Electronic Colloquium on Computational Complexity, Report No. 126 (2023)

AG codes have no list-decoding friends:
Approaching the generalized Singleton bound requires exponential alphabets

Omar Alrabiah* Venkatesan Guruswami† Ray Li‡

August 2023

Abstract

A simple, recently observed generalization of the classical Singleton bound to list-decoding asserts that rate R codes are not list-decodable using list-size L beyond an error fraction $\frac{L}{L+1}(1-R)$ (the Singleton bound being the case of $L = 1$, i.e., unique decoding). We prove that in order to approach this bound for any fixed $L > 1$, one needs exponential alphabets. Specifically, for every $L > 1$ and $R \in (0, 1)$, if a rate R code can be list-of-L decoded up to error fraction $\frac{L}{L+1}(1-R-\varepsilon)$, then its alphabet must have size at least $\exp(\Omega_{L,R}(1/\varepsilon))$. This is in sharp contrast to the situation for unique decoding where certain families of rate R algebraic-geometry (AG) codes over an alphabet of size $O(1/\varepsilon^2)$ are unique-decodable up to error fraction $(1 - R - \varepsilon)/2$.

Our lower bound is tight up to constant factors in the exponent—with high probability random codes (or, as shown recently, even random linear codes) over $\exp(O_L(1/\varepsilon))$-sized alphabets, can be list-of-L decoded up to error fraction $\frac{L}{L+1}(1 - R - \varepsilon)$.

1 Introduction

The Singeton bound [Sin64] states that a code of rate R cannot uniquely correct a fraction of worst-case errors exceeding $\frac{1}{2}(1 - R)$. The straightforward generalization of this bound to list decoding implies that one cannot do list-of-L decoding (where the decoder must output at most L codewords) beyond an error fraction of $\frac{L}{L+1}(1 - R)$. See Figure 1 for an illustration of this bound, which has been called the *generalized Singleton bound* [ST20].

Our main result is that approaching the generalized Singleton bound within ε requires an alphabet size exponential in $1/\varepsilon$. We say a code $C \subset \Sigma^n$ is (p, L)-list decodable if for every $y \in \Sigma^n$, there are at most L codewords of C within Hamming distance pn from y. Formally, we prove:

Theorem 1.1. *Let $L \geq 2$ be a fixed constant and $R \in (0, 1)$. There exists an absolute constant $\alpha_{L,R}$ such that the following holds for all $\varepsilon > 0$ and all sufficiently large $n \geq \Omega_{L,R}(1/\varepsilon)$. Let C be a code of length n with alphabet size q that is $(\frac{L}{L+1}(1 - R - \varepsilon), L)$-list-decodable. Then, $q \geq 2^{\alpha_{L,R}/\varepsilon}$.*

*Department of EECS, UC Berkeley, Berkeley, CA, 94709, USA. Email: oalrabiah@berkeley.edu. Research supported in part by a Saudi Arabian Cultural Mission (SACM) Scholarship, NSF CCF-2210823 and V. Guruswami's Simons Investigator Award.

†Departments of EECS and Mathematics, and the Simons Institute for the Theory of Computing, UC Berkeley, Berkeley, CA, 94709, USA. Email: venkatg@berkeley.edu. Research supported by a Simons Investigator Award and NSF grants CCF-2210823 and CCF-2228287.

‡Department of EECS, UC Berkeley, Berkeley, CA, 94709, USA. Email: rayyli@berkeley.edu. Research supported by the NSF Mathematical Sciences Postdoctoral Research Fellowships Program under Grant DMS-2203067, and a UC Berkeley Initiative for Computational Transformation award.

ISSN 1433-8092

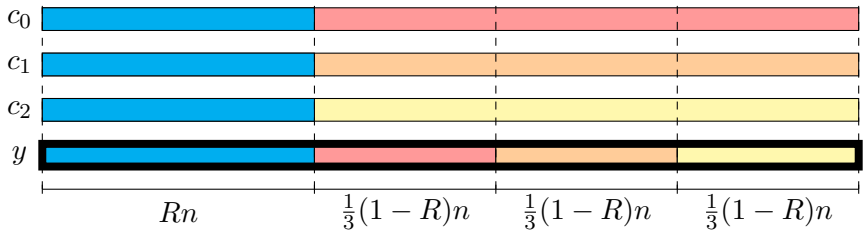

Figure 1: The generalized Singleton bound, illustrated for $L = 2$. In any code of rate R, by pigeonhole, there are three codewords c_0, c_1, c_2 that agree on the first $Rn - O(1)$ coordinates. Then there is a "list-decoding center" y that differs from each of c_0, c_1, c_2 on at most $\frac{2}{3}(1-R)n + O(1)$ coordinates, so the code is not $(\frac{2}{3}(1-R) + o(1), 2)$-list-decodable.

The exponential alphabet size lower bound in Theorem 1.1 is in sharp contrast with the situation for unique decoding (the $L = 1$ case), where, for any desired rate R, certain families of algebraic-geometric (AG) codes over an alphabet of size $O(1/\varepsilon^2)$ allow unique decoding up to an error fraction $(1 - R - \varepsilon)/2$ with rate R [TVZ82, GS95].

The Plotkin bound [Plo60] implies a lower bound of $\Omega(1/\varepsilon)$ for such unique decodability,[1] and AG codes come within a quadratic factor of this bound. However, for list decoding with any fixed list-size $L > 1$, there is no such AG-like polynomial convergence (as a function of alphabet size) to the optimal trade-off $\frac{L}{L+1}(1-R)$. In fact, the convergence is exponentially slow.

Remark 1.2. For a code $C \subset [q]^n$ of rate R, note that a random Hamming ball of radius pn has in expectation $q^{h_q(p)n - o(n)} q^{(R-1)n}$ codewords.[2] For C to be (p, L)-list decodable one must therefore have $h_q(p) \leq 1 - R + o(1)$. (This trade-off is the list-decoding capacity for codes of alphabet size q.) A straightforward computation then implies a lower bound of $q \geq 2^{\Omega_R(\min(L, 1/\varepsilon))}$ on the alphabet size of a family of $(\frac{L}{L+1}(1-R-\varepsilon), L)$-list-decodable codes. For a fixed L, this lower bound does not scale with ε. In comparison, we get an exponential in $1/\varepsilon$ lower bound for any fixed list size L.

Our work builds on the recent work of Brakensiek, Dhar, and Gopi [BDG22] who proved the following result for MDS codes that are list-decodable all the way up to the generalized Singleton bound. Recall that a (linear) MDS code is one whose dimension k, minimum distance d, and block length n satisfy $k + d = n + 1$.

Theorem 1.3 ([BDG22])**.** *Let $R \in (0, 1)$. Any linear MDS code that is $(\frac{2}{3}(1-R), 2)$-list-decodable must have alphabet size at least $2^{\Omega_R(n)}$.*

We present a comparison of our Theorem 1.1 to Theorem 1.3 in Section 3.1. To summarize this comparison, our result generalizes Theorem 1.3 in four ways: our result (i) applies to general (not necessarily linear) codes, (ii) incorporates the gap to capacity ε (iii) removes the MDS assumption (or more generally any assumption on the code distance), and (iv) generalizes to larger L. In essence, our proof distills the proof of Theorem 1.3 to its combinatorial core, and then adds new ideas to enable these generalizations.

