Singleton-type bounds for list-decoding and list-recovery, and related results

Eitan Goldberg[*], Chong Shangguan[†], and Itzhak Tamo[*]

[*]Department of Electrical Engineering–Systems, Tel-Aviv University, Tel-Aviv 39040, Israel
[†]Research Center for Mathematics and Interdisciplinary Sciences, Shandong University, Qingdao 266237, China
[*†]Emails: eitang1@mail.tau.ac.il, theoreming@163.com, zactamo@gmail.com

Abstract

List-decoding and list-recovery are important generalizations of unique decoding that received considerable attention over the years. However, the optimal trade-off among list-decoding (resp. list-recovery) radius, list size, and the code rate are not fully understood in both problems. This paper takes a step towards this direction when the list size is a given constant and the alphabet size is large (as a function of the code length). We prove a new Singleton-type upper bound for list-decodable codes, which improves upon the previously known bound by roughly a factor of $1/L$, where L is the list size. We also prove a Singleton-type upper bound for list-recoverable codes, which is to the best of our knowledge, the first such bound for list-recovery. We apply these results to obtain new lower bounds that are optimal up to a multiplicative constant on the list size for list-decodable and list-recoverable codes with rates approaching capacity.

Moreover, we show that list-decodable *nonlinear* codes can strictly outperform list-decodable linear codes. More precisely, we show that there is a gap for a wide range of parameters, which grows fast with the alphabet size, between the size of the largest list-decodable nonlinear code and the size of the largest list-decodable linear codes. This is achieved by a novel connection between list-decoding and the notion of sparse hypergraphs in extremal combinatorics. We remark that such a gap is not known to exist in the problem of unique decoding.

Lastly, we show that list-decodability or recoverability of codes implies in some sense good unique decodability.

1 Introduction

As a generalization of unique decoding, the notion of list-decoding was introduced independently by Elias [Eli57] and Wozencraft [Woz58] in the 1950s. In list-decoding, given a corrupted codeword, one can output a list of possible codewords, in contrast to unique decoding, where the

output is at most one codeword. The advantage of list-decoding is that it can handle more adversarial errors than unique decoding. List-recovery is a further generalization of list-decoding and was initially used as an intermediate step in the study of list-decoding and unique decoding (see [GI01, GI02, GI03, GI04] for example). Over the years, list-decoding and list-recovery have found many applications in information theory (see [Ahl73, Bli86, Bli97, Eli06] for example) and theoretical computer science (see [CPS99, GUV09, LP20, Siv99, STV01] for example).

Although extensively studied, many combinatorial properties of list-decoding and list-recovery are still far from being well-understood. In particular, the optimal trade-off between the radius of list-decoding/recovery, the list size, and the code rate are not known in both problems. In this paper, we take a step in this direction and consider this trade-off when the list size is constant and the alphabet size is large (as a function of the code length).

To move forward let us introduce some notations and definitions. For a positive integer q, let $[q] = \{1, \ldots, q\}$. The *Hamming distance* $d(x, y)$ between two vectors $x, y \in [q]^n$ is the number of coordinates where they differ, namely, for $x, y \in [q]^n$, let $d(x, y) = |\{i \in [n] : x_i \neq y_i\}|$. For an integer $1 \leq t \leq n$ and a vector $v \in [q]^n$, let $B_t(v)$ denote the Hamming ball of radius t centered at $v \in [q]^n$, which consists of all vectors in $[q]^n$ with Hamming distance at most t from v. A *code C* of block length n over an alphabet of size q is a subset $C \subseteq [q]^n$, whose vectors are called codewords. The *rate* of C is defined to be $R(C) := \log_q(|C|)/n$, and the *minimum distance $d(C)$* of C is defined to be the minimal Hamming distance between all pairs of distinct codewords in C, that is, $d(C) = \min\{d(x, y) : x, y \in C, \ x \neq y\}$. For simplicity, when C is understood from the context, we will drop the dependencies of C in $R(C)$ and $d(C)$ and just write R and d.

Unique decoding. It is well-known and easy to see that for any code with minimum distance d, a Hamming ball of radius $\lfloor \frac{d-1}{2} \rfloor$ centered at any vector in $[q]^n$ can contain at most one codeword of the code. This implies that given a received codeword with at most $\lfloor \frac{d-1}{2} \rfloor$ corrupted coordinates, it is possible to decode it and output the correct (transmitted) codeword, simply by outputting the unique codeword in the Hamming ball of radius $\lfloor \frac{d-1}{2} \rfloor$ around it. In the other direction, if one is requested to correct any fraction r of corrupted coordinates and output a unique codeword, the code must have a minimum distance of at least $2rn + 1$. The classical Singleton bound provides a bound on the parameters of a code and shows that they must satisfy $d + Rn \leq n + 1$. Then, it follows that for unique decoding, one must have $r \leq (1 - R)/2$ as n tends to infinity. Codes such as Reed-Solomon (RS) codes that attain this bound with equality are called MDS codes.

List-decoding and list-recovery. Next, we will give the formal definitions of list-decoding and list-recovery, which are both a generalization of unique decoding.

Definition 1.1. *A code $C \subseteq [q]^n$ is an (r, L) list-decodable for $r \in (0, 1)$, $L \in \mathbb{N}$ if $|B_{rn}(v) \cap C| \leq L$ for all $v \in [q]^n$, where r and L are called the list-decoding radius and the list size of the code, respectively.*

For positive integers ℓ ad q let $\binom{[q]}{\leq \ell}$ be the family of subsets of $[q]$ of size at most ℓ, i.e., $\binom{[q]}{\leq \ell} = \{S \subseteq [q] : |S| \leq \ell\}$.

Definition 1.2. *A code $C \subseteq [q]^n$ is an (r, ℓ, L) list-recoverable for $r \in (0, 1)$, and $\ell, L \in \mathbb{N}$ if for any sequence of lists $S_1, \ldots, S_n \in \binom{[q]}{\leq \ell}$ the size of the set $\{c \in C : c_i \notin S_i$ for at most rn coordinates$\}$ is at most L. Similarly, r and L are called the list-recovery radius and the list size, respectively.*

Clearly, one can easily see that unique decoding is a special case of list-decoding for $L = 1$, and list-decoding is a special case of list-recovery for $\ell = 1$.

Given parameters q and r, the list-decoding capacity cap_{LD} is the supremum over all rates of $q-$ary (r, L) list-decodable codes with L that is at most polynomial in n. Similarly, given q, r and ℓ, the list-recovery capacity cap_{LR} is the supremum over all rates of $q-$ary (r, ℓ, L) list-recoverable codes with L that is at most polynomial in n. By the list-decoding and list-recovery capacity theorems it is known that $\mathrm{cap}_{LD} = 1 - h_q(r)$ for $r \in [0, 1 - \frac{1}{q}]$ (see, e.g., [GRS19, Theorem 7.4.1]) and $\mathrm{cap}_{LR} = 1 - h_{q/\ell}(r) - \log_q(\ell)$ for $r \in [0, 1 - \frac{\ell}{q}]$ (see, e.g., [RW17]), where $h_q(x)$ is the q-ary entropy function defined in (1). Moreover, with high probability, a random code with rate $\mathrm{cap}_{LR} - \epsilon$ for $\epsilon > 0$ is list-recoverable with list size $L = O(\ell/\epsilon)$, and a similar result holds for list-decoding, simply by setting $\ell = 1$.

Next, we describe our main results and compare them to the previously known results.

1.1   Summary of main results

In the results reported below, we primarily consider the case of constant list size L, independent of n. Moreover, similar to the classical Singleton bound, our bounds (results) behave well, only when q is sufficiently large as a function of n. We begin with the first set of results that improve the Singleton-type bounds for list-decoding and recovery.

Singleton-type upper bounds for list-decoding and list-recovery.   Shangguan and Tamo [ST20a] proved the following generalization of the Singleton bound to list-decoding.

Theorem 1.3 (Theorem 1.2 in [ST20a]). *For integers $q \geq 2$, $L \geq 1$ and $r \in [0, \frac{L}{L+1}]$ with $rn \in \mathbb{N}$, every (r, L) list-decodable code $C \subseteq [q]^n$ has size at most $Lq^{n - \lfloor \frac{L+1}{L} rn \rfloor}$.*

It was also observed that one could prove a tighter upper bound for linear codes, as follows.

