A Lower Bound on List Size for List Decoding*

Venkatesan Guruswami†
Dept. of Computer Science & Engineering
University of Washington
Seattle, WA
venkat@cs.washington.edu

Salil Vadhan‡
Division of Engineering & Applied Sciences
Harvard University
Cambridge, MA
salil@eecs.harvard.edu

July 13, 2005

Abstract

A q-ary error-correcting code $C \subseteq \{1, 2, \ldots, q\}^n$ is said to be *list decodable* to radius ρ with list size L if every Hamming ball of radius ρ contains at most L codewords of C. We prove that in order for a q-ary code to be list-decodable up to radius $(1 - 1/q)(1 - \varepsilon)n$, we must have $L = \Omega(1/\varepsilon^2)$. Specifically, we prove that there exists a constant $c_q > 0$ and a function f_q such that for small enough $\varepsilon > 0$, if C is list-decodable to radius $(1 - 1/q)(1 - \varepsilon)n$ with list size c_q/ε^2, then C has at most $f_q(\varepsilon)$ codewords, independent of n. This result is asymptotically tight (treating q as a constant), since such codes with an exponential (in n) number of codewords are known for list size $L = O(1/\varepsilon^2)$.

A result similar to ours is implicit in Blinovsky [Bli] for the binary ($q = 2$) case. Our proof works for all alphabet sizes, and is technically and conceptually simpler.

1 Introduction

List decoding was introduced independently by Elias [Eli1] and Wozencraft [Woz] as a relaxation of the classical notion of error-correction by allowing the decoder to output a *list* of possible answers. The decoding is considered successful as long as the correct message is included in the list. We point the reader to the paper by Elias [Eli2] for a good summary of the history and context of list decoding.

The basic question raised by list decoding is the following: How many errors can one recover from, when constrained to output a list of small size? The study of list decoding strives to (1)

*An extended abstract of this paper will appear in *RANDOM '05* [GV].

†Supported in part by NSF Career Award CCF-0343672.

‡Supported by NSF grant CCF-0133096, ONR grant N00014-04-1-0478, US-Israel BSF grant 2002246, and a Sloan Research Fellowship. Work done in part while a Fellow at the Radcliffe Institute for Advanced Study.

understand the combinatorics underlying this question, (2) realize the bounds with explicit constructions of codes, and (3) list decode those codes with efficient algorithms. This work falls in the combinatorial facet of list decoding. Combinatorially, an error-correcting code has "nice" list-decodability properties if every Hamming ball of "large" radius has a "small" number of codewords in it. In this work, we are interested in exposing some combinatorial *limitations* on the performance of list-decodable codes. Specifically, we seek lower bounds on the list size needed to perform decoding up to a certain number of errors, or in other words, lower bounds on the number of codewords that must fall inside *some* ball of specified radius centered at some point. We show such a result by picking the center in a certain probabilistic way. We now give some background definitions and terminology, followed by a description of our main result.

1.1 Preliminaries

We denote the set $\{1, 2, \ldots, m\}$ by the shorthand $[m]$. For $q \geq 2$, a q-ary *code* of *block length* n is simply a subset of $[q]^n$. The elements of the code are referred to as *codewords*. The high-level property of a code that makes it useful for error-correction is its sparsity — the codewords must be well spread-out, so they are unlikely to distort into one another. One way to insist on sparsity is that the Hamming distance between every pair of distinct codewords is at least d. Note that this is equivalent to requiring that every Hamming ball of radius $\lfloor (d-1)/2 \rfloor$ has at most one codeword. Generalizing this, one can allow up to a small number, say L, of codewords in Hamming balls of certain radius. This leads to the notion of list decoding and a good list-decodable code. Since the expected Hamming distance of a random string of length n from any codeword is $(1 - 1/q) \cdot n$ for a q-ary code, the largest fraction of errors one can sensibly hope to correct is $(1 - 1/q)$. This motivates the following definition of a list-decodable code.

Definition 1 *Let $q \geq 2$, $0 < \rho < 1$, and L be a positive integer. A q-ary code C of block length n is said to be* (ρ, L)-list-decodable *if for every $y \in [q]^n$, the Hamming ball of radius $\rho \cdot (1 - 1/q) \cdot n$ centered at y contains at most L codewords of C.*

We will study (ρ, L)-list-decodable codes for $\rho = 1 - \varepsilon$ in the limit of $\varepsilon \to 0$. This setting is the one where list decoding is most beneficial, and is a clean setting to initially study the asymptotics. In particular, we will prove that, except for trivial codes whose size does not grow with n, $(1-\varepsilon, L)$-list-decodable codes require list size $L = \Omega(1/\varepsilon^2)$ (hiding dependence on q).

1.2 Context and Related Results

Before stating our result, we describe some of the previously known results to elucidate the broader context where our work fits. The *rate* of a q-ary code of block length n is defined to be $\frac{\log_q |C|}{n}$. For $0 \leq x \leq 1$, we denote by $H_q(x)$ the q-ary entropy function, $H_q(x) = x \log_q(q-1) - x \log_q x - (1 - x) \log_q(1 - x)$.