[1] The Plotkin bound is typically stated as follows: a code with relative distance at least $\frac{q-1}{q}$ has size $O(n)$. This alphabet dependent version follows by, in a rate R code, pigeonholing to find $O(n)$ codewords agreeing on the first roughly Rn coordinates, and finding (by the Plotkin bound) two codewords that additionally agree on $\frac{1}{q}$ fraction of the remaining coordinates, so that we need $q \geq \Omega(1/\varepsilon)$.

[2] Here, $h_q(x) := x \log_q(q-1) - x \log_q x - (1-x) \log_q(1-x)$ is the q-ary entropy function.

One can show that an exponential in $1/\varepsilon$ alphabet size suffices to approach the generalized Singleton bound within ε. The argument is a simple random coding argument with alterations (see, e.g., [AS16, Chapter 3]), which we present for completeness in Appendix A.1.

Proposition 1.4. *Let $L \geq 1$ be a fixed constant, let $R \in (0,1)$, and let $\varepsilon \in (0,1)$. There exists a code C over an alphabet size $q \leq 2^{O(1/\varepsilon)}$ that is $(\frac{L}{L+1}(1-R-\varepsilon), L)$-list-decodable.*

Thus the lower bound in Theorem 1.1 is tight up to the constant factor $\alpha_{L,R}$ in the exponent (which we did not try to optimize). This is perhaps surprising, because Proposition 1.4 considers the natural random construction, which does *not* give near-optimal alphabet size for unique decoding ($L = 1$). In fact, for unique decoding, the alphabet size exponentially far from optimal: $2^{O(1/\varepsilon)}$ compared to the optimal $\mathrm{poly}(1/\varepsilon)$ achieved by AG codes.

We also point out that, as a consequence of recent work [AGL23] on randomly punctured codes (building on [GZ23]), Proposition 1.4 can be achieved not just with random codes (with alterations), but also with *random linear codes*. [AGL23] shows that, for all $L \geq 2$ and $R, \varepsilon \in (0,1)$, random linear codes over alphabet size $2^{10L/\varepsilon}$ are with high probability $(\frac{L}{L+1}(1-R-\varepsilon), L)$-list-decodable.

2 Preliminaries

We use standard Landau notation $O(\cdot), \Omega(\cdot), \Theta(\cdot)$. We use the notations $O_x(\cdot), \Omega_x(\cdot), \Theta_x(\cdot)$ to mean that a multiplicative factor depending on the variable(s) x is suppressed. All logs are base-2 unless otherwise specified. For integers L, let $[L]$ denote the set $\{1, \ldots, L\}$. Let $h(x) := -x \log x - (1-x)\log(1-x)$ denote the binary entropy function. Let $\binom{n}{\leq r} := \sum_{i=0}^{r} \binom{n}{i}$. We have the binomial approximation

$$2^{h(\alpha)n - o(n)} \mathrm{poly}\, n \leq \binom{n}{\alpha n} \leq \binom{n}{\leq \alpha n} \leq 2^{h(\alpha)n}$$

We also use the Chernoff bound for binomials

$$\mathbf{Pr}[\mathrm{Binomial}(\alpha, m) > (1+\delta)\alpha m] \leq e^{-\frac{\delta^2}{2+\delta}\alpha m}. \tag{1}$$

A *code* $C \subseteq \Sigma^n$ is simply a collection of words of equal length over a fixed alphabet Σ. The *dimension* of a code C is defined to be $k(C) := \log_{|\Sigma|} |C|$ and the *rate* is $R(C) := \frac{k(C)}{n} = \frac{\log_{|\Sigma|}|C|}{n}$. The *distance* of a code C is defined to be $d(C) := \min_{c_1 \neq c_2 \in C} d(c_1, c_2)$, where $d(\cdot,\cdot)$ denotes the Hamming distance between two words. We say a code of length n and dimension k is *Maximum Distance Separable (MDS)* if it has minimum distance $n - k + 1$.

For a string $c \in \Sigma^n$ and a set $A \subset [n]$, we let $c|_A \in \Sigma^{|A|}$ denote the string c restricted to the indices in A.

To highlight the main ideas in Section 3, we prove special cases of Theorem 1.1 for *average-radius-list-decoding*: a code $C \subset [q]^n$ is (p, L)-*average-radius-list-decodable* if, for any distinct $L+1$ codewords $c^{(1)}, \ldots, c^{(L+1)}$ and any vector $y \in [q]^n$, the average Hamming distance from $c^{(1)}, \ldots, c^{(L+1)}$ to y is strictly greater than pn. We observe that average-radius-list-decoding is a strengthening of (ordinary) list-decoding: any (p, L)-average-radius-list-decodable code is also (p, L)-list-decodable.

3 Technical overview

We now introduce the ideas of our main result, Theorem 1.1. Our main result generalizes Theorem 1.3 in four ways, which we outline in Section 3.1. In Section 3.2, Section 3.3, and Section 3.4, we give a few warmup proofs that show (or give a taste of) how we achieve these generalizations. For exposition, we focus our warmup lower bounds on list size $L = 2$ and for average-radius-list-decoding. In Section 4, we give the full proof of our main result, Theorem 1.1.

3.1 Comparison to [BDG22]

Our work generalizes Theorem 1.3 in four ways.

1. **Removing linearity.** The proof of Theorem 1.3 uses that the codes in question have a "higher order MDS" (MDS(3)) property, which is a specific property of the columns of the parity-check matrix [Rot22, BGM22]. They then show that a small alphabet size contradicts the higher order MDS property, and hence the assumption of the code in Theorem 1.3. Higher order MDS codes can only be defined for linear codes, and furthermore, the proof of [BDG22] used several aspects of the linearity of the code. Thus, on the surface, it may seem like Theorem 1.3 could not be generalized to non-linear codes.

   We show that the linearity assumption is in fact not necessary. We show that one can avoid the linear-algebraic aspects of the proof in [BDG22], and that careful applications of the pigeonhole principle suffice to find a bad list-decoding configuration. In our first warmup (Section 3.2), we show how to do this (for average-radius-list-decoding).

2. **Incorporating gap to capacity.** While Theorem 1.3 only proves a lower bound for codes list-decodable exactly up to the generalized Singleton bound, we prove a lower bound even when the code has an ε gap to capacity, showing an alphabet size lower bound for codes approaching list-of-L capacity, for any L. In our second warmup (Section 3.3), we show how to do this (for average-radius-list-decoding).

3. **Removing MDS.** In the connection between list-decodable codes and higher order MDS codes, a code is "MDS($L+1$)" if and only if it exactly achieves the generalized Singleton bound for all $L' \leq L$ [BGM23]. The lower bound in Theorem 1.3 is proved for MDS(3), which requires Theorem 1.3 to assume our code both (i) meets the generalized Singleton bound for $L = 2$ *and* (ii) is MDS. Using arguments similar to the Johnson-bound we show that one can get away with a weaker distance assumption than MDS, and by adjusting our pigeonhole argument, we then can eliminate the assumption entirely. We give a taste of how to do this in our third warmup by showing how to remove the MDS/distance assumption for average-radius-list-decoding (Section 3.4) — it is only a taste, as removing the distance assumption is much easier for average-radius-list-decoding than for ordinary list-decoding.