Proposition 1.4 ([ST20a]). *For integers $q > L \geq 1$, if q is a prime power and $C \subseteq \mathbb{F}_q^n$ is a linear (r, L) list-decodable code, then $|C| \leq q^{n - \lfloor \frac{L+1}{L} rn \rfloor}$.*

[ST20a] also showed that the bound of Proposition 1.4 is tight by showing that certain RS codes attain it with equality, when $L = 2, 3$, $L \mid rn$, and q is sufficiently large as a function of n. Hence, for this set of parameters the largest size of an (r, L) list-decodable code C (not necessarily linear) satisfies

$$q^{n - \lfloor \frac{L+1}{L} rn \rfloor} \leq |C| \leq Lq^{n - \lfloor \frac{L+1}{L} rn \rfloor},$$

and the construction in [ST20a] is optimal up to a constant factor for $L = 2, 3$. We conjecture that the lower bound also holds for any L, r and n such that $L \mid rn$, and that it can be achieved by RS codes, see [ST20a]. Narrowing and even closing the gap between the conjectured lower bound and the upper bound is an interesting open question.

We improve the upper bounds in both Theorem 1.3 and Proposition 1.4, as detailed below.

- In Section 3.1 we show that the factor L in the upper bound of Theorem 1.3 can be replaced by $(1 + o(1))$, where $o(1)$ tends to zero as n tends to infinity. Thus, the new bound has the form $(1 + o(1))q^{n - \lfloor \frac{L+1}{L} rn \rfloor}$, which implies that that the constructions given by [ST20a] is asymptotically optimal for $L = 2, 3$ and $L \mid rn$, even among nonlinear codes. See Corollary 3.2 for the formal statement.

- As already mentioned, Proposition 1.4 is tight if L divides rn and $L = 2, 3$, and it is believed to be tight for any L as long as L divides rn. In Section 3.2, we improve the upper bound exactly when this does not hold. More precisely, we show that for sufficiently large n compared to r and L, every linear (r, L) list-decodable code has dimension at most $n - \lceil \frac{L+1}{L} rn \rceil$ (see Proposition 3.6 for the formal statement). Hence, if $L \nmid rn$, the bound on the dimension is improved by one, compared to Proposition 1.4. This improvement later enables us to separate linear and nonlinear codes by showing that there are nonlinear (r, L) list-decodable codes whose size exceeds the size of any linear (r, L) list-decodable code.

It is known that (r, L) list-decodable codes with alphabet size q and rates approaching $1 - h_q(r)(1 + \frac{1}{L})$ exist (see Theorem 5.5 in [Gur04] and [Eli91]). This implies that for alphabet of size at least $q \geq 2^{\Omega(1/\epsilon)}$, there exist (r, L) list-decodable codes with rates approaching $1 - \frac{L+1}{L} r$, as $h_q(r)$ approaches r when q is sufficiently large. This in turn implies that the Singleton-type bound for list decoding (Theorem 1.3) and its improvements obtained in this paper provide an asymptotically tight bound on the rate as n tends to infinity. However, for a fixed n, it was unclear whether these Singleton-type bounds are tight. In Section 4 we show that this is indeed true, by showing the existence of *nonlinear* codes with "dimension" $n - \frac{L+1}{L} rn - o(1)$, where the $o(1)$ term tends to zero for a fixed n and q tends to infinity (see Section 4 for more details).

Remark 1.5. *Recently, Roth [Rot21, Theorems 4,5] independently proved the same bound as in Proposition 3.6, but under different assumptions. Furthermore, the bounds in [Rot21] are stated as bounds on the list-decoding radius; however, the bounds are equivalent.*

The last result in this set of results is a new Singleton-type bound for (r, ℓ, L) list-recoverable codes, which reduces to Theorem 1.3 for $\ell = 1$. The reader is referred to Section 5 for details.

Lower bounds on the list size. The following is a typical question in the study of list-decoding and list-recovery [Bli86, Bli05, GN14, GLM$^+$20, GV10] for examples). It is stated for list-recovery, and the corresponding question for list-decoding is obtained by setting $\ell = 1$. .

Question 1.6. *Given q, r, ℓ and $\epsilon > 0$ that measures the gap between the code rate and the list-recovery capacity, what is the growth rate of the list size L?*

Previously, Question 1.6 has been studied by several works (mainly for list-decoding), as discussed below. Blinovsky [Bli86, Bli05] showed that any (r, L) list-decodable code with rate $\mathrm{cap}_{LD} - \epsilon$ must have $L = \Omega(\log(\frac{1}{\epsilon}))$. Guruswami and Narayanan [GN14] studied that problem for average-radius list-decoding, which is a strengthening of list-decoding, and showed that the list size must be $\Omega(\frac{1}{\sqrt{\epsilon}})$. Guruswami and Vadhan [GV10] studied the regime of codes with list decoding radius of $r = (1 - 1/q)(1 - \epsilon)$, approaching the upper limit of $1 - 1/q$, and proved that in this regime the list size must be $L = \Omega(1/\epsilon^2)$. Lower bounds on the list size for list-decoding and list-recovery of *random* codes were also studied in [GN14] and [GLM$^+$20]. Guruswami and Narayanan [GN14] proved that both for random codes and random linear codes of rate $\mathrm{cap}_{LD} - \epsilon$ has list size $L = \Omega(1/\epsilon)$, where the hidden leading constant tends to zero as r tends to $1 - 1/q$. Recently, Guruswami et al. [GLM$^+$20] improved the leading constant for random *linear* codes and showed that $L \geq \lfloor h_q(r)/\epsilon + 0.99 \rfloor$. For the binary case they proved this lower bound is tight up to an additive constant, pinning down the list size for random binary linear codes to a range of three values. Additionally for list recovery [GLM$^+$20] showed that for a random linear $(0, \ell, L)$ list-recoverable code of rate $1 - log_q(\ell) - \epsilon$, it holds that $L = \ell^{\Omega(1/\epsilon)}$.

It is known that if one allows the alphabet size q to grow, then there exist random codes with rate $1 - r - \epsilon$ that are list-decodable (resp. list-recoverable) from radius r and list size $L = O(1/\epsilon)$ (resp. $L = O(\ell/\epsilon)$). In fact, it is sufficient that $q \geq 2^{\Omega(1/\epsilon)}$. Hence, in this case the two capacities coincides, and we have $\mathrm{cap}_{LD} = \mathrm{cap}_{LR} = 1 - r$. In Proposition 5.4 we partially answer Question 1.6 by showing that an (r, ℓ, L) list-recoverable code of length n and rate at least $1 - r - \epsilon$ must satisfy $L \geq \frac{\ell r}{\epsilon} + \ell - 1 + o(1) = \Omega_r(\ell/\epsilon)$, where $o(1)$ tends to zero as n tends to infinity. The special case of list decoding is proved in Proposition 3.4.

Nonlinear codes outperform linear codes in list-decoding. A fundamental problem in combinatorial coding theory is to obtain an optimal trade-off between the rate of a code and its relative distance δ. However, this problem is far from being solved despite decades of research. Moreover, even the more modest problem of understanding the power of nonlinear codes is unknown. More precisely, it is unknown whether nonlinear codes perform better than linear codes under unique decoding or linear codes perform as well as their nonlinear counterpart. Indeed, for binary codes with relative distance $0 < \delta < 1/2$, the currently best known lower and upper bounds are the GV [Gil52, Var57] and MRRW [MRRW77] bounds, respectively. It is well known that linear code can achieve the GV bound, and there is no better upper bound for linear codes that is tighter than the MRRW bound. Although some stronger lower bounds are known for nonlinear codes (see [JV04, VW05]), it is not known whether there is a gap between the size of the largest linear code and nonlinear code for a given distance.

Surprisingly, considering the current state of the art for unique decoding, in Section 4 we show that such a separation between linear and nonlinear codes exists for list-decoding. In particular, we show that nonlinear codes can considerably outperform linear codes for sufficiently large q

as a function of n, L. Roughly speaking, we show that for given r, L with $rn \in \mathbb{N}$ and $L \nmid rn$, there is a constant $\theta \in [\frac{1}{L}, 1)$ so that the size of the largest linear (r, L) list-decodable code is at most a $q^{-\theta}$-fraction of the size of the largest nonlinear (r, L) list-decodable code. In particular, for $L = 2$, one can take $\theta = 1 - \epsilon$ with any $\epsilon > 0$ arbitrarily close to zero, provided that q is sufficiently large as a function of n, ϵ. The precise statement of this result can be found in Proposition 4.6 and Remark 4.7. We provide new constructions of list-decodable codes via a correspondence between codes and multi-partite hypergraphs to derive this result. In particular, the constructions are based on a notion of sparse-hypergraphs from extremal combinatorics. We use several known constructions of such sparse hypergraphs in the literature to construct the codes.

We note that results of similar flavor, i.e., that nonlinear codes perform better than linear codes, are known to exist; however, they are scarce. In particular, for the problem of *erasure list-decoding*, it is known that there exist nonlinear codes whose list size is exponentially smaller than the list size guaranteed for linear code (see [Gur04, Theorem 10.17]). Another example of this phenomenon is the recent result by [GLM+20] that showed that in the problem of zero-error list-recovery, random codes also have significantly smaller list sizes than random linear codes.

