_k\xi_k}]\right| = F\left(\frac{x_k t}{2\pi m}\right).$$

$\square$

The next claim proves some basic properties of the function F.

Claim 5.6. *The function F satisfies the following properties:*

1. *F is symmetric: $F(y) = F(-y)$ for all $y \in \mathbb{R}$.*

2. *F is invariant to shifts by integers: $F(y) = F(\{y\})$ for $y \in \mathbb{R}$.*

3. *F is bounded: $F(y) \in [0,1]$ for all $y \in \mathbb{R}$.*

4. *$F(y) \le G(my)$ for $y \in [0,1/2]$, where $G : \mathbb{R}_+ \to [0,1]$ is defined as follows:*

$$G(y) = \begin{cases} e^{-\eta y^2} & \text{if } y \in [0,1] \\ \frac{e^{-\eta}}{y} & \text{if } y \ge 1 \end{cases}$$

Here, $\eta > 0$ is an absolute constant. Note that G is decreasing.

Proof. The first three claims follow immediately from the definition of F in Claim 5.5. In order to prove the last claim, we will prove that $F(y) \le \frac{c_1}{my}$ for $y \in [1/m, 1/2]$ for some $c_1 \in (0,1)$; and that $F(y) \le \exp(-c_2(my)^2)$ for $y \in [0, 1/m]$ for some $c_2 > 0$. The claim then follows by taking $\eta = \min(\ln(1/c_1), c_2)$.

First, note that $F(y) \le \frac{1}{(2m+1)|\sin(\pi y)|}$. Using Taylor expansion at 0, we get for $y \in [0, 1/2]$ that

$$\sin(\pi y) \ge \pi y - \frac{\pi^3 y^3}{6} \ge \frac{\pi y}{2}.$$

In particular, $F(y) \le \frac{1}{\pi m y}$, which gives the desired bound for $c_1 = 1/\pi$.

Next, note that $F(y) = \frac{1}{2m+1}|\sin((2m+1)\pi y) \cdot \csc(\pi y)|$. The Laurent series of $\csc(x)$ at $x \ne 0$ is $\csc(x) = \frac{1}{x} + \frac{x}{6} + \frac{7x^3}{360} + \frac{31x^5}{15120} + \Theta(x^7)$ and the Taylor series for $\sin(x)$ is $\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} + \Theta(x^7)$. So for $y \in [0, 1/m]$ we have $F(y) \le 1 - c_2(my)^2 \le \exp(-c_2(my)^2)$. $\square$

We also need the following claim, which shows that incompressible vectors retain a large fraction of their norm when restricted to small coordinates. We use the following notation: given $x \in \mathbb{R}^n$ and a set of coordinates $J \subset [n]$, we denote by $x|_J \in \mathbb{R}^J$ the restriction of x to coordinates in J.

Claim 5.7. *Let $x \in S^{n-1}$ be (α, γ)-incompressible. Let $J = \left\{ i : x_i \le \frac{1}{\sqrt{\alpha n - 1}} \right\}$. Then*

$$\|x|_J\|_2^2 \ge \|x|_J\|_\infty^2 + \gamma^2.$$

Proof. Let $J^c = [n] \setminus J$. Since x is a unit vector, we have $|J^c| \le \alpha n - 1$. Let $j \in J$ be such that $|x_j|$ is maximal and take $K = J \setminus \{j\}$. Then $|K^c| \le \alpha n$, and since we assume that x is (α, γ)-incompressible, we have $\|x|_K\|_2 \ge \gamma$. This completes the proof, since

$$\|x|_J\|_2^2 - \|x|_J\|_\infty^2 = \|x|_J\|_2^2 - x_j^2 = \|x|_K\|_2^2 \ge \gamma^2.$$

$\square$

11

We would need the following lemma in the computations later on.