4. **Generalizing to larger L.** In contrast to lower bounds for higher order MDS codes, generalizing the list-decoding lower bounds to larger L is not immediate. [BDG22] proved an alphabet size lower bound of $q \geq 2^{\Omega_R(n)}$ for MDS(3) codes. Since all MDS(3) codes are also MDS(L) for $L \geq 3$, their lower bound $q \geq 2^{\Omega_R(n)}$ also held for MDS(L) for all $L \geq 3$. However, while the lower bounds for $L = 2$ imply lower bounds for larger L in higher order MDS codes, the same is *not* true for lower bounds for list-of-L decoding: optimal list-of-L decoding does not necessarily imply optimal list-of-L' decoding for $L' > L$.

Generalizing all the above machinery to larger L for average-radius list-decoding follows almost seamlessly from the warmup arguments. However, the list-decoding case is not as immediate. It requires new ideas to remove the distance assumption, care to find the bad list-decoding configuration, and a deliberate balancing of parameters to ensure that all distances from the codewords to the center are simultaneously below the list-decoding radius.

We now give several warmup proofs which show how to achieve these generalizations. Warmup 1 (Section 3.2) achieves the first generalization (removing linearity), Warmup 2 (Section 3.3) achieves the second generalization (incorporating gap to capacity), and Warmup 3 (Section 3.4) achieves the third generalization (removing MDS), all for the easier case of average-radius-list-decoding and $L = 2$. The full proof of Theorem 1.1 incorporates all of these ideas simultaneously and additionally achieves the fourth generalization (all L).

3.2 Warmup 1: A lower bound for exactly optimal list-of-2 decoding

First, we show how to prove a lower bound for all (not-necessarily-linear) codes. In other words, we generalize Theorem 1.3 to a lower bound for all codes. We state and prove the lower bound for average-radius-list-decoding, though, as demonstrated by our main result, Theorem 1.1, a similar lower bound holds for ordinary list-decoding.[3]

Proposition 3.1. *For all $R \in (0,1)$, there exists $\alpha_R > 0$ such that the following holds for sufficiently large n. Any MDS code that is $(\frac{2}{3}(1-R), 2)$-average-radius-list-decodable must have alphabet size $q \geq 2^{\alpha_R \cdot n}$.*

Proof. Fix $I_0 = \{1, 2\}$. Let $\mathcal{F}$ be the collection of all subsets of $[n] \setminus I_0$ of size $k-1$. Clearly $|\mathcal{F}| \geq 2^{\Omega_R(n)}$. Thus, it suffices to prove that $q^2 \geq |\mathcal{F}|/2$. Suppose for contradiction that

$$q^2 < |\mathcal{F}|/2. \tag{2}$$

Consider picking a uniformly random codeword $c \in C$. For each $A \in \mathcal{F}$, let $\mathcal{E}_A$ be the event that another codeword c' agrees with c on A, i.e., $c|_A = c'|_A$. For any $A \in \mathcal{F}$, there at most q^{k-1} possible values of $c|_A$, and thus at most q^{k-1} codewords c for which $c|_A$ uniquely determines c. Hence,

$$\mathbf{Pr}[\neg \mathcal{E}_A] < \frac{q^{k-1}}{q^k} = \frac{1}{q}\,. \tag{3}$$

For each codeword c, define the set $\mathcal{F}_c \coloneqq \{A \in \mathcal{F} : \mathcal{E}_A \text{ occurs}\}$. For each $A \in \mathcal{F}_c$, we can find, by definition, a codeword $f^A(c) \in C \setminus \{c\}$ such that $f^A(c)|_A = c|_A$. By linearity of expectation and (3), we find that

$$\mathbf{E}\left[|\mathcal{F}_c|\right] = \mathbf{E}\left[\sum_{A \in \mathcal{F}} \mathbb{1}\{\mathcal{E}_A\}\right] > \sum_{A \in \mathcal{F}} \left(1 - \frac{1}{q}\right) \geq \frac{|\mathcal{F}|}{2}\,.$$

Hence we can find a codeword $c \in C$ for which $|\mathcal{F}_c| > |\mathcal{F}|/2$. By pigeonhole and (2), there are 2 distinct sets $A_1, A_2 \in \mathcal{F}_c$ such that the codewords $f^{A_1}(c)$ and $f^{A_2}(c)$ agree on the coordinates I_0. These two codewords $f^{A_1}(c)$ and $f^{A_2}(c)$ are distinct: if not, then $f^{A_1}(c) = f^{A_2}(c)$ agrees with c on at least $|A_1 \cup A_2| \geq k$ coordinates, contradicting the assumption that the distance is at least $n - k + 1$.

[3] Our main result, Theorem 1.1, requires $\varepsilon \geq \Omega(1/n)$, but for such $\varepsilon = \Theta(1/n)$, a $(\frac{2}{3}(1-R), 2)$-list-decodable code is certainly $(\frac{2}{3}(1-R-\varepsilon), 2)$-list-decodable, so Theorem 1.1 implies $q \geq 2^{\Omega_{L,R}(n)}$.

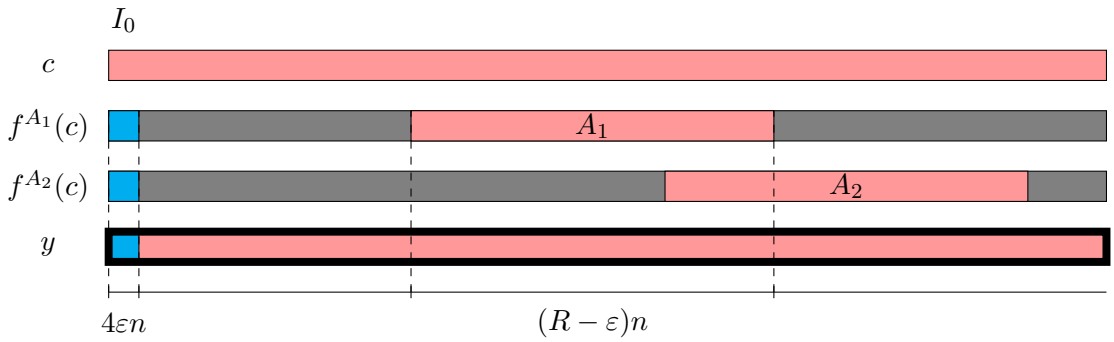

Figure 2: The bad average-radius-list-decoding configuration we search for in Proposition 3.2. The list-decoding center y has distances $4\varepsilon n$, $(1-R-3\varepsilon)n$, and $(1-R-3\varepsilon)n$ from codewords c_0, c_1, and c_2 respectively.

Let y be the word which agrees with $f^{A_1}(c)$ on I_0 (and thus $f^{A_2}(c)$ as well), and agrees with c everywhere else. Word y has total distance at most

$$|I_0| + |[n] \setminus (I_0 \cup A_1)| + |[n] \setminus (I_0 \cup A_2)| = 2 + (n-k-1) + (n-k-1) = 2(n-k) \tag{4}$$

from codewords $c, f^{A_1}(c), f^{A_2}(c)$, contradicting average-radius-list-decoding. Thus, (5) is false, which means $2q^2 \geq |\mathcal{F}| \geq 2^{\Omega_R(n)}$, and so $q \geq 2^{\Omega_R(n)}$. □

3.3 Warmup 2: Relaxing by ε

Next, we show how to prove an alphabet size lower bound of $2^{\Omega_R(1/\varepsilon)}$ (Proposition 3.2), assuming the code has a gap-to-capacity of ε, for average-radius-list-of-2 decoding. We additionally assume the code has near-optimal minimum distance, similar to Proposition 3.1 and Theorem 1.3, and then show how one can remove it in Section 3.4.