Using the probabilistic method, it can be shown that (ρ, L)-list-decodable q-ary codes of rate $1 - H_q((1 - 1/q)\rho) - 1/L$ exist [Eli2, GHSZ]. In particular, in the limit of large L, we can achieve a rate of $1 - H_q((1-1/q)\rho)$, which equals both the Hamming bound and the Shannon capacity of the q-ary channel that changes a symbol $\alpha \in [q]$ to a uniformly random element of $[q]$ with probability ρ and leaves α unchanged with probability $1 - (1 - 1/q)\rho$. When $\rho = 1 - \varepsilon$ for small ε, we have $H_q((1 - 1/q)\rho) = 1 - \Omega(q\varepsilon^2/\log q)$. Therefore, there exist $(1 - \varepsilon, L(q, \varepsilon))$-list-decodable q-ary codes with $2^{\Omega(q\varepsilon^2 n)}$ codewords and $L(q,\varepsilon) = O(\frac{\log q}{q\varepsilon^2})$. In particular, for constant q, list size of $O(1/\varepsilon^2)$ suffices for non-trivial list decoding up to radius $(1-1/q)\cdot(1-\varepsilon)$.

We are interested in whether this quadratic dependence on $1/\varepsilon$ in the list size is inherent. The quadratic bound is related to the $2\log(1/\varepsilon) - O(1)$ lower bound due to [RT] for the amount of "entropy loss" in *randomness extractors*, which are well-studied objects in the subject of pseudorandomness. In fact a lower bound of $\Omega(1/\varepsilon^2)$ on list size will implies such an entropy loss bound for ("strong") randomness extractors. However, in the other direction, the argument loses a factor of ε in the lower bound, yielding only a lower of $\Omega(1/\varepsilon)$ for list size (cf. [Vad]).

For the model of erasures, where up to a fraction $(1-\varepsilon)$ of symbols are erased by the channel, optimal bounds of $\Theta(\log(1/\varepsilon))$ are known for the list size required for binary codes [Gur]. This can be compared with the $\log\log(1/\varepsilon) - O(1)$ lower bound on entropy loss for dispersers [RT].

A lower bound of $\Omega(1/\varepsilon^2)$ for list size L for $(1-\varepsilon, L)$-list-decodable *binary* codes follows from the work of Blinovsky [Bli]. We discuss more about his work and how it compares to our results in Section 1.5.

1.3 Our Result

Our main result is a proof of the following fact: the smallest list size that permits list decoding up to radius $(1-1/q)(1-\varepsilon)$ is $\Theta(\varepsilon^{-2})$ (hiding constants depending on q in the Θ-notation). The formal statement of our main result is below.

Theorem 2 (Main) *For every integer $q \geq 2$ there exists $c_q > 0$ and $d_q < \infty$ such that for all small enough $\varepsilon > 0$, the following holds. If C is a q-ary $(1-\varepsilon, c_q/\varepsilon^2)$-list-decodable code, then $|C| \leq 2^{d_q \cdot \varepsilon^{-2}\log(1/\varepsilon)}$.*

1.4 Overview of Proof

We now describe the high-level structure of our proof. Recall that our goal is to exhibit a center z that has several (specifically $\Omega(1/\varepsilon^2)$) codewords of C with large correlation, where we say two codewords have correlation ε if they agree in $(1/q+\varepsilon)\cdot n$ locations. (The actual definition we use, given in Definition 3, is slightly different, but this version suffices for the present discussion.) Using the probabilistic method, it is not very difficult to prove the existence of such a center z and $\Omega(1/\varepsilon^2)$ codewords whose *average* correlation with z is at least $\Omega(\varepsilon)$. (This is the content of our Lemma 6.) This step is closely related to (and actually follows from) the known lower bound of Radhakrishnan and Ta-Shma [RT] on the "entropy loss" of "randomness extractors," by applying the known connection between randomness extractors and list-decodable error-correcting codes (see [Tre, TZ, Vad]).

However, this large average could occur due to about $1/\varepsilon$ codewords having a $\Omega(1)$ correlation with z, whereas we would like to find many more (i.e., $\Omega(1/\varepsilon^2)$) codewords with smaller (i.e., $\Omega(\varepsilon)$) correlation. We get around this difficulty by working with a large subcode C' of C where such a phenomenon cannot occur. Roughly speaking, we will use the probabilistic method to prove the existence of a large "L-pseudorandom" subcode C', for which looking at any set of L codewords of C *never* reveals any significant overall bias in terms of the most popular symbol (out of $[q]$). More formally, all ℓ-tuples, $\ell \leq L$, the average "plurality" (i.e., frequency of most frequent symbol) over all the coordinates isn't much higher than ℓ/q. (This is the content of our Lemma 7.) This in turn implies that for every center z, the sum of the correlations of z with all codewords that have "large" correlation (say at least $D\varepsilon$, for a sufficiently large constant D) is small. Together with the high average correlation bound, this means several codewords must have "intermediate" correlation with z (between ε and $D\varepsilon$). The number of such codewords is our lower bound on list size.

1.5 Comparison with Blinovsky [Bli]

As remarked earlier, a lower bound of $L = \Omega(1/\varepsilon^2)$ for *binary* $(1-\varepsilon, L)$-list-decodable codes follows from the work of Blinovsky [Bli]. He explores the tradeoff between ρ, L, and the relative rate γ of a (ρ, L)-list-decodable code, when all three of these parameters are constants and the block length n tends to infinity. A special case of his main theorem shows that if $\rho = 1 - \varepsilon$ and $L \le c/\varepsilon^2$ for a certain constant $c > 0$, then the rate γ must be zero asymptotically, which means that the code can have at most $2^{o(n)}$ codewords for block length n. A careful inspection of his proof, however, reveals an $f(\varepsilon)$ bound (independent of n) on the number of codewords in any such code. This is similar in spirit to our Theorem 2. However, our work compares favorably with [Bli] in the following respects.