Lemma 5.8. *Let* $\gamma, \delta > 0$. *Let* $x \in \mathbb{R}^n$ *be a vector such that* $\|x\|_\infty \leq \delta$ *and* $\|x\|_2^2 \geq \|x\|_\infty^2 + \gamma^2$. *Let* $T = \pi m / \delta$. *Then*

$$I = \int_0^T \prod_{i=1}^n F\left(\frac{x_i t}{2\pi m}\right) dt \leq \frac{c}{\gamma}.$$

Proof. To simplify the proof, we may assume by Claim 5.6(1) that $x_i \geq 0$ for all i. Reorder the coordinates of x so that $x_1 \geq x_2 \geq \ldots \geq x_n \geq 0$. Observe that for $x_i \in [0, T]$ we have $\frac{x_i t}{2\pi m} \in [0, 1/2]$ and hence we can apply Claim 5.6(4) and bound each term by $F\left(\frac{x_i t}{2\pi m}\right) \leq G\left(\frac{x_i t}{2\pi}\right)$. Thus

$$I \leq \int_0^T \prod_{i=1}^n G\left(\frac{x_i t}{2\pi}\right) dt = 2\pi \int_0^{T/2\pi} \prod_{i=1}^n G(x_i t) dt \leq 2\pi \int_0^\infty \prod_{i=1}^n G(x_i t) dt.$$

We bound this last integral. Let $t_i = 1/x_i$ so that $t_1 \leq t_2 \leq \ldots \leq t_n$. For simplicity of notation set $t_0 = 0, t_{n+1} = \infty$. We break the computation of the integral to intervals $[t_k, t_{k+1})$ for $k = 0, \ldots, n$, and denote by I_k the integral in each interval:

$$I_k = \int_{t_k}^{t_{k+1}} \prod_{i=1}^n G(x_i t) dt = \int_{t_k}^{t_{k+1}} \prod_{i=1}^k \frac{e^{-\eta}}{x_i t} \cdot \prod_{i=k+1}^n e^{-\eta t^2 x_i^2} dt = e^{-\eta k} \int_{t_k}^{t_{k+1}} \frac{e^{-\eta t^2 \sum_{i=k+1}^n x_i^2}}{t^k \prod_{i=1}^k x_i} dt.$$

Fix k and consider first the case that $\sum_{i=k+1}^n x_i^2 \geq \gamma^2/2$. In this case, using the fact that $x_i t \geq 1$ for $i \in [k]$ and $t \in [t_k, t_{k+1}]$, we can bound I_k by

$$I_k \leq e^{-\eta k} \int_{t_k}^{t_{k+1}} e^{-\eta \gamma^2 t^2/2} dt \leq e^{-\eta k} \int_0^\infty e^{-\eta \gamma^2 t^2/2} dt \leq \frac{c_1 e^{-\eta k}}{\gamma}.$$

Next, consider the case that $\sum_{i=k+1}^n x_i^2 < \gamma^2/2$, which means that $\sum_{i=1}^k x_i^2 > \|x\|_2^2 - \gamma^2/2 \geq \|x\|_\infty^2 + \gamma^2/2$. Observe that this is impossible for $k = 0$ or $k = 1$, and hence we may assume $k \geq 2$. In this case we bound

$$I_k \leq e^{-\eta k} \int_{t_k}^{t_{k+1}} \frac{1}{t^k \prod_{i=1}^k x_i} dt \leq e^{-\eta k} \int_{t_k}^\infty \frac{1}{t^k \prod_{i=1}^k x_i} dt = \frac{e^{-\eta k} x_k^{k-1}}{(k-1) \prod_{i=1}^k x_i} \leq \frac{e^{-\eta k}}{(k-1) x_1}.$$

By our assumption, $\sum_{i=1}^k x_i^2 \geq \gamma^2/2$ and hence $x_1^2 \geq \gamma^2/2k$. Thus we can bound

$$I_k \leq \frac{e^{-\eta k}}{(k-1)\gamma/\sqrt{2k}} \leq \frac{c_2 e^{-\eta k}}{\gamma}.$$

Overall, we bounded the integral by

$$I \leq 2\pi \sum_{k=0}^n I_k \leq 2\pi \max(c_1, c_2) \sum_{k=0}^n \frac{e^{-\eta k}}{\gamma} \leq \frac{c}{\gamma},$$

where we used the fact that $c_1, c_2, \eta > 0$ are all absolute constants. $\qquad \square$

Now we have all the ingredients to complete proof of Lemma 5.3.

Proof of Lemma 5.3. Let $Y = \langle X, x \rangle$. Lemma 5.4 and Claim 5.5 give the bound

$$\Pr[|Y| \le \varepsilon] \le c_1 \varepsilon I,$$

where I is the following integral:

$$I = \int_0^{1/\varepsilon} \prod_{i=1}^n F\left(\frac{x_i t}{2\pi m}\right) dt.$$

Let $T = \pi m \sqrt{\alpha n - 1}$. We will bound the integral in the regime $[0, T]$ and $[T, 1/\varepsilon]$, and denote the corresponding integrals by I_1, I_2.