The proof of Proposition 3.2 follows nearly the same blueprint as the proof of Proposition 3.1. The new addition will be increasing the size of I_0 to be $\Omega(\varepsilon n)$. That way, the bound in (4) decreases by a factor of $2\varepsilon n$ so that it still contradicts average-radius-list-decodability. In exchange, the bound of $2q^2 \leq |\mathcal{F}|$ is altered to become $2q^{|I_0|} \leq |\mathcal{F}|$.

Proposition 3.2. *For all $R \in (0,1)$, there exists $\alpha_R > 0$ such that the following holds for all $\varepsilon \in (0,1)$ and all sufficiently large $n \geq \Omega_R(1/\varepsilon)$. Let C be a code of rate R with alphabet size q that has minimum distance greater than $(1-R-\varepsilon)n$ and is $(\frac{2}{3}(1-R-\varepsilon), 2)$-average-radius-list-decodable. Then $q \geq 2^{\alpha_R/\varepsilon}$.*

Proof. By adjusting α_R, it suffices to consider ε sufficiently small compared to R. Fix $I_0 = \{1, 2, \ldots, 4\varepsilon n\}$. Let $\alpha := R - \varepsilon$ and $\beta := R + \varepsilon$. For any two subsets $A, B \subseteq [n] \setminus I_0$ satisfying $|A| = |B| = \alpha n$ and $|A \cup B| \leq \beta n$, notice that $|A \setminus B| = |B \setminus A| \leq (\beta - \alpha)n = 2\varepsilon n$. Since the tuple of sets $(A, A \setminus B, B \setminus A)$ determine B, the number of possible subsets B for any given A is therefore at most $\binom{(1-4\varepsilon)n}{\leq 2\varepsilon n}^2$. Thus, by greedily choosing subsets, we can find a family $\mathcal{F}$ of $\binom{(1-4\varepsilon)n}{\alpha n} / \binom{(1-4\varepsilon)n}{\leq 2\varepsilon n}^2 \geq 2^{\Omega_R(n)}$ subsets of $[n] \setminus I_0$ such that each subset has size αn and any pairwise union has size at least βn.

It suffices to prove that $q^{|I_0|} \geq |\mathcal{F}|/2$. Suppose for contradiction that

$$q^{|I_0|} < |\mathcal{F}|/2. \tag{5}$$

Consider picking a uniformly random codeword $c \in C$. For each $A \in \mathcal{F}$, let $\mathcal{E}_A$ be the event that another codeword c' agrees with c on A, i.e., $c|_A = c'|_A$. For any $A \in \mathcal{F}$, there at most $q^{\alpha n}$ possible values of $c|_A$, and thus at most $q^{\alpha n}$ codewords c for which $c|_A$ uniquely determines c. Hence,

$$\mathbf{Pr}[\neg \mathcal{E}_A] < \frac{q^{\alpha n}}{q^k} = \frac{1}{q^{\varepsilon n}} . \tag{6}$$

For each codeword c, define the set $\mathcal{F}_c \coloneqq \{A \in \mathcal{F} : \mathcal{E}_A \text{ occurs}\}$. For each $A \in \mathcal{F}_c$, we can find, by definition, a codeword $f^A(c) \in C \setminus \{c\}$ such that $f^A(c)|_A = c|_A$. By linearity of expectation and (6), we thus find that

$$\mathbf{E}\left[|\mathcal{F}_c|\right] = \mathbf{E}\left[\sum_{A \in \mathcal{F}} \mathbb{1}\{\mathcal{E}_A\}\right] > \sum_{A \in \mathcal{F}} \left(1 - \frac{1}{q^{\varepsilon n}}\right) \geq \frac{|\mathcal{F}|}{2} .$$

Hence we can find a codeword $c \in C$ for which $|\mathcal{F}_c| > |\mathcal{F}|/2$. By pigeonhole and (5), there are 2 distinct sets $A_1, A_2 \in \mathcal{F}_c$ such that the codewords $f^{A_1}(c)$ and $f^{A_2}(c)$ agree on the coordinates I_0. These two codewords $f^{A_1}(c)$ and $f^{A_2}(c)$ are distinct: if not, then $f^{A_1}(c) = f^{A_2}(c)$ agrees with c on at least $|A_1 \cup A_2| \geq \beta n = (R + \varepsilon)n$ coordinates, contradicting the assumption that the distance is greater than $(1 - R - \varepsilon)n$.

Let y be the word which agrees with $f^{A_1}(c)$ on I_0 (and thus $f^{A_2}(c)$ as well), and agrees with c everywhere else (see Figure 2). Word y has total distance at most

$$\begin{aligned} |I_0| + |[n] \setminus (I_0 \cup A_1)| + |[n] \setminus (I_0 \cup A_2)| &= 4\varepsilon n + (1 - R - 3\varepsilon)n + (1 - R - 3\varepsilon)n \\ &= 2(1 - R - \varepsilon)n \end{aligned}$$

from codewords $c, f^{A_1}(c), f^{A_2}(c)$, contradicting average-radius-list-decoding. Thus, (5) is false, which means $2q^{4\varepsilon n} \geq |\mathcal{F}| \geq 2^{\Omega_R(n)}$, and so $q \geq 2^{\Omega_R(1/\varepsilon)}$. □

3.4 Warmup 3: Removing the distance assumption

In this section, we prove Proposition 3.3, which is the same as Proposition 3.2 but with the distance assumption removed. To remove it, we simply observe that the minimum distance is already nearly satisfied in any $(\frac{2}{3}(1 - R - \varepsilon), 2)$-average-radius-list-decodable code. This is *not* true for ordinary list-decoding, so we need additional ideas to remove the minimum distance condition in the general lower bound, Theorem 1.1, but we include this much simplier proof to illustrate the high level structure of the proof.

Proposition 3.3. *For all $R \in (0, 1)$, there exists $\alpha_R > 0$ such that the following holds for all $\varepsilon \in (0, 1)$ and all sufficiently large $n \geq \Omega_R(1/\varepsilon)$. Let C be a code of rate R with alphabet size q that is $(\frac{2}{3}(1 - R - \varepsilon), 2)$-average-radius-list-decodable. Then $q \geq 2^{\alpha_R/\varepsilon}$.*

To prove Proposition 3.3, we need the following simple lemma.

Lemma 3.4. *In any $(p, 2)$-average-radius-list-decodable code, each codeword has at most 1 other codeword within distance $\frac{3p}{2}n$.*

Proof. If there are 2 codewords c_1 and c_2 both within distance $3pn/2$ of codeword c, then the word c has average distance pn from the codewords c, c_1, c_2, contradicting $(p, 2)$-average-radius-list-decodability. □

Now we can prove Proposition 3.3.