1. Our result also holds for q-ary codes for $q > 2$. The result in [Bli] applies only to binary codes, and it is unclear whether his analysis can be generalized to q-ary codes.

2. Our result is quantitatively stronger. The dependence $f(\varepsilon)$ of the bound on the size of the code in [Bli] is much worse than the $(1/\varepsilon)^{O(\varepsilon^{-2})}$ that we obtain. In particular, $f(\varepsilon)$ is at least an exponential tower of height $\Theta(1/\varepsilon^2)$ (and is in fact bigger than the Ackermann function of $1/\varepsilon$).

3. Our proof seems significantly simpler and provides more intuition about why and how the lower bound arises.

We now comment on the proof method in [Bli]. As with our proof, the first step in the proof is a bound for the case when the average correlation (w.r.t every center) for every set of $L+1$ codewords is small (this is Theorem 2 in [Bli]). Note that this is a more stringent condition than requiring no set of $L+1$ codewords lie within a small ball. Our proof uses the probabilistic method to show the existence of codewords with large average correlation in any reasonable sized code. The proof in [Bli] is more combinatorial, and uses a counting argument to bound the size of the code when all subsets of $L+1$ codewords have low average correlation (with every center). But the underlying technical goal of the first step in both the approaches is the same.

The second step in Blinovsky's proof is to use this bound to obtain a bound for list-decodable codes. The high-level idea is to pick a subcode of the list-decodable code with certain nice properties so that the bound for average correlation can be turned into one for list decoding. This is also similar in spirit to our approach (Lemma 7). The specifics of how this is done are, however, quite different. The approach in [Bli] is to find a large subcode which is $(L+1)$-*equidistant*, i.e., for every $k \le L+1$, all subsets of k codewords have the same value for their k'th order scalar product, which is defined as the sum over all coordinates of the product of the k symbols (from $\{0,1\}$) in that coordinate.[1] Such a subcode has the following useful property: in each subset of $L+1$ codewords, all codewords in the subset have the same agreement with the best center, i.e., the center obtained by taking their coordinate-wise majority, and moreover this value is independent of the choice of

[1] A slight relaxation of the $(L+1)$-equidistance property is actually what is used in [Bli], but this description should suffice for the discussion here.

the subset of $L+1$ codewords. This in turn enables one to get a bound for list decoding from one for average correlation. The requirement of being $(L+1)$-equidistant is a rather stringent one, and is achieved iteratively by ensuring k-equidistance for $k = 1, 2, \ldots, L+1$ successively. Each stage incurs a rather huge loss in the size of the code, and thus the bound obtained on the size of the original code is an enormously large function of $1/\varepsilon$. We make do with a much weaker property than $(L+1)$-equidistance, letting us pick a much larger subcode with the property we need. This translates into a good upper bound on the size of the original list-decodable code.

2 Proof of main result

We first begin with convenient measures of closeness between strings, the agreement and the correlation.

Definition 3 (Agreement and Correlation) *For strings $x, y \in [q]^n$, define their* agreement*, denoted* $\mathsf{agr}(x,y) = \frac{1}{n} \cdot \#\{i : x_i = y_i\}$*. Their* correlation *is the value* $\mathsf{corr}(x,y) \in [-1/(q-1), 1]$ *such that* $\mathsf{agr}(x,y) = \frac{1}{q} + \left(1 - \frac{1}{q}\right) \cdot \mathsf{corr}(x,y)$.[2]

The standard notion of correlation between two strings in $\{1,-1\}^n$ is simply their dot product divided by n; the definition above is a natural generalization to larger alphabets.

A very useful notion for us will be the plurality of a set of codewords.

Definition 4 (Plurality) *For symbols $a_1, \ldots, a_k \in [q]$, we define their* plurality $\mathsf{plur}(a_1, \ldots, a_k) \in [q]$ *to be the most frequent symbol among $a_1, \ldots, a_k$, breaking ties arbitrarily. We define the* plurality count $\#\mathsf{plur}(a_1, \ldots, a_k) \in \mathbb{N}$ *to be the number of times that* $\mathsf{plur}(a_1, \ldots, a_k)$ *occurs among $a_1, \ldots, a_k$.*

For vectors $c_1, \ldots, c_k \in [q]^n$, we define $\mathsf{plur}(c_1, \ldots, c_k) \in [q]^n$ *to be the componentwise plurality, i.e.* $\mathsf{plur}(c_1, \ldots, c_k)_i = \mathsf{plur}(c_{1i}, \ldots, c_{ki})$.

We define $\#\mathsf{plur}(c_1, \ldots, c_k)$ *to be the average plurality count over all coordinates; that is,* $\#\mathsf{plur}(c_1, \ldots, c_k) = (1/n) \cdot \sum_{i=1}^{n} \#\mathsf{plur}(c_{1i}, \ldots, c_{ki})$.