Large list-decoding radius implies large minimum Hamming distance. We study the relation between list-decodability and unique decodability of a code. In particular, whether a code with good list-decoding properties necessarily imply unique decoding properties. We divide the analysis into two cases, depending on whether the code is linear or nonlinear (see Section 6).

- For a general list-decodable code, i.e., not necessarily linear, we prove that it must contain a large subcode with a large minimum Hamming distance (see Theorem 6.1). As a corollary of this theorem, we obtain that a list-decodable code with a rate approaching the maximal rate given by Theorem 1.3 (and whose existence is guaranteed by Proposition 4.6) must contain a very large near MDS subcode (see Corollary6.2).

- For a linear code that is list-decodable or recoverable, we show that it must have a large Hamming distance. This result can be viewed as a generalization of the fact that an $(r, 1)$ list-decodable code (uniquely decodable) has Hamming distance of at least $2rn + 1$, to list-decoding and recovery. For details, see Theorem 6.4 and its derivatives Proposition 6.5 and Corollary 6.6.

As a final remark, we note that [Rot21] independently proved a result regarding the unique decodability of list-decodable linear codes (see [Rot21, Theorem 3]), which is equivalent to Corollary 6.6 when L divides rn. Moreover, it can be verified from the proof of [Rot21, Theorem 3] that Roth also proved Proposition 6.5. The proof argument of Roth is very similar to ours, though it is formulated differently.

6

1.2 Organization

The rest of this paper is organized as follows. In Section 2 we introduce some necessary notations and definitions. In Sections 3 and 5 we present the Singleton-type upper bounds for list-decoding and list-recovery, respectively. In Section 4 we introduce the notion of sparse hypergraphs and use them to show that for a wide range of parameters, the largest generic list-decodable codes must have much more codewords than the largest linear list-decodable codes. In Section 6 we show that if a linear code has a very large list-decoding or list-recovery radius, then it must also have a very large minimum Hamming distance.

2 Preliminaries and notations

We will use of the following notations. For positive integers $m \leq n$, we write $[n] = \{1, \ldots, n\}$, $[m, n] = \{m, \ldots, n\}$, $\binom{[n]}{m} = \{A \subseteq [n] : |A| = m\}$, and $\binom{[n]}{\leq m} = \{A \subseteq [n] : |A| \leq m\}$. We number vectors by superscripts, i.e., $x^1, x^2, \ldots$, and use subscripts to refer to their coordinates, e.g., x_i^j is the ith coordinate of x^j. For a subset $I \subseteq [n]$ and a vector x of length n, let x_I be the restriction of x to its coordinates with indices in I. For $x, y \in [q]^n$ let $I(x, y) = \{i : x_i = y_i\}$ be the set of indices for which x and y are equal, then it is clear that $d(x, y) + |I(x, y)| = n$, where $d(x, y)$ is the hamming distance between x and y. We will use $h_q(x)$ to denote the q-ary entropy,

$$h_q(x) := x \log_q(q-1) - x \log_q(x) - (1-x) \log_q(1-x). \tag{1}$$

For n subsets $S_1, \ldots, S_n \subseteq [q]$, let $S_1 \times \cdots \times S_n$ be the set of vectors $v \in [q]^n$ with $v_i \in S_i$ for all $i \in [n]$. For a set of vectors $D \subseteq [q]^n$ and a vector $v \in [q]^n$, let $d(v, D) := \min\{d(v, u) : u \in D\}$. Using the above notation, it is not hard to check by definition that a code C is (r, ℓ, L) list-recovery if and only if for every $D \in \binom{[q]}{\leq \ell}^n$,

$$|\{c \in C : d(c, D) \leq rn\}| \leq L$$

where we define $\binom{[q]}{\leq \ell}^n = \{S_1 \times \cdots \times S_n : S_i \in \binom{[q]}{\leq \ell} \text{ for all } i \in [n]\}$.

For a prime power q, let $\mathbb{F}_q$ be the finite field of q elements. A code $C \subseteq \mathbb{F}_q^n$ is linear if and only if it is a subspace of $\mathbb{F}_q^n$. RS codes [RS60] is an important family of linear codes, which is defined as follows: for integers $k \leq n \leq q$ and a vector $\alpha \in \mathbb{F}_q^n$ with distinct entries, the RS code with evaluation vector α is the k-dimensional subspace

$$\{(f(\alpha_1), \ldots, f(\alpha_n)) : f \in \mathbb{F}_q[x], \ deg(f) \leq k - 1\}.$$

Since for every $r' \leq r$ an (r, L) list-decodable code is also (r', L) list-decodable, by Theorem 1.3 we see that an (r, L) list-decodable code with $r \geq \frac{L}{L+1}$ has size at most L, which is obviously tight. So in order to avoid trivial cases, throughout the paper we assume that $r \in [0, \frac{L}{L+1})$.

3   Singleton-type bounds for list-decoding

3.1   Upper bounds for arbitrary codes

Below we will present several upper bounds that improve upon Theorem 1.3. We begin with the following theorem, which provides an improved upper bound on the cardinality of any list-decodable codes.

Theorem 3.1. *Let $C \subseteq [q]^n$ be an (r, L) list-decodable code with $r \in [0, \frac{L}{L+1})$ and $rn \in \mathbb{N}$, then for large enough n (as a function of r and L) the size of C satisfies*

$$|C| \leq \max\{q(1 + o_{r,L}(1)), L\} \cdot q^{n-(\lfloor \frac{L+1}{L} rn \rfloor + 1)},$$

where $o_{r,L}(1)$ is a function that tends to zero for fixed r, L and $n \to \infty$.

Note that by inspecting the precise function $o_{r,L}(1)$ to be given in the proof of the theorem, one can verify that Theorem 3.1 reduces to the classical Singleton bound for $L = 1$. Moreover, the theorem holds for any $n \geq L^2/r$.

Proof. We will need the following observation. Let $C \subseteq [q]^n$ be an (r, L) list-decodable code, then for $I \subseteq [n]$ and $w \in [q]^{|I|}$ the set of vectors $\{c_{\overline{I}} : c \in C, \ c_I = w\} \subseteq [q]^{|\overline{I}|}$ is $(\frac{rn}{n-|I|}, L)$ list-decodable.

Let $m := \left\lfloor \frac{L+1}{L} rn \right\rfloor + 1$ and note that $m \leq n$, since $r < \frac{L}{L+1}$, and let $b \in [0, L-1]$ such that $b \equiv rn \pmod{L}$. We claim that for any vector $w \in [q]^{n-m}$ there are less than $M := 1 + \max\{q + \lfloor f(n)q \rfloor, L\}$, where $f(n) = \frac{(L-b-1)}{2(\lfloor \frac{L+1}{L} rn \rfloor + 1) - (L-b-1)}$ codewords $c \in C$ with $c_{[m+1,n]} = w$, then the result will follow since

$$|C| = \sum_{w \in [q]^{n-m}} |\{c \in C : c_{[m+1,n]} = w\}| \leq \max\{q + \lfloor f(n)q \rfloor, L\} \cdot q^{n-m}$$

$$= \max\{q(1 + o_{r,L}(1)), L\} \cdot q^{n-(\lfloor \frac{L+1}{L} rn \rfloor + 1)}.$$

Assume towards a contradiction that the claim is false, then C contains M codewords whose last $n-m$ coordinates are all identical. Let $v^1, \ldots, v^M \in [q]^m$ be the restriction of these codewords to their first m coordinates. The contradiction will follow by showing that the set of vectors $\{v^i : i \in [M]\} \subseteq [q]^m$ is *not* $(\frac{rn}{m}, L)$ list-decodable, together with the observation above.

Towards this end, let us construct a multi-graph whose vertices are the vectors $v^i, \ i \in [M]$, and draw an edge between distinct v^i and v^j for every coordinate they agree on. It is not hard to verify that each coordinate $i \in [m]$ contributes at least $M - q \geq 0$ edges to the multi-graph. Therefore, the multi-graph has average degree at least $\frac{2(M-q)m}{M} > L - b - 1$, where the inequality follows since $M > q + f(n)q$. As the degree of a vertex must be an integer, there exists a vertex (vector) $v \in \{v^i : i \in [M]\}$ of degree at least $L - b$. Equivalently,

$$\sum_{u \in U} |I(u, v)| \geq L - b, \tag{2}$$

8

where $U = \{v^i : \ i \in [M]\} \setminus \{v\}$. By the value of M, $|U| \geq L$, then it is possible to remove vertices from U to make it the set of vertices $\{v^i : i \in [L]\}$ (possibly by changing the indices of the vertices), while still maintaining (2). Suppose next, that n is sufficiently large so that $\lfloor \frac{rn}{L} \rfloor + 1 \geq L$, then there exists a subset of coordinates $A \subseteq [m]$ be of size $\lfloor \frac{rn}{L} \rfloor + 1$ with $\sum_{u \in U} |A \cap I(v,u)| \geq L - b$.