Proof of Proposition 3.3. Let C be a $(\frac{2}{3}(1 - R - \varepsilon), 2)$-average-radius-list-decodable code. By Lemma 3.4, each codeword has at most 1 other codeword within distance $(1 - R - \varepsilon)n$. Thus, by choosing codewords greedily, C has a subcode C' of size at least $|C|/2$ that both is $(\frac{2}{3}(1-R-\varepsilon), 2)$-average-radius-list-decodable and has minimum distance greater than $(1 - R - \varepsilon)n$. Subcode C' has rate at least $R' = R - (1/n)$. Applying Proposition 3.2 with subcode C', rate $R' = R - (1/n)$, and $\varepsilon' = \varepsilon + (1/n)$ gives the desired bound on the alphabet size q. □

4 The full lower bound: all L and (ordinary) list-decoding.

We now present the full proof of our main result, Theorem 1.1. To do so, we need to combine the ideas in the Warmups 1, 2, and 3, and add some additional ideas.

4.1 Additional Ingredients

First, we need to generalize the warmups from average-radius list-decoding to (ordinary) list-decoding. To do so, we use similar ideas to [BDG22], but distill those ideas down to their combinatorial essence. The idea is to choose our bad list-decoding configuration by first choosing a bad average-radius-list-decoding configuration $y, c_0, c_1, \ldots, c_L$, as in the warmups. However, because the list-decoding center y was much closer to c_0 than to each of $c_1, \ldots, c_L$, we will instead balance out the distances between them by "transferring" agreements between y and c_0 (of which there are almost n) to agreements between y and $c_1, \ldots, c_L$, until the y has a similar number of agreements with each of $c_0, c_1, \ldots, c_L$. Specifically, the parts where we will transfer agreements from c_0 to $c_1, \ldots, c_L$ will be the intervals $I_1, \ldots I_L$ that we define in the proof: for $p = \frac{L}{L+1}(1 - R - \varepsilon)$, we set aside the first pn coordinates for the intervals $I_0, I_1, \ldots, I_L$, meaning that the agreement sets $A_1, \ldots, A_L$ will have to be subsets of $\{pn + 1, \ldots, n\}$.

Next, we need to remove the distance requirement for (ordinary) list-decoding, which is more difficult than removing the distance requirement for average-radius-list-decoding in Section 3.4. To do so, we need a lemma similar to Lemma 3.4 that shows that a (p, L)-list-decodable code has a subcode with large distance and essentially the same rate. Clearly, (p, L)-list-decoding implies that every codeword is within distance pn of at most L codewords, so we can essentially assume our code has distance pn. For technical reasons, this is not good enough. In Lemma 4.1, we show, more strongly, that a (p, L)-list-decodable code has a large subcode with distance $(p + \frac{p^L}{2L})n$. Therefore, to show Theorem 1.1, it suffices for us to show it with the additional assumption that the minimum distance is $(p + \frac{p^L}{2L})n$, which is what we show in Theorem 4.3.

To prove Theorem 4.3, we need additional ideas. In the warmup arguments, we needed to assume near-optimal minimum distance (namely $(1 - R - O(\varepsilon))n$), to show that the codewords $c_1, \ldots, c_L$ we find via the pigeonhole argument are pairwise distinct. To accommodate our weaker assumption of distance $p + \frac{p^L}{2L}$, we note that our pigeonholing actually gives substantially more than L codewords, and in particular, certainly at least $L \cdot W$ codewords for some constant W. Then, it is enough to show that our pigeonholing never produces W codewords that are all equal (rather than showing we never get two equal codewords). Here, a relaxed condition on our set system $\mathcal{F}$

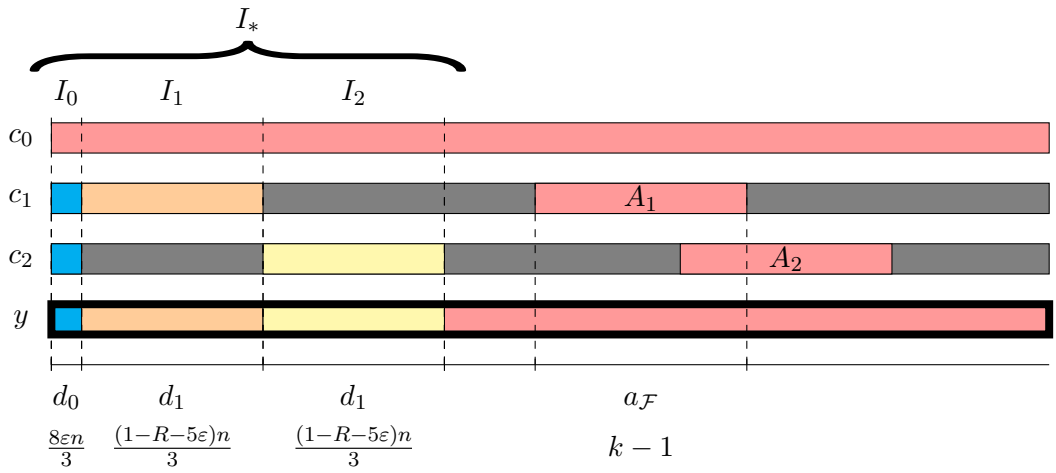

Figure 3: The agreement pattern we search for via pigeonhole in our upper bound, for $L = 2$. Codeword c_0 differs from y in at most $d_0 + 2d_1$ places and codewords c_1 and c_2 differ from y in at most $n - d_0 - d_1 - a_{\mathcal{F}}$ places.

suffices, namely that the sets have very large W-wise, rather than pairwise, unions (see Lemma 4.2 below).

We now present the lemmas described above. The first one implies that any (p, L)-list-decodable code has a large subcode with distance $(p+\frac{p^L}{2L})n$. A similar lemma appears in [GN14, Theorem 15], and a lemma simliar in spirit appears in [GST22, Theorem 6.1]. We defer the proof to Appendix A.2.

Lemma 4.1. *Let $p \in (0, 1)$. In any (p, L)-list-decodable code C, every codeword is within relative distance $\alpha := p + \frac{p^L}{2L}$ of at most $L' = O(L^2/p)$ codewords. Consequently, C has a subcode of size at least $|C|/(L' + 1)$ that has relative distance at least α.*

The next lemma says that, for sufficiently large W, we can choose a large family of sets with very large W-wise unions. The proof is a straightforward probabilistic argument, and we defer it to Appendix A.3.

Lemma 4.2. *For all $1 > \beta > \alpha > 0$, for all positive integers m, there exists a constant $W = O\left(\log(1-\beta)/\log(1-\alpha)\right)$ and a family $\mathcal{F}$ of $2^{\Omega(m(1-\beta)\log(1-\alpha)/\log(1-\beta))}$ subsets of $[m]$, each of size αm, such that all W-wise unions of subsets are of size at least βm.*

4.2 Proof

Now, we put the above ideas all together, and generalize to all L, to give the full theorem. By Lemma 4.1, the following theorem implies Theorem 1.1

Theorem 4.3. *Let $L \geq 2$ be a fixed constant and $R \in (0, 1)$ and $\varepsilon \in (0, 1)$, and let n be sufficiently large. Let C be a code of length n with alphabet size q that is (p, L)-list-decodable for $p = \frac{L}{L+1}(1 - R - \varepsilon)$. Suppose also that C has minimum distance at least $(p + \frac{p^L}{2L})n$. Then $q \geq 2^{\Omega_{L,R}(1/\varepsilon)}$.*

Proof. Since we suppress factors of R and L, it suffices to consider ε sufficiently small. With

hindsight, define the following parameters

$$a_{\mathcal{F}} \coloneqq k-1, \tag{7}$$
$$a_{\cup} \coloneqq \left(1-p-\frac{p^L}{4L}\right)n, \tag{8}$$
$$d_1 \coloneqq \left(\frac{1-R-5\varepsilon}{L+1}\right)n,$$
$$d_0 \coloneqq \left(\frac{4L\varepsilon}{L+1}\right)n,$$

We remark that the parameters d_0 and d_1 are chosen to satisfy the following equation and inequalities specifically:

$$d_0 + Ld_1 = pn, \tag{9}$$
$$n - d_0 - d_1 - a_{\mathcal{F}} \leq pn, \tag{10}$$
$$d_0 \leq 4\varepsilon n. \tag{11}$$

Now, let $I_0, \ldots, I_L$ be consecutive subintervals of $[n]$ (in that order), such that interval $I_0 = \{1, \ldots, d_0\}$ has size d_0, and intervals $I_1, \ldots, I_L$ have size d_1. Define $I_* \coloneqq I_0 \cup I_1 \cup \cdots \cup I_L$. Note that $|I_*| = d_0 + Ld_1 =^{(9)} pn$.