The reason pluralities will be useful to us is that they capture the maximum average correlation any vector has with a set of codewords:

Lemma 5 *For all $c_1, \ldots, c_k \in [q]^n$,*

$$\begin{aligned}
\max_{z \in [q]^n} \sum_{i=1}^{k} \mathsf{agr}(z, c_i) &= \sum_{i=1}^{k} \mathsf{agr}(\mathsf{plur}(c_1, \ldots, c_k), c_i) \\
&= \#\mathsf{plur}(c_1, \ldots, c_k)
\end{aligned}$$

Note that our goal of proving lower bound on list size is the same as proving that in every not too small code, there must be some center z that has several (i.e. $\Omega(1/\varepsilon^2)$) close-by codewords, or in other words several codewords with large (i.e., at least ε) correlation. We begin by showing the existence of a center which has a large *average* correlation with a collection of several codewords. By Lemma 5, this is equivalent to finding a collection of several codewords whose total plurality count is large.

[2] Note that we find it convenient to work with agreement and correlation that are normalized by dividing by the length n.

Lemma 6 *For all integers $q \geq 2$ and $t \geq 37q$, there exists a constant $b_q > 0$ such that for every positive integer t and every code $C \subseteq [q]^n$ with $|C| \geq 2t$, there exist t distinct codewords $c_1, c_2, \dots, c_t \in C$ such that*

$$\#\mathsf{plur}(c_1, \dots, c_t) \geq \frac{t}{q} + \Omega\left(\sqrt{\frac{t}{q}}\right).$$

Equivalently, there exists a $z \in [q]^n$ such that

$$\sum_{i=1}^{t} \mathsf{corr}(z, c_i) \geq \Omega\left(\sqrt{\frac{t}{q}}\right) \quad . \tag{1}$$

Proof: Without loss of generality, assume $|C| = 2t$. Pick a subset $\{c_1, c_2, \dots, c_t\}$ from C, chosen uniformly at random among all t-element subsets of C. For $j = 1, \dots, n$, define the random variable $P_j = \#\mathsf{plur}(c_{1j}, \dots, c_{tj})$ to be the plurality of the j'th coordinates. By definition, $\#\mathsf{plur}(c_1, \dots, c_t) = (1/n) \cdot \sum_{j=1}^{n} P_j$. Notice that P_j is always at least t/q, and we would expect the plurality to occasionally deviate from the lower bound. Indeed, Lemma 15 shows that for any sequence of $2t$ elements of $[q]$, if we choose a random subset of half of them, the expected plurality count is $t/q + \Omega(\sqrt{t/q})$. Thus, $\mathbf{E}[P_j] = t/q + \Omega(\sqrt{t/q})$. So $\mathbf{E}[(1/n) \cdot \sum_j P_j] = t/q + \Omega(\sqrt{t/q})$, and thus the lemma follows by taking any $c_1, \dots, c_t$ that achieves the expectation. The equivalent reformulation in terms of correlation follows from Lemma 5 and the definition of correlation in terms of agreement (Definition 3). ∎

For any $\varepsilon > 0$, the above lemma gives a center z and $t = \Omega(1/q\varepsilon^2)$ codewords $c_1, \dots, c_t$ such that the average correlation between z and $c_1, \dots, c_t$ is at least ε. This implies that at least $(\varepsilon/2) \cdot t$ of the c_i's have correlation at least $\varepsilon/2$ with z. Thus we get a list-size lower bound of $(\varepsilon/2) \cdot t = \Omega(1/q\varepsilon)$ for decoding from correlation $\varepsilon/2$. (This argument has also appeared in [LTW].)

Now we would like to avoid the ε factor loss in list size in the above argument. The reason it occurs is that the average correlation can be ε due to the presence of $\approx \varepsilon t$ of the c_i's having extremely high correlation with z. This is consistent with the code being list-decodable with list size $o(1/(q\varepsilon^2))$ for correlation ε, but it means that it code has very poor list-decoding properties at some higher correlations — e.g., having $\varepsilon t = \Omega(1/(q\varepsilon))$ codewords at correlation $\Omega(1)$, whereas we'd expect a "good" code to have only $O(1)$ such codewords. In our next (and main) lemma, we show that we can pick a subcode of the code where this difficulty does not occur. Specifically, if C has good list-decoding properties at correlation ε, we get a subcode that has good list-decoding properties at every correlation larger than ε.

Lemma 7 (Main technical lemma) *For all positive integers L, t, $m \geq 2t$ and $q \geq 2$, and all small enough $\varepsilon > 0$, the following holds. Let C be a $(1 - \varepsilon, L)$-list-decodable q-ary code of block length n with $|C| > 2L \cdot t \cdot m!/(m-t)!$. Then there exists a subcode $C' \subseteq C$, $|C'| \geq m$, such that for all positive integers $\ell \leq t$ and every $c_1, c_2, \dots, c_\ell \in C'$,*

$$\#\mathsf{plur}(c_1, \dots, c_\ell) \leq \left(\frac{1}{q} + \left(1 - \frac{1}{q}\right) \cdot \varepsilon + O\left(\frac{q^{3/2}}{\sqrt{\ell}}\right)\right) \cdot \ell.$$

Equivalently, for every $z \in [q]^n$ and every $c_1, \dots, c_\ell \in C'$, we have

$$\sum_{i=1}^{\ell} \mathsf{corr}(z, c_i) \leq \left(\varepsilon + O\left(\frac{q^{3/2}}{\sqrt{\ell}}\right)\right) \cdot \ell \ . \tag{2}$$

Notice that the lemma implies a better upper bound on list size for correlations much larger than ε. More precisely, for every $\delta > 0$, it implies that the number of codewords having correlation at least $\varepsilon + \delta$ with a center z is at most $\ell = O(q^3/\delta^2)$. In fact, any ℓ codewords must even have *average* correlation at most $\varepsilon + \delta$.