Partition the set $[m] \setminus A$ arbitrarily to L sets $P^i, i = 1, \ldots, L$, each of size at least $\lfloor \frac{rn}{L} \rfloor + 1 - |A \cap I(v^i, v)|$. This is possible since

$$\sum_{i=1}^{L} \left\lfloor \frac{rn}{L} \right\rfloor + 1 - |A \cap I(v^i, v)| \leq L\left(\left\lfloor \frac{rn}{L} \right\rfloor + 1 \right) - (L - b) = rn = m - |A|.$$

To complete the proof it suffices to construct a vector $y \in [q]^m$ such that the ball $B_{rn}(y)$ contains the vectors $v, v^1, \ldots, v^L$. Define the vector y by

$$y_A = v_A, \text{ and } y_{P^i} = v^i_{P^i} \text{ for } i \in [L].$$

It is clear that $d(y, v) \leq m - |A| = rn$. Furthermore, since y and v agree on the coordinates in A, then v^i agrees with y on $|I(v, v^i) \cap A|$ coordinates in A (which is possibly zero), and by construction on at least $\lfloor \frac{rn}{L} \rfloor + 1 - |I(v, v^i) \cap A|$ coordinates in $[m] \setminus A$. Therefore

$$d(y, v^i) \leq m - |I(v, v^i) \cap A| - \left(\left\lfloor \frac{rn}{L} \right\rfloor + 1 - |I(v, v^i) \cap A| \right) = rn,$$

and we have arrived at the desired contradiction which completes the proof of the theorem. $\square$

The following corollary, which is an easy consequence of Theorem 3.1, shows that for sufficiently large n (as a function of r, L), the factor L in the upper bound in Theorem 1.3 can be replaced by $1 + o(1)$, where $o(1)$ tends to zero as n tends to infinity.

Corollary 3.2. *For integers $q \geq 2$, $1 \leq L \leq q$ and $r \in [0, \frac{L}{L+1})$ with $rn \in \mathbb{N}$, if n is sufficiently large with respect to r, L, then every (r, L) list-decodable code in $[q]^n$ has size at most*

$$(1 + o(1))q^{n - \lfloor \frac{L+1}{L} rn \rfloor},$$

where $o(1)$ tends to zero as n tends to infinity.

Proof. Apply Theorem 3.1 and note that for $L \leq q$, one has $\max\{q + \lfloor (f(n)q \rfloor, L\} \leq (1 + (f(n))q$ and $f(n) = o(1)$. $\square$

Corollary 3.2, which is an improvement over Theorem 1.3, is of interest since it provides an asymptotically optimal bound on the size of such codes. Indeed, the last two authors showed in [ST20a] that over sufficiently large finite fields, $L = 2, 3$ and $L \mid rn$, there exist RS codes of size $q^{n - \frac{L+1}{L} rn}$.

With an additional condition, one can remove the $o(1)$ term in the statement of Corollary 3.2, and obtain a cleaner bound, as follows.

Corollary 3.3. *For integers $q \geq 2$, $1 \leq L \leq q$ and $r \in [0, \frac{L}{L+1})$ with $rn \in \mathbb{N}$, $rn \equiv L - 1$ (mod L), if n is sufficiently large with respect to r, L, then every (r, L) list-decodable code in $[q]^n$ has size at most $q^{n - \lfloor \frac{L+1}{L} rn \rfloor}$.*

Proof. Apply Theorem 3.1 and note that $f(n) = 0$ if $rn \equiv L - 1$ (mod L). $\square$

The next result, which can be deduced from either Theorem 1.3 or Theorem 3.1, gives a lower bound on the list size of list-decodable codes. Moreover, it partially answers Question 1.6.

Proposition 3.4. *Any q-ary (r, L) list-decodable code of length n and rate at least $1 - r - \epsilon$ satisfies $L \geq \frac{r}{\epsilon} + o(1)$, where $o(1)$ tends to zero as n tends to infinity.*

Proof. Let C be a code that satisfies the assumption of the proposition, then by Theorem 1.3

$$q^{(1-r-\epsilon)n} = |C| \leq Lq^{n - \lfloor \frac{L+1}{L} rn \rfloor}.$$

Equivalently, $(1 - r - \epsilon)n \leq n - \lfloor \frac{L+1}{L} rn \rfloor + \log_q L$ and the result follows by rearranging. $\square$

3.2 Improved upper bounds for linear codes

This section shows that the upper bounds obtained in Section 3.1 can be further improved when restricted to linear codes. We begin with the following lemma.

Lemma 3.5. *For a prime power q, positive integers $n, L \geq 2$, and $r \in [0, \frac{L}{L+1})$ satisfying $rn \in \mathbb{N}$, and $n - \lceil \frac{L+1}{L} rn \rceil + 1 > (L-1)\frac{q}{q-1}$, any $[n, n - \lceil \frac{L+1}{L} rn \rceil + 1]_q$ is not (r, L) list-decodable.*

As $(L-1)\frac{q}{q-1}$ is decreasing with q, the lemma in fact holds for all large enough n satisfying $n + 1 - \lceil \frac{L+1}{L} rn \rceil > 2(L-1)$, where we set $q = 2$. In [Rot21] Roth gives examples of two $[n, n - \lceil \frac{L+1}{L} rn \rceil + 1]_q$ codes that *are* (r, L) list-decodable, seemingly contradicting Lemma 3.5. The codes are the $[n, n-1]_q$ parity code, which is $(1/n, n)$ list-decodable and its dual, the $[n, 1]_q$ repetition code for $n = (L+1)u - 1$ for some $u, L \in \mathbb{N}$, which is $(\frac{Ln-1}{(L+1)n}, L)$ list-decodable. One can easily verify that the parameters of these two codes do not satisfy the assumptions of Lemma 3.5.

Proof. Since $rn \in \mathbb{N}$ we can write $rn = La + b$ for integers a, b with $b \in [0, L-1]$. The following can be easily verified

$$\left\lceil \frac{L+1}{L} rn \right\rceil - 1 = \begin{cases} (L+1)a + b = rn + a & \text{if } L \nmid rn \\ (L+1)(a-1) + L = rn + a - 1 & \text{if } L \mid rn \end{cases}$$

Let $C \subseteq \mathbb{F}_q^n$ be an $[n, k]$-linear code with $k = n - (\lceil \frac{L+1}{L} rn \rceil - 1)$. As before, to prove the lemma it suffices to show that there exist $L + 1$ distinct codewords that are contained in a ball of radius rn.

Towards this end, assume without loss of generality that the first k coordinates of the code form an information set, and consider the $k(q-1)$ codewords of the code whose restriction

10

to this information set is a vector of weight one, i.e., it has only one nonzero coordinate. As $k(q-1) > (L-1)q$, by the pigeonhole principle, among these $k(q-1)$ codewords there are L codewords, say $c^1, ..., c^L \in C$, agree on their $k+1$ coordinate, i.e., $c_{k+1}^i = c_{k+1}^j$ for $i, j \in [L]$.

We will consider two cases, $L \nmid rn$ and $L \mid rn$ and notice that $|[k+2, n]| = n - k - 1 = \lceil \frac{L+1}{L} rn \rceil - 2$. In the first case $n - k - 1 = (L+1)a + b - 1 \geq (L+1)a$, hence there is a partition of $[k+2, n]$ into $L+1$ pairwise disjoint subsets, say $P_1, \ldots, P_{L+1}$, with $|P_j| \geq a$ for each $j \in [L+1]$. Next, let $y \in \mathbb{F}_q^n$ be the vector satisfying

$$y_{[k]} = 0, \ y_{k+1} = c_{k+1}^2, y_{P_j} = c_{P_j}^j \text{ for each } j \in [L], \text{ and } y_{P_{L+1}} = 0.$$

It is routine to check that $d(y, 0), d(y, c^j) \leq rn$ for each $j \in [L+1]$, and we have obtained the desired contradiction.

Similarly, if $L|rn$ then $|[k+2, n]| = \lceil \frac{L+1}{L} rn \rceil - 2 = (L+1)(a-1) + L - 2$, hence there is a partition of $[k+2, n]$ into $L+1$ pairwise disjoint subsets, with $|P_j| \geq a - 1$ for each $j \in [L+1]$. Let $y \in \mathbb{F}_q^n$ be the vector satisfying

$$y_{[k]} = 0, \ y_{k+1} = c_{k+1}^2, y_{P_j} = c_{P_j}^j \text{ for each } j \in [L], \text{ and }.$$

As before $d(y, 0), d(y, c^j) \leq rn$ for each $j \in [L+1]$, which contradicts the assumptions of list-decodability. $\square$

The reformulation of Lemma 3.5 gives the following proposition.