Since C has positive rate R and minimum distance $(p + p^L/2L)n$, then for small enough ε, we find by applying (9) and the Singleton bound that

$$1 - \left(\frac{1-R}{2}\right)^{L-1} \geq \frac{a_{\cup}}{n-pn} \geq \frac{R}{1-p} \geq \frac{a_{\mathcal{F}}}{n-pn} \geq \frac{(L+1)R}{1+(L+1)R}\ . \tag{12}$$

That is, if we define $\alpha \triangleq a_{\mathcal{F}}/(n-pn)$ and $\beta \triangleq a_{\cup}/(n-pn)$, then (12) implies that α and β lie in an interval contained in $(0,1)$ that is completely determined by R and L and not on ε. By applying Lemma 4.2 on the ground set $[n] \setminus I_*$, we get a constant $W = O_{L,R}(1)$ and a family $\mathcal{F}$ of $2^{\Omega_{L,R}(n)}$ subsets of $[n] \setminus I_*$, each of size $a_{\mathcal{F}}$, such that every W-wise union of sets in $\mathcal{F}$ has size at least $a_{\cup}$. Here, (12) ensures that W and the number of subsets $2^{\Omega_{L,R}(n)}$ do not depend on ε.

Consider picking a uniformly random codeword $c \in C$. For each $A \in \mathcal{F}$, let $\mathcal{E}_A$ be the event that another codeword c' agrees with c on A, i.e., $c|_A = c'|_A$. For any $A \in \mathcal{F}$, there at most $q^{a_{\mathcal{F}}}$ possible values of $c|_A$, and thus at most $q^{a_{\mathcal{F}}}$ codewords c for which $c|_A$ uniquely determines c. Hence,

$$\mathbf{Pr}[\neg\mathcal{E}_A] < \frac{q^{a_{\mathcal{F}}}}{q^k} \overset{(7)}{=} \frac{1}{q}\ . \tag{13}$$

For each codeword c, define the set $\mathcal{F}_c \coloneqq \{A \in \mathcal{F} : \mathcal{E}_A \text{ occurs}\}$. For each $A \in \mathcal{F}_c$, we can find, by definition, a codeword $f^A(c) \in C \setminus \{c\}$ such that $f^A(c)|_A = c|_A$. By linearity of expectation and (13), we thus find that

$$\mathbf{E}\left[|\mathcal{F}_c|\right] = \mathbf{E}\left[\sum_{A\in\mathcal{F}} \mathbb{1}\{\mathcal{E}_A\}\right] > \sum_{A\in\mathcal{F}}\left(1-\frac{1}{q}\right) \geq \frac{|\mathcal{F}|}{2}\ .$$

Hence we can find a codeword $c \in C$ for which $|\mathcal{F}_c| > |\mathcal{F}|/2$. Fix this codeword $c_0 = c$. To prove Theorem 4.3, it suffices to prove that $2 \cdot W \cdot L \cdot q^{d_0} \geq |\mathcal{F}|$. Suppose for contradiction that

$$W \cdot L \cdot q^{d_0} < |\mathcal{F}|/2 \tag{14}$$

By pigeonhole and (14), there are WL sets $A_1, \ldots, A_{WL} \in \mathcal{B}_{c_0}$ such that $f^{A_1}(c_0), \ldots, f^{A_{WL}}(c_0)$ agree on the coordinates in I_0. Further, no W of the codewords can be equal: if, for example, $f^{A_1}(c_0) = f^{A_2}(c_0) = \cdots = f^{A_W}(c_0)$, then this codeword $f^{A_1}(c_0)$ agrees with c_0 on the coordinates $\cup_{i=1}^{W} A_i$, which by construction of $\mathcal{F}$ has size at least $a_\cup =^{(8)} (1-p-p^L/4L)n$. Thus, we have found two distinct codewords that disagree on at most $(p + p^L/4L)n$ positions, which contradicts the minimum distance assumption of our code. Thus, no W codewords among $f^{A_1}(c_0), \ldots, f^{A_{WL}}(c_0)$ are equal. In particular, this implies that we have at least L distinct codewords. Without loss of generality, say the codewords $c_1 \coloneqq f^{A_1}(c_0), c_2 \coloneqq f^{A_2}(c_0), \ldots, c_L \coloneqq f^{A_L}(c_0)$ are pairwise distinct.

Let $y \in [q]^n$ be the list-decoding center that agrees with c_1 on coordinates I_0 (and thus $c_2, \ldots, c_L$), agrees with c_j on coordinates I_j for $j = 1, \ldots, L$, and agrees with c_0 elsewhere (see Figure 3). Let us analyze the distance of y to the $L+1$ codewords $c_0, c_1, \ldots, c_L$ in two cases:

1. First, by construction of y, the codeword c_0 can only disagree with y on I_*. Thus the distance between y and c_0 is most $|I_*| = m$, which is at most pn by (9).
2. For $j = 1, \ldots, L$, by construction of y, codeword c_j agrees with y on I_0, I_j, and A_j. Thus, c_j disagrees with y on at most $n - |I_0| - |I_j| - |A_j| = n - d_0 - d_1 - a_\mathcal{F}$ coordinates, which is at most pn by (10).

Thus, we have found $L+1$ distinct codewords each with Hamming distance at most pn from y, contradicting that C is (p, L)-list-decodable. Hence, (14) is false, giving us our desired lower bound $q \geq (\frac{|\mathcal{F}|}{2WL})^{1/d_0} \geq^{(11)} 2^{\Omega_R(1/\varepsilon)}$. □

Proof of Theorem 1.1. Let C be a code that is (p, L)-list-decodable for $p = \frac{L}{L+1}(1 - R - \varepsilon)$. By Lemma 4.1 and for ε sufficiently small, every codeword C has a subcode C' of C with minimum distance $(p + \frac{p^L}{2L})n$ and rate $R' = R - \frac{\log(L'+1)}{n} = R' - o(1)$. Apply Theorem 4.3 to the subcode C' with rate R' and $\varepsilon' = \varepsilon + \frac{\log(L'+1)}{n} \leq \varepsilon + o(1)$, and use that n is sufficiently large to obtain the result. □

5 Concluding Remarks

As alluded to in Remark 1.2, Theorem 1.1 focuses on the case when L is a fixed constant independent of $1/\varepsilon$. It nonetheless leaves open the question of showing an alphabet size lower bound for L growing with $1/\varepsilon$.