Proof: We will pick a subcode $C' \subseteq C$ of size m at random from all m-element subsets of C, and prove that C' will fail to have the claimed property with probability less than 1.

For now, however, think of the code C' as being fixed, and we will reduce proving the desired properties above to bounding some simpler quantities. Let $(c_1, c_2, \ldots, c_\ell)$ be an arbitrary ℓ-tuple of codewords in C'. We will keep track of the average plurality count $\#\mathsf{plur}(c_1, \ldots, c_i)$ as we add each codeword to this sequence. To describe how this quantity can change at each step, we need a couple of additional definitions. We say a sequence $(a_1, \ldots, a_i) \in [q]^i$ has a *plurality tie* if at least two symbols occur $\#\mathsf{plur}(a_1, \ldots, a_i)$ times among $a_1, \ldots, a_i$. For vectors $c_1, \ldots, c_i \in [q]^n$, we define $\#\mathsf{ties}(c_1, \ldots, c_i)$ to be the fraction of coordinates $j \in [n]$ such that $(c_{1j}, \ldots, c_{ij})$ has a plurality tie. Then:

Claim 8 *For every* $c_1, \ldots, c_i \in [q]^n$, $\#\mathsf{plur}(c_1, \ldots, c_i) \leq \#\mathsf{plur}(c_1, \ldots, c_{i-1}) + \mathsf{agr}(c_i, \mathsf{plur}(c_1, \ldots, c_{i-1})) + \#\mathsf{ties}(c_1, \ldots, c_{i-1})$.

Proof of claim: Consider each coordinate $j \in [n]$ separately. Clearly,

$$\#\mathsf{plur}(c_{1j}, \ldots, c_{ij}) \leq \#\mathsf{plur}(c_{1j}, \ldots, c_{(i-1)j}) + 1 \ .$$

Moreover, if $(c_{1j}, \ldots, c_{(i-1)j})$ does not have a plurality tie, then the plurality increases iff c_{ij} equals the unique symbol $\mathsf{plur}(c_{1j}, \ldots, c_{(i-1)j})$ achieving the plurality. Thus,

$$\#\mathsf{plur}(c_{1j}, \ldots, c_{ij}) \leq \#\mathsf{plur}(c_{1j}, \ldots, c_{(i-1)j}) + A_j + T_j,$$

where T_j is the indicator variable for $(c_{1j}, \ldots, c_{(i-1)j})$ having a plurality tie, and A_j for c_{ij} agreeing with $\mathsf{plur}(c_{1j}, \ldots, c_{(i-1)j})$. The claim follows by averaging over $j = 1, \ldots, n$.
$\square$

Thus, our task of bounding $\#\mathsf{plur}(c_1, \ldots, c_\ell)$ reduces to bounding $\mathsf{agr}(c_i, \mathsf{plur}(c_1, \ldots, c_{i-1}))$ and $\#\mathsf{ties}(c_1, \ldots, c_{i-1})$ for each $i = 1, \ldots, \ell$. The first term we bound using the list-decodability of C and the random choice of the subcode C'.

Claim 9 *There exists a choice of the subcode C' such that $|C'| = m$ and for every $i \leq t$ and every (ordered) sequence $c_1, \ldots, c_i \in C'$, we have*

$$\mathsf{agr}(c_i, \mathsf{plur}(c_1, \ldots, c_{i-1})) \leq 1/q + (1 - 1/q) \cdot \varepsilon \ .$$

Proof of claim: We choose the subcode C' uniformly at random from all m-subsets of C. We view C' as a sequence of m codewords selected randomly from C without replacement. Consider any i of the codewords $c_1, \ldots, c_i$ in this sequence. By the list-decodability of the C, for any $c_1, \ldots, c_{i-1}$, there are at most L choices for c_i having agreement larger than $(1/q + (1 - 1/q) \cdot \varepsilon)$ with $\mathsf{plur}(c_1, \ldots, c_{i-1})$. Conditioned on $c_1, \ldots, c_{i-1}$, c_i is distributed uniformly on the remaining $|C| - i + 1$ elements of C, so the probability of c_i being one of the $\leq L$ bad codewords is at most $L/(|C| - i + 1)$.

By a union bound, the probability that the claim fails for at least one subsequence $c_1, \ldots, c_i$ of at most t codewords in C' is at most

$$\sum_{i=1}^{t} \frac{m!}{(m-i)!} \cdot \frac{L}{|C| - i + 1} \leq t \cdot \frac{m!}{(m-t)!} \cdot \frac{L}{|C| - t + 1} < 1.$$

Thus, there exists a choice of subcode C' satisfying the claim. □

For the $\#\mathsf{ties}(c_1, \ldots, c_{i-1})$ terms, we consider the codewords $c_1, \ldots, c_\ell$ in a random order.