Consider first the integral I_1 in $[0, T]$. Let $\delta = 1/\sqrt{\alpha n - 1}$ and take $J = \{i : x_i \le \delta\}$. Observe that by Claim 5.6(3), we can bound $F\left(\frac{x_i t}{2\pi m}\right) \le 1$ for $i \notin J$. Thus

$$I_1 = \int_0^T \prod_{i=1}^n F\left(\frac{x_i t}{2\pi m}\right) dt \le \int_0^T \prod_{i \in J} F\left(\frac{x_i t}{2\pi m}\right) dt.$$

Next, as we assume that x is (α, γ)-incompressible, Claim 5.7 gives that $\|x|_J\|_2^2 \ge \|x|_J\|_\infty^2 + \gamma^2$. Applying Lemma 5.8 to $x|_J$, we bound the first integral by

$$I_1 \le \frac{c_2}{\gamma}.$$

Next, consider the second integral I_2 in $[T, 1/\varepsilon]$. We will apply the LCD assumption to uniformly bound the integrand in this range. Given $t \in [T, 1/\varepsilon]$, let $y(t) = \left\{\frac{xt}{2\pi m}\right\} \in [-1/2, 1/2]^n$, $\beta(t) = \beta \min(t/\sqrt{n}, 1)$ and $J(t) = \{i \in [n] : |y(t)_i| \ge \beta(t)\}$. As $t \le 1/\varepsilon \le 2\pi m D$, we have that $\frac{t}{2\pi m} \le D = \mathrm{LCD}_{\alpha,\beta}(x)$, and hence $|J(t)| \ge \alpha n$. Applying Claim 5.6, we bound the integrand by

$$\prod_{i=1}^n F\left(\frac{x_i t}{2\pi m}\right) = \prod_{i=1}^n F(y_i) \le \prod_{i \in J(t)} F(y_i) \le \prod_{i \in J(t)} G(m y_i) \le \prod_{i \in J(t)} G(m\beta(t)) \le G(m\beta(t))^{\alpha n}.$$

Following up on this, we have

$$\beta(t) \ge \beta(T) = \beta \frac{\sqrt{\alpha n - 1}}{\sqrt{n}} \ge \beta \sqrt{\alpha/2} \ge \alpha\beta,$$

where we used the assumptions that $\alpha n \ge 2$ and $\alpha \le 1/2$. We may assume that $\alpha\beta m \ge 1$, otherwise the conclusion of the lemma is trivial. In that case we have by Claim 5.6(4) that

$$G(m\beta(t)) \le G(\alpha\beta m) \le \frac{1}{\alpha\beta m}.$$

Thus we can bound the integral I_2 by

$$I_2 = \int_T^{1/\varepsilon} \prod_{i=1}^n F\left(\frac{x_i t}{2\pi m}\right) dt \le \frac{1/\varepsilon}{(\alpha\beta m)^{\alpha n}}.$$

Overall, we get

$$\Pr[|Y| \le \varepsilon] \le c_1 \varepsilon I = c_1 \varepsilon (I_1 + I_2) \le \frac{c_1 c_2 \varepsilon}{\gamma} + \frac{c_1}{(\alpha\beta m)^{\alpha n}}.$$

$\square$

6 Bounding the LCD

Our main goal in this section is to prove that a random normal vector X^* has large LCD with high probability. Let M' denote the $(n-1) \times n$ matrix with rows $X_1, \ldots, X_{n-1}$. Let $D_0 = \sqrt{\alpha n}$ and $D_1 = \beta(\alpha\beta m)^{\alpha n}$ in this section.

Lemma 6.1. *Let $\alpha \in (0, 1/40)$, $\beta \in (0, 1/2)$ and $D \in (1, D_1)$. Then*

$$\Pr[LCD_{\alpha,\beta}(X^*) \leq D] \leq D^2 \left(1/\alpha c\right)^n \beta^{cn}$$

for some absolute constant $c \in (0, 1)$.