Proposition 3.6. *For a prime power q, an integer $2 \leq L$, and $r \in [0, \frac{L}{L+1})$ with $rn \in \mathbb{N}$, there exists an integer $n(r, L)$ such that for all $n \geq n(r, L)$ any $[n, k]_q$ code that is (r, L) list-decodable satisfies $k \leq n - \lceil \frac{L+1}{L} rn \rceil$.*

During the work on this paper, we became aware of a recent paper by Roth [Rot21] who proved a result similar to Proposition 3.6 in Theorems 4,5 of [Rot21]. These theorems provide the same bound on the dimension of the code as Proposition 3.6 does; however, they assume a bit stronger assumptions. More precisely,

- Theorem 4 assumes that the code is MDS with rate greater than $1 - \frac{2}{L} - \frac{(n-k) \mod (L+1)}{n}$.

- Theorem 5 assumes that the code is MDS, alphabet $q > \binom{n}{k+1}$ and list size $n - k - 1 \leq L < \binom{n}{k}$.

Proposition 3.6 almost subsumes Theorem 4, except for a small number of cases for which Theorem 4 holds and Proposition 3.6 does not hold. However, it does not subsume Theorem 5 since Proposition 3.6 assumes that n is large enough compared to L, so it does not hold for L too large, for example, when $L \approx \binom{n}{k}$.

Proof. This is just the contrapositive of Lemma 3.5, since linear codes have integer dimensions. $\square$

We note that the method of [ST20a] also gives the following result, whose proof is omitted.

Proposition 3.7. *For any sufficiently large q and any real $r \in [0, \frac{2}{3})$ with $2 \nmid rn$ there exist $[n, n - \frac{3rn+1}{2}]$-RS codes that are also $(r, 2)$ list-decodable.*

Proposition 3.6 implies that any linear code that satisfies the parameters of Proposition 3.7 has dimension at most $n - \lceil \frac{3rn}{2} \rceil = n - \frac{3rn+1}{2}$, and hence shows that the construction given by Proposition 3.7 is also optimal among all linear codes (in the corresponding parameter regime).

4 Nonlinear codes outperform linear codes in list-decoding

In this section, we show that there exist *nonlinear* codes whose list-decodability outperform any other linear code, i.e., they strictly outperform their linear counterpart. Our method will exploit some known results in the area of extremal (hyper)-graph theory and equivalence between certain "sparse hypergraphs" and codes with "good" list-decodability properties. We begin first by the equivalence.

4.1 An equivalence between codes and multi-partite hypergraphs

Let us begin with some needed definitions. A *hypergraph H* is an ordered pair $H = (V, E)$, where the *vertex set V* is a finite set and the *edge set E* is a family of distinct subsets of V. A hypergraph is called *n-uniform* if all of its edges are of size n. An n-uniform hypergraph is further called *n-partite* if its vertex set V admits a partition $V = \cup_{i=1}^n V_i$ such that every edge intersects each vertex set V_i in *exactly* one vertex.

We will define a natural bijection between the family of n-uniform n-partite hypergraphs with equal part size q (i.e., $|V_i| = q$ for all $i \in [n]$) and the family of q-ary codes of length n, as follows. For $i \in [n]$, let $V_i = \{(i, a) : a \in [q]\}$ and $V = \cup_{i=1}^n V_i$. For an n-uniform n-partite hypergraph $H = (V, E)$ (with equal part size q) we associate a code of size $|E|$ in the following way: for each edge $e = \{(i, x_i) : i \in [n], x_i \in [q]\} \in E$ define the codeword

$$\psi(e) := (x_1, \ldots, x_n) \in [q]^n,$$

and the code

$$C_H := \{\psi(e) : e \in E\} \subseteq [q]^n.$$

Clearly, the mapping ψ is a bijection, and for a vector $x \in [q]^n$ define $\psi^{-1}(x) = \{(i, x_i) : i \in [n], x_i \in [q]\}$. Then, one can easily verify that the Hamming distance between any two vectors $x, y \in [q]^n$ satisfies

$$|\psi^{-1}(x) \cap \psi^{-1}(y)| + d(x, y) = n. \tag{3}$$

Next, we will introduce the notion of sparse hypergraphs, which will later be used to construct nonlinear list-decodable codes, using the mapping ψ defined above.

For positive integers $v \geq 2$, $e \geq 2$, an n-uniform hypergraph H is called (v,e)-sparse if for any e distinct edges $A_1, \ldots, A_e \in H$, it holds that $|\cup_{i=1}^{e} A_i| > v$. Let $g_n(q,v,e)$ denote the maximum number of edges in any (v,e)-sparse n-uniform n-partite hypergraph with equal part size q. In what follows, we will list some known lower bounds on $g_n(q,v,e)$.

Lemma 4.1 (Section 4, Brown, Erdős, Sós [BES73])**.** *Given positive integers q, n, e, there is a $\mu_1 > 0$ depending only on n, e such that $g_n(q,v,e) \geq \mu_1 q^{\frac{en-v}{e-1}}$.*

Lemma 4.2 (Theorem 3, Shangguan, Tamo [ST20b])**.** *Given positive integers q, n, e with $\gcd(en - v, e-1) = 1$, there is $\mu_2 > 0$ depending only on n, e such that $g_n(q,v,e) \geq \mu_2 q^{\frac{en-v}{e-1}} \cdot \log^{\frac{1}{e-1}} q$.*

Lemma 4.3 (Theorem 1, Alon, Shapira [AS06])**.** *Given integers s, n with $2 \leq s < n$ and $\epsilon > 0$, there is an integer $q(n, \epsilon)$ such that for all $q \geq q(n, \epsilon)$ it holds that $g_n(q, 3n - 2s + 1, 3) \geq q^{s-\epsilon}$.*

We remark that although all of the lower bounds in Lemmas 4.1, 4.2 and 4.3 were initially proved for arbitrary n-uniform hypergraph, which is not necessarily n-partite, one can easily convert them to lower bounds on n-uniform n-partite hypergraphs by the following variant of the Erdős-Kleitman lemma (see Theorem 1, [EK68]).

Lemma 4.4. *Any n-uniform hypergraph H with m edges has an n-uniform n-partite subhypergraph with equal part size and at least $\frac{n!}{n^n} \cdot m$ edges.*

Proof. Let $H = (V, E)$. To prove the lemma, we can assume without loss of generality that $n \mid |V|$, as if $n \nmid |V|$, by adding to H $n - |V| \pmod{n}$ isolated vertices, the following proof still works. Take a uniformly chosen random partition of V, such that every subset in the partition has equal size $q := \frac{|V|}{n}$. Let F be the n-uniform n-partite hypergraph given by such a random partition. It is clear that F has equal part size q, and moreover, every n-subset of V lies in F with equal probability $\frac{q^n}{\binom{nq}{n}} > \frac{n!}{n^n}$. Therefore, by the linearity of expectation, the expected number of edges contained in F is at least $\frac{n!}{n^n} \cdot m$, as needed. $\square$

4.2 Constructions of list-decodable codes via sparse hypergraphs

The following lemma shows that one can construct "good" list-decodable codes from sparse hypergraphs.

Lemma 4.5. *If $H = (V, E)$ is an $(n + (L+1)rn, L+1)$-sparse n-uniform n-partite hypergraph, then the code $C_H = \{\psi(e) : e \in E\} \subseteq [q]^n$ is (r, L) list-decodable.*

Proof. Suppose for the sake of contradiction that C_H is not (r, L) list-decodable. Then there exist $L + 1$ distinct codewords $c^1, \ldots, c^{L+1} \in C_H$ and a vector $y \in [q]^n$ such that $d(c^i, y) \leq rn$ for $1 \leq i \leq L + 1$. It therefore follows that $|\psi^{-1}(c^i) \setminus \psi^{-1}(y)| \leq rn$, and moreover

$$|\bigcup_{i=1}^{L+1} \psi^{-1}(c^i)| \leq |\psi^{-1}(y)| + \sum_{i=1}^{L+1} |\psi^{-1}(c^i) \setminus \psi^{-1}(y)| \leq n + (L+1)rn,$$

which contradicts the assumption that H is $(n + (L+1)rn, L+1)$-sparse. $\square$

Applying Lemma 4.5 in concert with the lower bounds of $g_n(q, v, e)$ listed in Lemmas 4.1, 4.2, 4.3 and 4.5 gives the main result of this section, whose proof is omitted as it follows straightforwardly.