Claim 10 *For every sequence of $c_1, \ldots, c_\ell \in [q]^n$, there exists a permutation $\sigma : [\ell] \to [\ell]$ such that*

$$\sum_{i=1}^{\ell} \#\mathsf{ties}(c_{\sigma(1)}, \ldots, c_{\sigma(i)}) = O(q^{3/2} \cdot \sqrt{\ell}).$$

Proof of claim: We choose σ uniformly at random from all permutations $\sigma : [\ell] \to [\ell]$ and show that the expectation of the left side is at most $O(q^{3/2} \cdot \sqrt{\ell})$. By linearity of expectations, it suffices to consider the expected number of plurality ties occurring in each coordinate $j \in [n]$. That is, we read the symbols $c_{1j}, \ldots, c_{\ell j} \in [q]$ in a random order σ and count the number of prefixes $c_{\sigma(1)j}, \ldots, c_{\sigma(i)j}$ having a plurality tie. If this prefix were i symbols chosen independently according to some (arbitrary) distribution, then it is fairly easy to show that the probability of a tie is $O(1/\sqrt{i})$ (ignoring the dependence on q), and summing this from $i = 1, \ldots, \ell$ gives $O(\sqrt{\ell})$ expected ties in each coordinate. Since they are not independently chosen, but rather i distinct symbols from a fixed sequence of ℓ symbols, the analysis becomes a bit more involved, but nevertheless the bound remains essentially the same. Specifically, in Lemma 17, the expected number of ties is shown to be $O(q^{3/2} \cdot \sqrt{\ell})$, yielding the claim. □

Now to complete the proof of Lemma 7, let C' be as in Claim 9, and let $c_1, \ldots, c_\ell$ be an arbitrary sequence of distinct codewords in C'. Let σ be permutation guaranteed by Claim 10. Then, by Claim 8, we have

$$\begin{aligned}
\#\mathsf{plur}(c_1, \ldots, c_\ell) &= \#\mathsf{plur}(c_{\sigma(1)}, \ldots, c_{\sigma(\ell)}) \\
&\leq \sum_{i=1}^{\ell} \left[\mathsf{agr}(c_{\sigma(i)}, \mathsf{plur}(c_{\sigma(1)}, \ldots, c_{\sigma(i-1)})) + \#\mathsf{ties}(c_{\sigma(1)}, \ldots, c_{\sigma(i-1)})\right] \\
&\leq \ell \cdot (1/q + (1 - 1/q) \cdot \varepsilon) + O(q^{3/2} \cdot \sqrt{\ell}),
\end{aligned}$$

as desired. The equivalent reformulation in terms of correlation again follows from Lemma 5 and the definition of correlation in terms of agreement (Definition 3). ■

The following corollary of Lemma 7 will be useful in proving our main result.

Corollary 11 *Let L, t, m, q, ε, C, and C' be as in Lemma 7 for a choice of parameters satisfying $t \geq L$. Then for all $z \in [q]^n$ and all $D \geq 2$,*

$$\sum_{\substack{c \in C' \\ \mathsf{corr}(z,c) \geq D\varepsilon}} \mathsf{corr}(z, c) \leq O\left(\frac{q^3}{D\varepsilon}\right) . \tag{3}$$

Proof: Let $c_1, c_2, \ldots, c_r$ be all the codewords of C' that satisfy $\mathsf{corr}(z,c) \geq D\varepsilon$. Since C, and hence C', is $(1-\varepsilon, L)$-list-decodable, we have $r \leq L \leq t$. Using (2) for the codewords $c_1, c_2, \ldots, c_r$, we have

$$D\varepsilon \leq \frac{1}{r}\sum_{i=1}^{r} \mathsf{corr}(z, c_i) \leq \varepsilon + O\Big(\frac{q^{3/2}}{\sqrt{r}}\Big)$$

which gives $r = O(q^3/((D-1)^2\varepsilon^2)) = O(q^3/(D^2\varepsilon^2))$, since $D \geq 2$. Applying (2) again,

$$\begin{aligned}
\sum_{\substack{c \in C' \\ \mathsf{corr}(z,c) \geq D\varepsilon}} \mathsf{corr}(z,c) &= \sum_{i=1}^{r} \mathsf{corr}(z, c_i) \\
&\leq \varepsilon r + O(q^{3/2}\sqrt{r}) \\
&\leq O\Big(\frac{q^3}{D^2\varepsilon}\Big) + O\Big(\frac{q^3}{D\varepsilon}\Big) \\
&\leq O\Big(\frac{q^3}{D\varepsilon}\Big) .
\end{aligned}$$

■

We are now ready to prove our main result, Theorem 2, which restate (in slightly different form) below.

Theorem 12 (Main) *There exist constants $c > 0$, $d < \infty$, such that for all small enough $\varepsilon > 0$, the following holds. Suppose C is a q-ary $(1-\varepsilon, L)$-list-decodable code with $|C| > 1/(q\varepsilon^2)^{d/(q\varepsilon^2)}$ Then $L \geq c/(q^5\varepsilon^2)$.*

Proof: Let T be a large enough constant to be specified later. Let $t = \lfloor \frac{1}{Tq\varepsilon^2} \rfloor$. If $L > t$, then there is nothing to prove. So assume that $t \geq L \geq 1$ and set $m = 2t$. Then

$$2L \cdot t \cdot \frac{m!}{(m-t)!} \leq 2t^2 \cdot (2t)^t = \left(\frac{1}{q\varepsilon^2}\right)^{O(1/(q\varepsilon^2))} < |C|,$$

for a sufficiently large choice of the constant d. Let C' be a subcode of C of size $m = 2t$ guaranteed by Lemma 7.