We set $\gamma = \sqrt{\beta}$ throughout the section. We first condition on a number of bad events not holding. Define:

$$E_1 = \left[\|M\| \geq \sqrt{n\log(1/\beta)}\right]$$
$$E_2 = [X^* \text{ is } (5\alpha, \beta)\text{-compressible}]$$
$$E_3 = [X^* \text{ is } (\alpha, \gamma)\text{-compressible}]$$

Applying Claim 3.1 for E_1, and Lemma 4.2 for E_2, E_3, we get that

$$\Pr[E_1 \text{ or } E_2 \text{ or } E_3] \leq \beta^{cn}.$$

Thus, we will assume in this section that none of E_1, E_2, E_3 hold. Assuming $\neg E_2$, Claim 5.2 yields that $\mathrm{LCD}_{\alpha,\beta}(X^*) \geq D_0$. For $D \geq D_0$ define

$$\mathcal{S}_D = \left\{x \in S^{n-1} : \mathrm{LCD}_{\alpha,\beta}(x) \in [D, 2D] \text{ and } x \text{ is } (\alpha, \gamma)\text{-incompressible}\right\}.$$

The following is an analog of Lemma 7.2 in [21].

Claim 6.2. *Let $D \geq D_0$ and set $\nu = 6\beta\sqrt{n}/D$. There exists a ν-net $\mathcal{N}_D \subset \mathcal{S}_D$ of size*

$$|\mathcal{N}_D| \leq (D/\beta) \left(\frac{cD}{\sqrt{\alpha n}}\right)^n (1/\beta)^{\alpha n}.$$

Namely, for each $x \in \mathcal{S}_D$ there exists $y \in \mathcal{N}_D$ that satisfies $\|x - y\|_2 \leq \nu$.

Proof. Let $x \in \mathcal{S}_D$ and shorthand $D(x) = \mathrm{LCD}_{\alpha,\beta}(x)$. By definition, we can decompose $\{D(x)x\} = u + v$ where u is (αn)-sparse and $\|v\|_2 \leq \beta \min(D, \sqrt{n}) \leq \beta\sqrt{n}$.

Let W denote the set of (αn)-sparse vectors $w \in [-1/2, 1/2]^n$ such that each w_i is an integer multiple of β. Then $|W| \leq \binom{n}{\alpha n}(1/\beta)^{\alpha n}$, and there exists $w \in W$ such that $\|u - w\|_\infty \leq \beta$, which implies $\|u - w\|_2 \leq \beta\sqrt{n}$. This implies that

$$\|\{D(x)x\} - w\|_2 \leq 2\beta\sqrt{n}.$$

Next, consider $[D(x)x] \in \mathbb{Z}^n$. As $|[a]| \leq 2|a|$ for all $a \in \mathbb{Z}$, we have $\|[D(x)x]\|_2 \leq 2D(x)\|x\|_2 \leq 4D$. Let $Z = \{z \in \mathbb{Z}^n : \|z\|_2 \leq 4D\}$. Then $[D(x)x] \in Z$, and Claim 3.6 bounds $|Z| \leq \left(1 + \frac{c_1 D}{\sqrt{n}}\right)^n$. So there is $z \in Z$ such that

$$\|D(x)x - z - w\|_2 \leq 2\beta\sqrt{n}.$$

Next, let R be set of integer multiples of β in the range $[D, 2D]$, so that $|R| \le D/\beta$ and there exists $r \in R$ with $|D(x) - r| \le \beta$. As $\|x\|_2 = 1$ we have

$$\|rx - z - w\|_2 \le 2\beta\sqrt{n} + \beta \le 3\beta\sqrt{n}.$$

Finally, define the set

$$Y = \{(z + w)/r : z \in Z, w \in W, r \in R\}.$$

Then there exists $y \in Y$ such that

$$\|x - y\|_2 \le 3\beta\sqrt{n}/D = \nu/2.$$

Take a maximal set $\mathcal{N}_D \subset \mathcal{S}_D$ which is ν-separated. That is, for any $x', x'' \in \mathcal{N}_D$ we have $\|x' - x''\|_2 > \nu$. Note that by maximality, $\mathcal{N}_D$ is a ν-net in $\mathcal{S}_D$. Next, note that $|\mathcal{N}_D| \le |Y|$, as any point $x \in \mathcal{N}_D$ must be $(\nu/2)$-close to a distinct point in Y. To conclude, we need to bound $|Y|$. We have

$$|Y| \le |W||Z||R| \le \binom{n}{\alpha n}(1/\beta)^{\alpha n} \cdot \left(1 + \frac{cD}{\sqrt{n}}\right)^n \cdot (D/\beta).$$