Proposition 4.6.   *1. For integers $q \geq 2, L \geq 1$ and a real $r \in [0, \frac{L}{L+1})$ with $rn \in \mathbb{N}$ there exists an (r, L) list-decodable code $C \subseteq [q]^n$ with $|C| \geq \mu_1 q^{n - \frac{rn(L+1)}{L}}$, where $\mu_1 > 0$ depends only on n, L;*

2. For integers $q \geq 2, L \geq 1$ and a real $r \in [0, \frac{L}{L+1})$ with $rn \in \mathbb{N}$ and $\gcd(L, rn) = 1$, there exists an (r, L) list-decodable code $C \subseteq [q]^n$ with $|C| \geq \mu_2 q^{n - \frac{rn(L+1)}{L}} \cdot \log^{\frac{1}{L}} q$, where $\mu_2 > 0$ depends only on n, L;

3. For positive integer n and a real $r \in [0, \frac{2}{3})$ with $rn \in \mathbb{N}$ odd and $\epsilon > 0$, there is an integer $q(n, \epsilon)$ such that for all $q \geq q(n, \epsilon)$ there exist an $(r, 2)$ list-decodable code $C \subseteq [q]^n$ with $|C| > q^{n - \frac{3rn - 1}{2} - \epsilon}$.

Remark 4.7. *Proposition 4.6 shows the existence of nonlinear codes that are better list-decodable than all other linear codes with the same parameters. Indeed, it shows that for $L \geq 2$, rn not divisible by L and sufficiently large q (as a function of n, L), there are (r, L) list-decodable codes of size $\Omega_{n,L}(q^{n - \frac{rn(L+1)}{L}})$. However, Proposition 3.6 shows that for sufficiently large $n \geq n(r, L)$, every linear (r, L) list-decodable code over $\mathbb{F}_q$ has dimension at most $n - \lceil \frac{L+1}{L} rn \rceil$, equivalently it is of size at most $q^{n - \lceil \frac{L+1}{L} rn \rceil}$. Hence, the upper bound given in Proposition 3.6 does not hold in general.*

We find this phenomenon quite surprising; although other instances of nonlinear codes outperform all other linear codes are known in coding theory, they are few and far between.

5   Singleton-type bound for list-recovery

In this section, we prove the following Singleton-type bound for list-recovery.

Theorem 5.1. *For integers $q \geq 2$, $\ell \geq 1, L \geq \ell$ and $r \in [0, 1 - \frac{\ell}{L+1})$ with $rn \in \mathbb{N}$, every (r, ℓ, L) list-recoverable code $C \subseteq [q]^n$ satisfies*

$$|C| \leq L\Big(\frac{q}{\ell}\Big)^{n - \lfloor \frac{L+1}{L+1-\ell} rn \rfloor}.$$

Note that when $\ell = 1$ Theorem 5.1 recovers Theorem 1.3, hence it is can be seen as a generalization of the Singleton bound. Additionally, this bound recovers the list-recovery capacity bound on the rate $R \leq 1 - log_q(\ell)$, for zero error (i.e., $r = 0$) and constant list size L.

Proof. Let $t := \lfloor \frac{\ell}{L+1-\ell} rn \rfloor$ and $m := rn + t$. The proof of the theorem will follow from the following two claims.

Claim 5.2. *For every $B \in \binom{[q]}{\leq \ell}^{n-m}$ there are at most L codewords $c \in C$ with $c_{[m+1,n]} \in B$.*

14

Proof. Recall that for a positive integer n, $\left(\binom{[q]}{\leq \ell}\right)^n = \{S_1 \times \cdots \times S_n : S_i \in \binom{[q]}{\leq \ell} \text{ for all } i \in [n]\}$, where $\binom{[q]}{\leq \ell}$ is the family of subsets of $[q]$ of size at most ℓ. Suppose towards a contradiction that there exist $B \in \left(\binom{[q]}{\leq \ell}\right)^{n-m}$ and $L+1$ codewords $c^1, \ldots, c^{L+1} \in C$ such that $c^i_{[m+1,n]} \in B$, for all $i \in [L+1]$. To obtain the desired contradiction, it is enough to show that there exist sets $S_1, \ldots, S_m \in \binom{[q]}{\leq \ell}$ such that $d(c^i, S_1 \times \cdots \times S_m \times B) \leq rn$ for all $i \in [L+1]$. Since $m = rn + t$ and $c^i_{[m+1,n]} \in B$, it suffices to construct sets S_i such that $d(c^i_{[m]}, S_1 \times \cdots \times S_m) \leq rn$. In other words, for each i there exists at least t indices $j \in [m]$ such that $c^i_j \in S_j$. The sets S_j be can be easily constructed since the number of constraints they have to satisfy is $(L+1)t$, whereas there are m sets S_i, each of size ℓ, and the choice of the parameters $(L+1)t \leq \ell m$. Table 1 shows an example of a selection of such sets for certain values of the parameters.

S_1	S_2	S_3	S_4
c^1_1	c^1_2	c^1_3	c^2_4
c^2_1	c^2_2	c^3_3	c^3_4
c^3_1	c^4_2	c^4_3	c^4_4

Table 1: An example of the sets S_i for $rn = 1$, $L = 3$, $\ell = 3$, hence $m = 4, t = 3$. The sets are $S_1 = \{c^1_1, c^2_1, c^3_1\}$, $S_2 = \{c^1_2, c^2_2, c^4_2\}$, $S_3 = \{c^1_3, c^3_3, c^4_3\}$ and $S_4 = \{c^2_4, c^3_4, c^4_4\}$.

$\square$

Claim 5.3. *For any $t \leq n - m$, and any $S_1, \ldots, S_{n-m-t} \in \binom{[q]}{\leq \ell}$, there are at most $L(\frac{q}{\ell})^t$ codewords $c \in C$ with $c_i \in S_i$.*

Proof. Let us apply induction on t. For the base case $t = 0$, the statement follows from Claim 5.2. Now in order to prove the claim for $t \leq n - m$, let us assume that we have proved it for $t - 1$. Fix $S_1, \ldots, S_{n-m-t} \in \binom{[q]}{\leq \ell}$. For $j \in [q]$ let a_j be the number of codewords $c \in C$ with $c_i \in S_i$ for all $i \in [n - m - t]$ and $c_{n-m-t+1} = j$. It is easy to see that

$$\{c \in C : c_i \in S_i \text{ for all } i \in [n - m - t]\} = \sum_{j \in [q]} a_j,$$

so to prove the claim it suffices to show $\sum_{j \in [q]} a_j \leq L(\frac{q}{\ell})^t$. By induction hypothesis for $t - 1$, for any $A \in \binom{[q]}{\ell}$, $\sum_{j \in A} a_j \leq L(\frac{q}{\ell})^{t-1}$. Therefore, by averaging over all such sets $A \in \binom{[q]}{\ell}$ it can be easily seen that $\sum_{j \in [q]} a_j \leq L(\frac{q}{\ell})^t$, completing the proof of the claim. $\square$

Returning to the proof of Theorem 5.1, one can see that it follows directly from Claim 5.3 with $t = n - m$. $\square$

The following corollary, which can be deduced easily from Theorem 5.1, gives a lower bound on the list size of list-recoverable codes. It also provides a partial answer to Question 1.6, as explained in the introduction.

Corollary 5.4. *Any q-ary (r, ℓ, L) list-recoverable code of length n and rate at least $1 - r - \epsilon$ satisfies $L \geq \frac{\ell r}{\epsilon} + \ell - 1 + o(1)$, where $o(1)$ tends to zero as n tends to infinity.*

Proof. By Theorem 5.1 the size of any such code C satisfies

$$q^{(1-r-\epsilon)n} \leq |C| \leq L\left(\frac{q}{\ell}\right)^{n - \lfloor \frac{(L+1)rn}{L+1-\ell} \rfloor} \leq Lq^{n - \lfloor \frac{(L+1)rn}{L+1-\ell} \rfloor}.$$

Hence, $\frac{\ell}{L+1-\ell}rn \leq \epsilon n + log_q(L) + 1 + \frac{L+1}{L+1-\ell}$, and as $\left(log_q(L) + 1 + \frac{L+1}{L+1-\ell}\right) \Big/ n$ tends to zero as n tends to infinity, the result follows. $\square$

6 Large list-decoding radius implies a large minimum distance

In this section, we show that good list-decoding property would imply in some sense good unique decoding property, i.e., large minimum distance. We divide our analysis into two cases; first, for general codes (not necessarily linear), we show that such a statement can not hold in general, but it is undoubtedly true for a large subcode of the code. Then, we proceed to consider the case of linear codes, where we are able to prove that even for the general problem of list-recovery, the code itself (and not its subcode) must have a large minimum distance.

General codes: It is clear that for general codes, one can *not* hope that good list-decodability would imply a large minimum distance. Indeed, given an $(r, L - 1)$ list-decodable code, one can "ruin" the minimum distance by adding a new codeword that is of distance one from one of its codewords (and thereby possibly making it a nonlinear code). The new code is (r, L) list-decodable code and thus retains its good list-decoding property; however, it has a poor minimum distance. On the other hand, one needs only to remove the newly added codeword to obtain back the (possibly) large minimum distance of the code. In other words, a small number of codewords needs to be removed to have also a large minimum distance (and, of course, retain the good list-decoding property). The following theorem shows that this is the only case in general, i.e., any large enough code with good list-decoding property must contain a large subcode with a large minimum distance. The formal details follow.