By Lemma 6, there exist t codewords c_i, $1 \leq i \leq t$, in C', and a center $z \in [q]^n$ such that

$$\sum_{i=1}^{t} \mathsf{corr}(z, c_i) = \Omega\left(\sqrt{\frac{t}{q}}\right) \quad . \tag{4}$$

Also, for any $D \geq 2$, we have $\sum_{i=1}^{t} \mathsf{corr}(z, c_i)$ equals

$$\sum_{i:\mathsf{corr}(z,c_i) < \varepsilon} \mathsf{corr}(z, c_i) + \sum_{i:\varepsilon \leq \mathsf{corr}(z,c_i) < D\varepsilon} \mathsf{corr}(z, c_i) + \sum_{i:\mathsf{corr}(z,c_i) \geq D\varepsilon} \mathsf{corr}(z, c_i)$$

$$\leq \varepsilon t + D\varepsilon L + O\Big(\frac{q^3}{D\varepsilon}\Big) \tag{5}$$

where to bound the second part we used that C' is $(1-\varepsilon, L)$-list-decodable, and to bound the third part we used the fact that C' satisfies (3).

Putting these together, and setting $D = q^4 \cdot T$, we have

$$\begin{aligned} \Omega\left(\sqrt{\frac{t}{q}}\right) &\leq \varepsilon t + D\varepsilon L + O\Big(\frac{q^3}{D\varepsilon}\Big) \\ &\leq \sqrt{\frac{t}{Tq}} + D\varepsilon L + O\Big(\sqrt{\frac{t}{Tq}}\Big). \end{aligned}$$

For a sufficiently large choice of the constant T, this gives

$$L \geq \frac{1}{D\varepsilon} \cdot \Omega\left(\sqrt{\frac{t}{q}}\right) = \Omega\left(\frac{1}{\sqrt{T} \cdot q^5 \cdot \varepsilon^2}\right).$$

as desired. ■

3 Open questions

Several questions are open in the general direction of exhibiting limitations on the performance of list-decodable codes. We mention some of them below.

- We have not attempted to optimize the dependence on the alphabet size q in our bound on list size (i.e. the constant c_q in Theorem 2), and this leaves a gap between the upper and lower bounds. The probabilistic code construction of [Eli2, GHSZ] achieves a nearly linear dependence on q (specifically, list size $L = O(\log q/(q\varepsilon^2))$, whereas our lower bound (Theorem 12) has a polynomial dependence on q (namely, it shows $L = \Omega(1/(q^5\varepsilon^2))$.

- It should be possible to use our main result, together with an appropriate "filtering" argument (that focuses, for example, on a subcode consisting of all codewords of a particular Hamming weight) to obtain upper bounds on rate of list-decodable q-ary codes. In particular, can one confirm that for each fixed L, the maximum rate achievable for list decoding up to radius p with list size L is strictly less than the capacity $1 - H_q(p)$? Such a result is known for binary codes [Bli]. Also, it is an interesting question whether some of the ideas in this paper can be used to improve the rate upper bounds of Blinovsky [Bli] for the binary case.

- Can one prove a lower bound on list size as a function of distance from "capacity"? In particular, does one need list size $\Omega(1/\gamma)$ to achieve a rate that is within γ of capacity?

- Can one prove stronger results for linear codes?

Acknowledgments

We thank Michael Mitzenmacher, Madhu Sudan, and Amnon Ta-Shma for helpful conversations.

References

[Bli] V. M. Blinovsky. Bounds for codes in the case of list decoding of finite volume. *Problems of Information Transmission*, 22(1):7–19, 1986.

[DS] D. Dubhashi and S. Sen. Concentration of Measure for Randomized Algorithms: Techniques and Analysis. In S. R. et al., editor, *Handbook of Randomized Computing, Vol. I.*, chapter 3, pages 35–100. Kluwer, 2001.

[Eli1] P. Elias. List decoding for noisy channels. *Technical Report 335, Research Laboratory of Electronics, MIT*, 1957.

[Eli2] P. Elias. Error-correcting codes for list decoding. *IEEE Transactions on Information Theory*, 37:5–12, 1991.

[Gur] V. Guruswami. List decoding from erasures: Bounds and code constructions. *IEEE Transactions on Information Theory*, 49(11):2826–2833, 2003.

[GHSZ] V. Guruswami, J. Hastad, M. Sudan, and D. Zuckerman. Combinatorial bounds for list decoding. *IEEE Transactions on Information Theory*, 48(5):1021–1035, 2002.

[GV] V. Guruswami and S. Vadhan. A Lower Bound on List Size for List Decoding. In *Proceedings of the 8th International Workshop on Randomization and Computation (RANDOM '05)*, number 3624 in Lecture Notes in Computer Science, Berkeley, CA, August 2005. Springer. To appear.

[LTW] C.-J. Lu, S.-C. Tsai, and H.-L. Wu. On the complexity of hardness amplification. In *Proceedings of the 20th Annual IEEE Conference on Computational Complexity*, San Jose, CA, June 2005. To appear.

[RT] J. Radhakrishnan and A. Ta-Shma. Bounds for dispersers, extractors, and depth-two superconcentrators. *SIAM Journal on Discrete Mathematics*, 13(1):2–24 (electronic), 2000.

[TZ] A. Ta-Shma and D. Zuckerman. Extractor codes. *IEEE Transactions on Information Theory*, 50(12):3015–3025, 2004.