As $D \ge D_0 = \sqrt{\alpha n}$ we can simplify $1 + \frac{cD}{\sqrt{n}} \le \frac{(c+1)D}{\sqrt{\alpha n}}$. We can trivially bound $\binom{n}{\alpha n} \le 2^n$. Hence

$$|\mathcal{N}_D| \le |Y| \le (D/\beta)\left(\frac{2(c+1)D}{\sqrt{\alpha n}}\right)^n (1/\beta)^{\alpha n}.$$

$\square$

Claim 6.3. *For any $D \in [D_0, D_1]$ we have*

$$\Pr\left[X^* \in \mathcal{S}_D \text{ and } \neg E_1\right] \le D^2 (c/\alpha)^n \beta^{n/8}.$$

Proof. First, note that we may assume $\beta \le \beta_0$ for any absolute constant $\beta_0 \in (0, 1)$, by choosing the constant $c > 0$ large enough to compensate for that (namely, taking $c \ge 1/\beta_0$). In particular, setting $\beta_0 = 2^{-20}$ works.

If $X^* \in \mathcal{S}_D$ then there exists $y \in \mathcal{N}_D$ such that $\|X^* - y\|_2 \le \nu$ for $\nu = 6\beta\sqrt{n}/D$. By definition of X^* we have $M'X^* = 0$, and as we assume that $\neg E_1$ hold, we have

$$\|M'y\|_2 \le \|M'\|\|X^* - y\|_2 \le \nu\sqrt{n\log(1/\beta)}.$$

Set $\beta_1 = 6\beta\sqrt{\log(1/\beta)}$. The assumption $\beta \le \beta_0$ implies that $\beta_1 \le \beta^{3/4}$. Set $\delta = \beta^{3/4}\sqrt{n}/D$. We will bound the probability that there exists $y \in \mathcal{N}_D$ such that $\|M'y\|_2 \le \delta\sqrt{n}$.

Fix $y \in \mathcal{N}_D$, let $X \sim \mathcal{D}^n$, and define $p(\varepsilon) = \Pr[|\langle X, y \rangle|] \le \varepsilon$. As $y \in \mathcal{N}_D \subset \mathcal{S}_D$ we have that y is (α, γ)-incompressible, and hence we can apply Lemma 5.3, which gives

$$p(\varepsilon) \le c_1 \left(\frac{\varepsilon}{\gamma} + \frac{1}{(\alpha\beta m)^{\alpha n}}\right) \qquad \text{for all } \varepsilon \ge 1/2\pi mD.$$

Next, we restrict attention to only $\varepsilon \ge \delta$, and note that in this regime the first term is dominant (since $D \le D_1$ we have $\delta \ge \beta^{3/4}\sqrt{n}/D_1 \ge 1/(\alpha\beta m)^{\alpha n}$). We can then simplify the bound as

$$p(\varepsilon) \le \frac{c_2\varepsilon}{\gamma} \qquad \text{for all } \varepsilon \ge \delta.$$

15

Applying Claim 3.4, and recalling that we set $\gamma = \sqrt{\beta}$, gives

$$\Pr\left[\|M'y\|_2 \leq \delta\sqrt{n}\right] \leq \left(\frac{c_3\delta}{\gamma}\right)^{n-1} = \left(\frac{c_4\beta^{1/4}\sqrt{n}}{D}\right)^{n-1}.$$

Union bounding over all $y \in \mathcal{N}_D$, using Claim 6.2 to bound its size, gives

$$\Pr[\exists y \in \mathcal{N}_D, \|M'y\|_2 \leq \delta\sqrt{n}] \leq (D/\beta)\left(\frac{cD}{\sqrt{\alpha n}}\right)^n (1/\beta)^{\alpha n} \cdot \left(\frac{c_4\beta^{1/4}\sqrt{n}}{D}\right)^{n-1}$$

$$\leq D^2\left(c_5/\sqrt{\alpha}\right)^n \beta^{n/4-\alpha n-2}.$$

Our assumption $\alpha < 1/40$ and the implicit assumption $\alpha n \geq 2$ imply that $\alpha n + 2 \leq n/8$, which simplifies the above bound to the claimed bound. □

We are now in place to prove Lemma 6.1.