Theorem 6.1. *Let $L \geq 1$ be an integer, $\gamma \in (0, 1)$, $r \in [0, \frac{L}{L+1})$, $n \in \mathbb{N}$ with $rn \in \mathbb{N}$, and $q \geq q(n, r, L, \gamma)$, then every (r, L) list-decodable code $C \subseteq [q]^n$ with $|C| = q^{n - \lfloor \frac{L+1}{L}rn \rfloor - \epsilon}$, where $\epsilon := \epsilon(n) \geq 0$ is an integer valued function such that $(L-1)(\epsilon+1) \leq \lfloor rn/L \rfloor$, contains a subcode of size at least $\gamma|C|$ and minimum distance at least $\lfloor \frac{L+1}{L}rn \rfloor - (L-1)(\epsilon+1) + 1$.*

We note that in the more general case where ϵ is not necessarily an integer, a slightly weaker result holds, where one can show that the subcode distance is weakened to be at least $\lfloor \frac{L+1}{L}rn \rfloor - (L-1)(\lfloor \epsilon \rfloor + 2) + 1$, if $(L-1)(\lfloor \epsilon \rfloor + 2) \leq \lfloor rn/L \rfloor$. We omit the details.

Applying Theorem 6.1 to a sequence of codes with $\epsilon(n) = o(n)$ and a fixed L gives the following corollary.

Corollary 6.2. *Let $L \geq 1$ be an integer, $\gamma \in (0,1)$, $r \in [0, \frac{L}{L+1})$, $n \in \mathbb{N}$ with $rn \in \mathbb{N}$, and $q \geq q(n, r, L, \gamma)$, then every (r, L) list-decodable code $C \subseteq [q]^n$ with $|C| = q^{n-\lfloor \frac{L+1}{L}rn \rfloor - o(n)}$ contains a subcode of size at least $\gamma|C|$ and minimum distance at least $\lfloor \frac{L+1}{L}rn \rfloor - o(n)$.*

Notice that the guaranteed subcode by Corollary 6.2 has near optimal rate-distance tradeoff. Indeed, the rate is at least $1 - \delta + \frac{1}{n} - o(1)$, whereas the rate of an MDS code is $1 - \delta + \frac{1}{n}$. Moreover, Proposition 4.6 shows the existence of such sequence of codes with $\epsilon < 1$ and large enough q.

We proceed with the proof Theorem 6.1.

Proof. Let $C \subseteq [q]^n$ be a code that satisfies the assumptions of the theorem. The required subcode will be constructed by identifying and removing a small subset of codewords from the code that cause its minimum distance to be relatively small. Therefore, the remained codewords, i.e., the subcode, will have a large minimum distance. We proceed with the formal proof.

Let $m := \lfloor \frac{L+1}{L}rn \rfloor + \epsilon + 1$, and say that a vector $w \in [q]^{n-m}$ is bad for the subset $I \subseteq [n], |I| = n - m$ if there exist two codewords $c^1, c^2 \in C$ such that $c_I^1 = c_I^2 = w$, and $c_j^1 = c_j^2$ for at least $L(\epsilon + 1)$ additional coordinates $j \notin I$, and note that by the assumption on ϵ, $L(\epsilon + 1) \leq m$. Next, we get a bound on the number of codewords whose projection on a fixed set I is a bad vector for I.

Claim 6.3. *If $w \in [q]^{n-m}$ is bad for the set I, then there are at most L codewords $c \in C$ with $c_I = w$.*

Now we can proceed to construct the desired subcode C' with large size and distance, as follows. C' is obtained from C by removing from it all codewords c such that there exists a subset $I \subseteq [n]$ of size $n - m$, and a vector $w \in q^{n-m}$ that is bad for I and $c_I = w$. We claim that C' has distance at least $\lfloor \frac{L+1}{L}rn \rfloor - (L-1)(\epsilon + 1) + 1$, and size at least $\gamma|C|$.

Size: As there are $\binom{n}{n-m}$ sets $I \subseteq [n]$ of size $n - m$, and for each such set I, there are at most q^{n-m} bad vectors w for it. By Claim 6.3 there are at most L codewords $c \in C$ such that $c_I = w$. Hence at most

$$\binom{n}{n-m}q^{n-m}L,$$

codewords were removed and the size of C' is a least

$$q^{n-m+1} - L\binom{n}{n-m}q^{n-m} = \left(1 - \frac{L\binom{n}{n-m}}{q}\right)|C| \geq \gamma|C|,$$

for large enough $q \geq q(n, r, L, \gamma)$.

Minimum distance: By construction, any two codewords of C that agreed on at least $n - m + L(\epsilon + 1)$ coordinates were removed, thus the minimum distance is at least

$$m - L(\epsilon + 1) + 1 = \left\lfloor \frac{L+1}{L}rn \right\rfloor - (L-1)(\epsilon + 1) + 1,$$

as needed. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

It remains to prove Claim 6.3

Proof of Claim 6.3. Let w be a bad vector for I, where we assume without loss of generality that $I = [n - m + 1, n]$, and assume towards contradiction that there are $L + 1$ codewords c^j such that $c^j_I = w$. Further, assume that $c^1, c^2 \in C$ $c^1_{[(L\epsilon+1)]} = c^2_{[(L\epsilon+1)]}$.

As before, write $rn = La + b$ with integers a, and $0 \le b < L$. Thus, $m = (L+1)a + b + \epsilon + 1$, and since we assumed that $(L-1)(\epsilon+1) \le \lfloor rn/L \rfloor$, then $L(\epsilon+1) \le a + \epsilon + 1$. Next, partition that set $[2(a + \epsilon + 1) - L(\epsilon + 1) + 1, m]$ into $L - 1$ disjoint sets I_j for $j \in [3, L + 1]$, each of size at least $a + \epsilon + 1$, which is possible since

$$|[2(a + \epsilon + 1) - L(\epsilon + 1) + 1, m]|$$
$$= m - \big(2a - (L - 2)(\epsilon + 1) + 1\big) + 1$$
$$= (L + 1)a + b + \epsilon + 1 - \big(2a - (L - 2)(\epsilon + 1) + 1\big) + 1$$
$$= (L - 1)a + b + (L - 1)(\epsilon + 1) = (L - 1)(a + \epsilon + 1) + b$$

Next, we show that C is not (r, L) list-decodable. To obtain the desired contradiction we consider the following vector $y \in [q]^n$

$$y_i = \begin{cases} c^1_i & i \in [a + \epsilon + 1] \\ c^2_i & i \in [a + \epsilon + 2, 2(a + \epsilon + 1) - L(\epsilon + 1)] \\ c^j_i & i \in I_j \text{ and } j \in [3, L + 1] \\ w_{i-m} & i \in [m + 1, n]. \end{cases}$$

It is fairly straightforward to check that $c^1, \dots, c^{L+1} \in B_{rn}(y)$, and we arrive at the desired contradiction. $\square$

Linear codes: By utilizing the additional structure of linear codes which general codes might not possess, we can prove that *any* linear code that is list-recoverable must have good unique decoding properties, i.e., large minimum distance. Furthermore, for $\ell = 1, L = 1$, this result can also be seen as a generalization of the fact that codes that are unique decodable from relative radius r, have a Hamming distance greater than $2rn$.

Theorem 6.4. *For a prime power q, integers $1 \le \ell \le q$, $\ell \le L < \ell q$ and $r \in [0, 1 - \frac{\ell}{L+1})$ with $rn \in \mathbb{N}$, if $C \subseteq \mathbb{F}_q^n$ is a linear (r, ℓ, L) list-recoverable code of dimension at least 2, then $d(C) > rn + \lfloor \frac{\ell}{L+1-\ell} rn \rfloor$.*

Note that it is easy to prove that an (r, ℓ, L) list-recoverable linear code has minimum distance $d(C) > rn$ if $L < q$. Otherwise, all the q multiples of a minimum weight codeword would violate the list-recoverability of the code. Theorem 6.4 improves on this observation by utilizing the list-recovery property even further.

Proof. Assume towards contradiction that $d(C) \leq rn + \lfloor \frac{\ell}{L+1-\ell} rn \rfloor$, and let $m := rn + \lfloor \frac{\ell}{L+1-\ell} rn \rfloor$. Since C is a linear code, there exists a nonzero codeword $c \in C$ with weight at most m, and assume without loss of generality that $c_{[m+1,n]} = 0$. By the dimension of the code, let $c^1 \dots, c^{\ell-1} \in C$ be ℓ distinct codewords which are scalar multiples of each other, but *not* a scalar multiple of c, and let $c^\ell = 0$. Next, note that any of the $\ell q > L$ distinct codewords of the form $v = \lambda c + c^j$ for $\lambda \in \mathbb{F}, j \in [\ell]$ satisfy $v_i \in S_i := \{c_i^j : j \in [\ell]\}$ for $i \in [m+1, n]$, where clearly $|S_i| \leq \ell$. This contradicts Claim 5.2 with $B = S_{m+1} \times \cdots \times S_n$. $\square$

The following proposition is a similar result to that of Theorem 6.4 but specialized for list-decodable codes and without the constraint on the code's dimension to be at least 2. We omit its proof, as it is very similar to that of Theorem 6.4.