Question 5.1. Can we show that all codes (for sufficiently large n) that are $(\frac{L}{L+1}(1 - R - \varepsilon), L)$-list-decodable require alphabet size $q \geq 2^{\Omega_R(1/\varepsilon)}$ (independent of L)?

In the most general case, our current methods give a constant $\alpha_{L,R}$ in Theorem 1.1 that is at most $\exp(-O_R(L))$. However, in special cases, we get better bounds.

For average-radius list-decoding, such an L-independent alphabet size lower bound of $q \geq 2^{\Omega_R(1/\varepsilon)}$ follows from the warmup arguments in Sections 3.2, 3.3, and 3.4. In particular, one can check that Proposition 3.2 holds if we assume $(\frac{L}{L+1}(1-R-\varepsilon), L)$-average-radius-list-decoding, again with minimum distance $(1 - R - \varepsilon)n$, and we again get an alphabet size lower bound of $q \geq 2^{\alpha_R/\varepsilon}$, independent of L (the lower bound will in fact be $\binom{n}{Rn}^{1-\delta}$ for some $\delta = \delta(\varepsilon) \to 0$ decreasing with

$\varepsilon \to 0$.).[4] This implies that Proposition 3.3 also generalizes to give an optimal alphabet size lower bound of $q \geq 2^{\Omega_R(1/\varepsilon)}$ for $(\frac{L}{L+1}(1-R-\varepsilon), L)$-average-radius-list-decoding (without an additional distance assumption).

For ordinary list-decoding, if we can assume a fixed relative distance $\delta > p$ depending only on R, then $\alpha_{L,R}$ can be improved to $\Omega_R(1/L)$ by following the same argument as in Theorem 4.3. Combining this with the list-decoding capacity theorem (Remark 1.2) gives an L-independent alphabet size lower bound of $q \geq 2^{\Omega_R(1/\sqrt{\varepsilon})}$.

References

[AGL23] Omar Alrabiah, Venkatesan Guruswami, and Ray Li. Randomly punctured Reed–Solomon codes achieve list-decoding capacity over linear-sized fields. *arXiv preprint arXiv:2304.09445*, 2023.

[AS16] Noga Alon and Joel H Spencer. *The probabilistic method*. John Wiley & Sons, 2016.

[BDG22] Joshua Brakensiek, Manik Dhar, and Sivakanth Gopi. Improved field size bounds for higher order MDS codes. *arXiv preprint arXiv:2212.11262*, 2022.

[BGM22] Joshua Brakensiek, Sivakanth Gopi, and Visu Makam. Lower bounds for maximally recoverable tensor codes and higher order MDS codes. *IEEE Transactions on Information Theory*, 68(11):7125–7140, 2022.

[BGM23] Joshua Brakensiek, Sivakanth Gopi, and Visu Makam. Generic Reed-Solomon codes achieve list-decoding capacity. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing (STOC)*, pages 1488–1501, 2023.

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

[GS95] Arnaldo Garcia and Henning Stichtenoth. A tower of Artin-Schreier extensions of function fields attaining the Drinfeld-Vlădut bound. *Inventiones mathematicae*, 121(1):211–222, 1995.

[GST22] Eitan Goldberg, Chong Shangguan, and Itzhak Tamo. Singleton-type bounds for list-decoding and list-recovery, and related results. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 2565–2570. IEEE, 2022.

[GZ23] Zeyu Guo and Zihan Zhang. Randomly punctured Reed-Solomon codes achieve the list decoding capacity over polynomial-size alphabets. In *FOCS 2023, to appear, arXiv preprint arXiv:2304.01403*, 2023.

[Plo60] Morris Plotkin. Binary codes with specified minimum distance. *IRE Transactions on Information Theory*, 6(4):445–450, 1960.

[4] More details: We still take $|I_0| = 4\varepsilon, \alpha = R - \varepsilon, \beta = R + \varepsilon$, and $\mathcal{F}$ to be the same family of subsets. We prove $q^{|I_0|} \geq |\mathcal{F}|/L$, and, assuming not (for the sake of contradiction), we find a codeword c and L codewords $f^{A_1}(c), \ldots, f^{A_L}(c)$ agreeing with each other on I_0 and where $f^{A_i}(c)$ agrees with c on A_i. As $|A_i \cup A_j| \geq (R+\varepsilon)n$, these codewords are pairwise distinct or else we contradict the code distance. Now we choose the list-decoding center y in the same way, and it has total distance $|I_0| + L \cdot (n - |I_0| - \alpha n) = 4\varepsilon n + L(1 - R - 3\varepsilon)n < L(1 - R - \varepsilon)n$ to these codewords, contradiction.

[Rot22] Ron M Roth. Higher-order MDS codes. *IEEE Transactions on Information Theory*, 68(12):7798–7816, 2022.

[Sin64] Richard Singleton. Maximum distance q-nary codes. *IEEE Trans. Inform. Theory*, 10(2):116–118, April 1964.

[ST20] 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.

[TVZ82] Michael A Tsfasman, SG Vlădut, and Th Zink. Modular curves, Shimura curves, and Goppa codes, better than Varshamov-Gilbert bound. *Mathematische Nachrichten*, 109(1):21–28, 1982.

A Deferred Proofs

A.1 Proof of Proposition 1.4

In this appendix, we prove Proposition 1.4. We remark that the proof we present seamlessly extends to the notion of average-radius-list-decodabililty.

Proof of Proposition 1.4. Fix an alphabet Σ of size q, and set $N := \lfloor q^{Rn} \rfloor$ and $p := \frac{L}{L+1}(1-R-\varepsilon)$. Consider a code $C = \{c^{(1)}, \ldots, c^{(N)}\} \subseteq \Sigma^n$ where each $c^{(\ell)}$ is chosen independently and uniformly at random from Σ^n. Set $p := \frac{L}{L+1}(1-R-\varepsilon)$. For any set $I \subseteq [N]$ of size $L+1$ and word $w \in \Sigma^n$, let $\mathcal{B}_I^w$ be the event that $d(c^{(\ell)}, w) \leq pn$ for all $\ell \in I$. Define the sets $A_\ell := \{i \in [n] : c_i^{(\ell)} = w_i\}$ and $E_i := \{\ell \in I : c_i^{(\ell)} = w_i\}$. Then by double counting, we find that

$$\sum_{i=1}^{n} |E_i| = \sum_{\ell \in I} |A_\ell| = \sum_{\ell \in I} (n - d(c^{(\ell)}, w)) \geq (1 + LR + L\varepsilon)n \ . \tag{15}$$

Consider the hypergraph $\mathcal{H}$ with vertices $V(\mathcal{H}) = I$ and hyperedges $E(\mathcal{H}) = \{E_1, \ldots, E_n\}$). Let $\mathcal{X}_I^{\mathcal{H}}$ be the event that $c_i^{(\ell)} = w_i$ for all $\ell \in E_i$ and $i \in [n]$. Since each $c^{(\ell)}$ is independently and uniformly chosen from Σ^n, each hyperedge E_i 'imposes' $|E_i|$ constraints. Thus by using Inequality (15) and union bounding over all choices of $\mathcal{H}$, we find that

$$\begin{aligned} \mathbf{Pr}[\mathcal{B}_i^w] &\leq \mathbf{Pr}\left[\exists \text{ hypergraph } \mathcal{H} \text{ such that } \mathcal{X}_I^{\mathcal{H}} \text{ occurs}\right] \\ &\leq 2^{(L+1)n} q^{-\sum_{i=1}^{n} |E_i|} \\ &\leq 2^{(L+1)n} q^{-(1+LR+L\varepsilon)n} \ . \end{aligned} \tag{16}$$