Proof of Lemma 6.1. We may assume that non of E_1, E_2, E_3 hold, as the probability that any of them hold is at most $\beta^{c_1 n}$ for some absolute constant $c_1 \in (0, 1)$. This in particular implies that $\text{LCD}_{\alpha,\beta}(X^*) \geq D_0$. Fix $D \in [D_0, D_1]$. As $D \leq D_1$ we can applying Claim 6.3 to $D_i = 2^i D_0$ as long as $D_i \leq D/2$. Summing the results we get

$$\Pr\left[\text{LCD}_{\alpha,\beta}(X^*) \leq D \text{ and } \neg E_1, \neg E_2, \neg E_3\right] \leq (2D)^2 (c_2/\alpha)^n \beta^{n/8}.$$

Thus overall we have

$$\Pr\left[\text{LCD}_{\alpha,\beta}(X^*) \leq D\right] \leq \beta^{c_1 n} + (2D)^2 (c_2/\alpha)^n \beta^{n/8}.$$

The lemma follows by taking $c \in (0, 1)$ small enough. □

7 Completing the proof

We now prove Theorem 1.4.

Proof of Theorem 1.4. Fix $\alpha = 1/50, \beta = 1/\sqrt{m}$ and assume $m \geq m_0$ for a large enough constant m_0 to be determined soon. Let D to be determined soon. Lemma 6.1 gives

$$\Pr[\text{LCD}_{\alpha,\beta}(X^*) \leq D] \leq D^2(1/\alpha c_1)^n \beta^{c_1 n}.$$

As α is constant, and using the choice $\beta = 1/\sqrt{m}$, we can simplify the bound as follows. For a small enough constant $c \in (0, 1)$, setting $D = m^{cn}$ and $c_2 = 1/\alpha c_1$, we have

$$\Pr[\text{LCD}_{\alpha,\beta}(X^*) \leq m^{cn}] \leq m^{2cn} c_2^n m^{-(c_1/2)n} \leq c_2^n m^{-(c_1/2-2c)n} \leq c_2^n m^{-cn}.$$

Assuming $m \geq m_0$ for a large enough constant m_0, we can simplify this bound further as

$$\Pr[\text{LCD}_{\alpha,\beta}(X^*) \leq m^{cn}] \leq m^{-(c/2)n}.$$

Next, assume $D = \text{LCD}_{\alpha,\beta}(X^*) \geq m^{cn}$. In this case, Lemma 5.3 for $\varepsilon = 1/2\pi m D$ gives that

$$\Pr[\langle X^*, X \rangle = 0] \leq \Pr[|\langle X^*, X \rangle| \leq \varepsilon] \leq c_3 \left(\frac{\varepsilon}{\gamma} + \frac{1}{(\alpha\beta m)^{\alpha n}}\right) \leq m^{-c'n}$$

for some $c' \in (0, 1)$. Overall we obtain the desired bound. □

Acknowledgement

References

[1] S. Balaji, M. N. Krishnan, M. Vajha, V. Ramkumar, B. Sasidharan, and P. V. Kumar. Erasure coding for distributed storage: An overview. *Science China Information Sciences*, 61(10):100301, 2018.

[2] S. Ball. On sets of vectors of a finite vector space in which every subset of basis size is a basis. *Journal of the European Mathematical Society*, 14(3):733–748, 2012.

[3] J. Bourgain, V. H. Vu, and P. M. Wood. On the singularity probability of discrete random matrices. *Journal of Functional Analysis*, 258(2):559–603, 2010.

[4] M. Braverman. Towards deterministic tree code constructions. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 161–167, 2012.

[5] A. Carbery and J. Wright. Distributional and L^q norm inequalities for polynomials over convex bodies in $\mathbb{R}^n$. *Mathematical research letters*, 8(3):233–248, 2001.

[6] G. Cohen, B. Haeupler, and L. J. Schulman. Explicit binary tree codes with polylogarithmic size alphabet. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 535–544, 2018.

[7] S. H. Dau, W. Song, and C. Yuen. On the existence of MDS codes over small fields with constrained generator matrices. In *2014 IEEE International Symposium on Information Theory*, pages 1787–1791. IEEE, 2014.