Proposition 6.5. *For a prime power q, an integer $1 \leq L < q$ and $r \in [0, \frac{L}{L+1})$ with $rn \in \mathbb{N}$, if $C \subseteq \mathbb{F}_q^n$ is a linear (r, L) list-decodable code, then $d(C) > rn + \lfloor \frac{rn}{L} \rfloor$.*

We want to emphasize that the lower bound on the minimum distance given by Proposition 6.5 is tight for certain parameters. Indeed, in [ST20a] it was shown that over sufficiently large finite fields, a positive integer rn, $L = 2, 3$ and $L \mid rn$ there exist an $[n, n - \frac{L+1}{L} rn]$-RS codes, which are (r, L) list-decodable, and whose minimum distance is $rn + \frac{rn}{L} + 1$, which attains the lower bound of Proposition 6.5.

Proposition 6.5 was also implicitly proved by Roth recently (See [Rot21, Theorem 3]). Specifically, he shows that an (r, L) list-decodable code of dimension $n - \lfloor \frac{L+1}{L} rn \rfloor$ is an MDS code. This result follows directly from Proposition 6.5, as shown in the next Corollary, which shows that a linear code that attains with equality the bound on the dimension given in Proposition 3.6 has to be a near MDS or an MDS code.

Corollary 6.6. *For a prime power q, an integer $1 \leq L < q$ and $r \in [0, \frac{L}{L+1})$ satisfying $rn \in \mathbb{N}$, any $[n, k]$ linear (r, L) list-decodable code whose dimension attains the bound in Proposition 3.6 with equality, i.e., $k = n - \lceil \frac{L+1}{L} rn \rceil$, has distance at least $n - k$. Furthermore, if $L | rn$ the distance is at least $n - k + 1$, hence it is an MDS code.*

Proof. This is a direct consequence of Proposition 6.5. $\square$

Acknowledgements

References

[Ahl73]   R. Ahlswede. Channel capacities for list codes. *Journal of Applied Probability*, 10(4):824–836, 1973.

[AS06]   N. Alon and A. Shapira. On an extremal hypergraph problem of Brown, Erdős and Sós. *Combinatorica*, 26(6):627–645, 2006.

[BES73]   W. G. Brown, P. Erdős, and V. T. Sós. On the existence of triangulated spheres in 3-graphs, and related problems. *Period. Math. Hungar.*, 3(3-4):221–228, 1973.

[Bli86]   V. M. Blinovskiĭ. Bounds for codes in decoding by a list of finite length. *Problemy Peredachi Informatsii*, 22(1):11–25, 1986.

[Bli97]   V. M. Blinovsky. *Asymptotic Combinatorial Coding Theory*. Kluwer Academic Publishers, Norwell, MA, USA, 1997.

[Bli05]   Vladimir M Blinovsky. Code bounds for multiple packings over a nonbinary finite alphabet. *Problems of Information Transmission*, 41(1):23–32, 2005.

[CPS99]   Jin-Yi Cai, A. Pavan, and D. Sivakumar. On the hardness of permanent. In *STACS 99 (Trier)*, volume 1563 of *Lecture Notes in Comput. Sci.*, pages 90–99. Springer, Berlin, 1999.

[EK68]   P. Erdős and D. J. Kleitman. On coloring graphs to maximize the proportion of multicolored *k*-edges. *J. Combinatorial Theory*, 5:164–169, 1968.

[Eli57]   Peter Elias. List decoding for noisy channels. *Wescon Convention Record, Part 2, Institute of Radio Engineers*, pages 99–104, 1957.

[Eli91]   Peter Elias. Error-correcting codes for list decoding. *IEEE Transactions on Information Theory*, 37(1):5–12, 1991.

[Eli06]   P. Elias. Error-correcting codes for list decoding. *IEEE Trans. Inform. Theory*, 37(1):5–12, September 2006.

[GI01]   Venkatesan Guruswami and Piotr Indyk. Expander-based constructions of efficiently decodable codes. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pages 658–667. IEEE, 2001.

[GI02]   Venkatesan Guruswami and Piotr Indyk. Near-optimal linear-time codes for unique decoding and new list-decodable codes over smaller alphabets. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 812–821, 2002.

[GI03]   Venkatesan Guruswami and Piotr Indyk. Linear time encodable and list decodable codes. In *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pages 126–135, 2003.

[GI04]   Venkatesan Guruswami and Piotr Indyk. Efficiently decodable codes meeting gilbert-varshamov bound for low rates. In *SODA*, volume 4, pages 756–757. Citeseer, 2004.

[Gil52]   Edgar N Gilbert. A comparison of signalling alphabets. *The Bell system technical journal*, 31(3):504–522, 1952.

[GLM+20]   Venkatesan Guruswami, Ray Li, Jonathan Mosheiff, Nicolas Resch, Shashwat Silas, and Mary Wootters. Bounds for list-decoding and list-recovery of random linear codes. *CoRR*, abs/2004.13247, 2020.

[GN14]   Venkatesan Guruswami and Srivatsan Narayanan. Combinatorial limitations of average-radius list-decoding. *IEEE Transactions on Information Theory*, 60(10):5827–5842, 2014.

[GRS19]   Venkatesan   Guruswami,   Atri   Rudra,   and   Madhu   Sudan.   Essential   coding   theory.   *Draft   available   at* *http://cse.buffalo.edu/faculty/atri/courses/coding-theory/book/*, 2019.

[Gur04]   Venkatesan Guruswami. *List decoding of error-correcting codes: winning thesis of the 2002 ACM doctoral dissertation competition*, volume 3282. Springer Science & Business Media, 2004.

[GUV09]   Venkatesan Guruswami, Christopher Umans, and Salil Vadhan. Unbalanced expanders and randomness extractors from parvaresh–vardy codes. *Journal of the ACM (JACM)*, 56(4):1–34, 2009.

[GV10]   Venkatesan Guruswami and Salil Vadhan. A lower bound on list size for list decoding. *IEEE Transactions on Information Theory*, 56(11):5681–5688, 2010.

[JV04]   Tao Jiang and Alexander Vardy. Asymptotic improvement of the gilbert-varshamov bound on the size of binary codes. *IEEE Trans. Inform. Theory*, 50(8):1655–1664, 2004.

[LP20]   Ben Lund and Aditya Potukuchi. On the list recoverability of randomly punctured codes. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2020)*, volume 176, pages 30:1–30:11, 2020.

[MRRW77]   Robert McEliece, Eugene Rodemich, Howard Rumsey, and Lloyd Welch. New upper bounds on the rate of a code via the delsarte-macwilliams inequalities. *IEEE Transactions on Information Theory*, 23(2):157–166, 1977.

[Rot21]   Ron M. Roth. Higher-order mds codes, 2021.

[RS60]   Irving S. Reed and Gustave Solomon. Polynomial codes over certain finite fields. *Journal of the Society for Industrial and Applied Mathematics*, 8(2):300–304, 1960.

[RW17]   Atri Rudra and Mary Wootters. Average-radius list-recovery of random linear codes: it really ties the room together. *CoRR*, abs/1704.02420, 2017.

[Siv99]   D. Sivakumar. On membership comparable sets. *Journal of Computer and System Sciences*, 59(2):270–280, 1999.

[ST20a]   Chong Shangguan and Itzhak Tamo. Combinatorial list-decoding of Reed-Solomon codes beyond the Johnson radius. In *Proceedings of the 52nd Annual ACM Symposium on Theory of Computing*, STOC 2020, pages 538–551, 2020.

[ST20b]   Chong Shangguan and Itzhak Tamo. Sparse hypergraphs with applications to coding theory. *SIAM J. Discrete Math.*, 34(3):1493–1504, 2020.

[STV01]   Madhu Sudan, Luca Trevisan, and Salil Vadhan. Pseudorandom generators without the xor lemma. *Journal of Computer and System Sciences*, 62(2):236–266, 2001.

[Var57]   Rom Rubenovich Varshamov. Estimate of the number of signals in error correcting codes. *Docklady Akad. Nauk, SSSR*, 117:739–741, 1957.

[VW05]   Van Vu and Lei Wu. Improving the gilbert-varshamov bound for q-ary codes. *IEEE Trans. Inform. Theory*, 51(9):3200–3208, 2005.

[Woz58]   John M. Wozencraft. List decoding. *Quarterly Progress Report, Research Laboratory of Electronics, MIT*, 48:90–95, 1958.