Now, pick $q \geq 2^{3/\varepsilon}$. Using Inequality (16), we conclude that

$$\begin{aligned} \mathbf{E}\left[\sum_{\substack{w \in \Sigma^n \\ I \subseteq [N], |I| = L+1}} \mathbb{1}\{\mathcal{B}_I^w\}\right] &\leq q^n \cdot q^{(L+1)Rn} \cdot 2^{(L+1)n} q^{-(1+LR+L\varepsilon)n} \\ &= q^{Rn} \cdot 2^{(L+1)n} \cdot q^{-L\varepsilon n} \\ &\leq q^{Rn} \cdot 2^{-Ln} \ . \end{aligned}$$

Thus we can find a code C of rate R such that $\mathcal{B}_I^w$ occurs for at most $q^{Rn} \cdot 2^{-Ln}$ choices of w and I. For each such pair (w, I), fix an index $i_{w,I} \in I$ and consider the expurgated code $C' \coloneqq C \setminus C_b$. Then $|C'| \geq q^{Rn}(1 - 2^{-Ln}) = q^{n(R-o(1))}$. Furthermore, since we removed all the 'bad' codewords from, none of the events $\mathcal{B}_I^w$ now occur in C', implying that C' is a $\left(\frac{L}{L+1}(1 - R - \varepsilon), L\right)$-list-decodable code. □

A.2 Proof of Lemma 4.1

Proof of Lemma 4.1. Let $M = \lceil \frac{L^2}{p} \rceil$. We may assume without loss of generality that the all-0s string 0^n is in the code, and by symmetry it suffices to show that there are at most $L + M - 1$ codewords within distance at most αn from 0^n, i.e., (Hamming) weight at most αn.

There are clearly at most L codewords of weight at most pn be the list-decoding property (centered at 0). Suppose for contradiction there are M codewords $c_1, \ldots, c_M$ of weight between pn and αn.

We claim there are L nonzero codewords $c_{j_1}, \ldots, c_{j_L}$ such that

$$|\operatorname{supp}(c_{j_1}) \cap \cdots \cap \operatorname{supp}(c_{j_L})| \geq p^L n.$$

For $i = 1, \ldots, n$, let a_i denote the number of $j \in [M]$ such that $i \in \operatorname{supp}(c_j)$. Let T denote the number of tuples $(i, j_1, \ldots, j_L) \in [n] \times [M]^L$ with $j_1 < j_2 < \cdots < j_L$ such that $i \in \operatorname{supp}(c_{j_1}) \cap \cdots \cap \operatorname{supp}(c_{j_L})$. By double counting,

$$\binom{M}{L} \max_{j_1 < \cdots < j_L} |\operatorname{supp}(c_{j_1}) \cap \cdots \cap \operatorname{supp}(c_{j_L})| \geq T = \sum_{i=1}^{n} \binom{a_i}{L} \geq n \cdot \binom{pM}{L}$$

The last inequality uses convexity of $\binom{\cdot}{L}$ and that $\sum_{i=1}^n a_i = \sum_{j=1}^M |\operatorname{supp}(c_j)| \geq M \cdot pn$. Rearranging, we have

$$\begin{aligned}
\max_{j_1 < \cdots < j_L} |\operatorname{supp}(c_{j_1}) \cap \cdots \cap \operatorname{supp}(c_{j_L})| &\geq n \cdot \frac{\binom{pM}{L}}{\binom{M}{L}} \\
&\geq n \cdot \frac{(pM)(pM-1)\cdots(pM-L+1)}{M^L} \\
&\geq n \cdot p^L \cdot \left(1 - \frac{\binom{L}{2}}{pM}\right) \\
&\geq n \cdot \frac{p^L}{2}. \qquad (17)
\end{aligned}$$

by the bound on M. The third inequality uses that $(1-a_1)(1-a_2)\cdots(1-a_n) \geq 1-(a_1+\cdots+a_n)$.

Without loss of generality, (17) is realized by

$$|\operatorname{supp}(c_1) \cap \operatorname{supp}(c_2) \cap \cdots \cap \operatorname{supp}(c_L)| \geq \frac{p^L}{2} n. \qquad (18)$$

Now consider the codewords $0, c_1, c_2, \ldots, c_L$, and let $S_1, S_2, \ldots, S_L$ be pairwise disjoint subsets of $\operatorname{supp}(c_1) \cap \cdots \cap \operatorname{supp}(c_L)$ of size $(\alpha - p)n = \frac{p^L}{2L} n$. There sets exist because of (18). Consider the word w such that w agrees with c_j on S_j for $j = 1, \ldots, L$, and is 0 otherwise. Note that the distance from w to 0 is at most $|S_1 \cup S_2 \cdots \cup S_j| \leq \frac{p^L}{2} n < pn$. The distance from w to c_1 is at most

$$|\operatorname{supp}(c_1) \cup S_2 \cup S_3 \cup \cdots \cup S_L \setminus S_1| = |\operatorname{supp}(c_1) \setminus S_1| = |\operatorname{supp}(c_1)| - |S_1| \leq pn.$$

Thus, we've found $L+1$ codewords $0, c_1, c_2, \ldots, c_L$ within distance pn of w, contradicting list-decodability. Hence, there are at most $M-1$ codewords with weight between pn and αn, so there are at most $L+M-1$ codewords with weight at most αn, as desired.

The subcode of C with minimum distance αn can be chosen greedily from C. □

A.3 Proof of Lemma 4.2

Proof of Lemma 4.2. With hindsight, let W be the positive integer such that $(1-\beta)/2 \le (1-\alpha)^W < (1-\beta)/(2-2\alpha)$, and let $M = 2^{(1-\beta)m/6W}$. Pick M sets independently by including each element of $[m]$ independently with probability $\alpha_0 = \alpha - \frac{1}{m^{1/3}}$. By standard concentration arguments, with high probability, at most $o(M)$ of the sets are of size more than αm.

The probability we have some W-wise union of size less than βm is at most

$$\begin{aligned}\binom{M}{W} \cdot \mathbf{Pr}\left[\text{Binomial}(1-(1-\alpha_0)^W, m) < \beta m\right] &= \binom{M}{W} \cdot \mathbf{Pr}\left[\text{Binomial}\left((1-\alpha_0)^W, m\right) > (1-\beta)m\right] \\ &\le \binom{M}{W} \cdot \mathbf{Pr}\left[\text{Binomial}\left((1-\beta)/2, m\right) > (1-\beta)m\right] \\ &\le M^W \cdot e^{-(1-\beta)m/6} \ll 1.\end{aligned}$$

where above used the Chernoff bound (1) with $\delta = 1$. Thus, with high probability, all W-wise unions have size at least βm and at least $M - o(M)$ sets are of size at most αm. Hence, some choice of sets exists. Taking the $M - o(M)$ sets of size at most αm, and appending arbitrary elements to them until they have size exactly αm, gives our desired family. □

ECCC ISSN 1433-8092
https://eccc.weizmann.ac.il