[8] C. Esseen. On the Kolmogorov-Rogozin inequality for the concentration function. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 5(3):210–216, 1966.

[9] S. Gopi, V. Guruswami, and S. Yekhanin. Maximally recoverable LRCs: A field size lower bound and constructions for few heavy parities. *IEEE Transactions on Information Theory*, 2020.

[10] A. Heidarzadeh and A. Sprintson. An algebraic-combinatorial proof technique for the GM-MDS conjecture. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 11–15. IEEE, 2017.

[11] S. Hoang Dau, W. Song, and C. Yuen. On the existence of MDS codes over small fields with constrained generator matrices. *arXiv*, pages arXiv–1401, 2014.

[12] J. Kahn, J. Komlós, and E. Szemerédi. On the probability that a random ±1-matrix is singular. *Journal of the American Mathematical Society*, 8(1):223–240, 1995.

[13] D. Kane, S. Lovett, and S. Rao. The independence number of the Birkhoff polytope graph, and applications to maximally recoverable codes. *SIAM Journal on Computing*, 48(4):1425–1435, 2019.

[14] Y. Katznelson. Singular matrices and a uniform bound for congruence groups of $SL_n(\mathbb{Z})$. *Duke Mathematical Journal*, 69(1):121–136, 1993.

[15] J. Komlós. On determinant of $(0,1)$ matrices. *Studia Science Mathematics Hungarica*, 2:7–21, 1967.

[16] L. Leindler. On a certain converse of Hölder's inequality. In *Proceedings of the 1971 Oberwolfach Conference, BirkhHuser Verlag. Basel-Stuttgart*, 1972.

[17] S. Lovett. MDS matrices over small fields: A proof of the GM-MDS conjecture. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 194–199. IEEE, 2018.

[18] V. D. Milman and G. Schechtman. *Asymptotic theory of finite dimensional normed spaces: Isoperimetric inequalities in riemannian manifolds*, volume 1200. Springer, 2009.

[19] C. Moore and L. J. Schulman. Tree codes and a conjecture on exponential sums. In *Proceedings of the 5th conference on Innovations in theoretical computer science*, pages 145–154, 2014.

[20] A. Prékopa. On logarithmic concave measures and functions. *Acta Scientiarum Mathematicarum*, 34:335–343, 1973.

[21] M. Rudelson. Lecture notes on non-asymptotic theory of random matrices. 2013.

[22] M. Rudelson and R. Vershynin. The Littlewood–Offord problem and invertibility of random matrices. *Advances in Mathematics*, 218(2):600–633, 2008.

[23] M. Rudelson and R. Vershynin. Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010) (In 4 Volumes) Vol. I: Plenary Lectures and Ceremonies Vols. II–IV: Invited Lectures*, pages 1576–1602. World Scientific, 2010.

[24] L. J. Schulman. Deterministic coding for interactive communication. In *Proceedings of the twenty-fifth annual ACM symposium on Theory of computing*, pages 747–756, 1993.

[25] L. J. Schulman. Coding for interactive communication. *IEEE transactions on information theory*, 42(6):1745–1756, 1996.

[26] J. T. Schwartz. Fast probabilistic algorithms for verification of polynomial identities. *Journal of the ACM (JACM)*, 27(4):701–717, 1980.

[27] B. Segre. Curve razionali normali ek-archi negli spazi finiti. *Annali di Matematica Pura ed Applicata*, 39(1):357–379, 1955.

[28] R. Singleton. Maximum distance q-nary codes. *IEEE Transactions on Information Theory*, 10(2):116–118, 1964.

[29] T. Tao and V. Vu. On random ± 1 matrices: singularity and determinant. *Random Structures & Algorithms*, 28(1):1–23, 2006.

[30] T. Tao and V. Vu. On the singularity probability of random Bernoulli matrices. *Journal of the American Mathematical Society*, 20(3):603–628, 2007.

[31] K. Tikhomirov. Singularity of random Bernoulli matrices. *Annals of Mathematics*, 191(2):593–634, 2020.

[32] H. Yildiz and B. Hassibi. Further progress on the GM-MDS conjecture for Reed-Solomon codes. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 16–20. IEEE, 2018.

[33] R. Zippel. Probabilistic algorithms for sparse polynomials. In *International symposium on symbolic and algebraic manipulation*, pages 216–226. Springer, 1979.