[Tre] L. Trevisan. Extractors and Pseudorandom Generators. *Journal of the ACM*, 48(4):860–879, July 2001.

[Vad] S. P. Vadhan. Randomness Extractors and their Many Guises. Tutorial at IEEE Symposium on Foundations of Computer Science, November 2002. Slides available at `http://eecs.harvard.edu/~salil`.

[Woz] J. M. Wozencraft. List Decoding. *Quarterly Progress Report, Research Laboratory of Electronics, MIT*, 48:90–95, 1958.

A Appendix

Lemma 13 *For all integers n, k such that $0 < k < n$, we have*

$$\binom{n}{k} = \Theta\left(\sqrt{\frac{n}{k \cdot (n-k)}} \cdot 2^{H(k/n) \cdot n}\right).$$

Proof: Use Stirling's approximation for the factorials. ∎

Lemma 14 *For all positive integers $t, s \le t/2$,*

$$\sum_{i=\lfloor\sqrt{s}/2\rfloor}^{\lfloor\sqrt{s}\rfloor} \frac{\binom{t}{s+i}\binom{t}{s-i}}{\binom{2t}{2s}} = \Omega(1)$$

Proof: Without loss of generality, we may assume that $s \le t/2$ (otherwise replace s with $t-s$). Let $A_i = \binom{t}{s+i}\binom{t}{s-i}/\binom{2t}{2s}$. Using Lemma 13, we see that

$$\begin{aligned} A_0 &= \Theta\left(\sqrt{\frac{t}{s\cdot(t-s)}}\right) \\ &= \Omega\left(\frac{1}{\sqrt{s}}\right) \end{aligned}$$

For $0 < i \le \lfloor\sqrt{s}\rfloor$, we have

$$\begin{aligned} A_i &= \frac{(t-s-i+1)\cdot(s-i+1)}{(s+i)\cdot(t-s+i)} \cdot A_{i-1} \\ &= \left(1 - \frac{2i-1}{t-s+i}\right)\cdot\left(1-\frac{2i-1}{s+i}\right)\cdot A_{i-1} \\ &\ge \left(1-\frac{2}{\sqrt{s}}\right)^2 \cdot A_{i-1} \\ &\ge \left(1-\frac{2}{\sqrt{s}}\right)^{2i} A_0 \\ &= \Omega(A_0) \end{aligned}$$

Therefore,

$$\sum_{i=\lfloor\sqrt{s}/2\rfloor}^{\lfloor\sqrt{s}\rfloor} A_i = (\lfloor\sqrt{s}\rfloor - \lfloor\sqrt{s}/2\rfloor)\cdot\Omega(A_0) = \Omega(1).$$

■

Lemma 15 *Let t, q be integers such that $q \ge 2$ and $t \ge 37q$. Let $a_1, \ldots, a_{2t} \in [q]$, and let T be chosen uniformly at random from all subsets of $[2t]$ of size t. Then*

$$\mathop{\mathbf{E}}_{T}[\#\mathsf{plur}(a_j : j \in T)] = \frac{t}{q} + \Omega\Big(\sqrt{\frac{t}{q}}\Big).$$

Proof: Notice that $\#\mathsf{plur}(a_j : j \in T)$ is always at least t/q. Thus it suffices to show that with constant probability over the choice of T, there exists an $\alpha \in [q]$ such that $\#\{i \in T : a_i = \alpha\} \ge t/q + \Omega(\sqrt{t/q})$. In fact, we restrict our attention to a single value of α, namely the most frequent symbol among $a_1, \ldots, a_{2t}$. Then setting $s = \lfloor t/q\rfloor$, α occurs at least $2s$ times among $a_1, \ldots, a_{2t}$, so let $S \subseteq [2t]$ be any set of $2s$ indices j such that $a_j = \alpha$. Then,

$$\Pr_T[\#\{j \in T : a_j = \alpha\} = s+i] \ge \Pr_T[|T \cap S| = s+i] = \frac{\binom{t}{s+i}\binom{t}{s-i}}{\binom{2t}{2s}}.$$

(To see the last equation, note that $|T \cap S|$ has the same distribution whether T is a random and S is fixed, or T is fixed and S is random.) Thus by Lemma 14, with probability $\Omega(1)$ over T, we have:

$$\begin{aligned}\#\{j \in T \mid a_j = \alpha\} &\geq s + \lfloor \sqrt{s}/2 \rfloor \\ &\geq \frac{t}{q} + \frac{\sqrt{t/q}}{2} - 3 \\ &\geq \frac{t}{q} + \Omega\left(\sqrt{\frac{t}{q}}\right),\end{aligned}$$

where in the last inequality we use the fact that $t \geq 37q$. ■

Lemma 16 *For integers* $0 < i < a$, $0 < j < b$,

$$\frac{\binom{a}{i}\binom{b}{j}}{\binom{a+b}{i+j}} \leq O\left(\sqrt{\frac{a \cdot b \cdot (i+j) \cdot (a+b-i-j)}{i \cdot (a-i) \cdot j \cdot (b-j) \cdot (a+b)}}\right).$$

Proof: Applying Lemma 13 to each of the binomial coefficients yields the bound above times 2^t, where

$$t = H(i/a) \cdot a + H(j/b) \cdot b - H((i+j)/(a+b)) \cdot (a+b) \leq 0,$$

where the last inequality follows by concavity of the entropy function. ■

Lemma 17 *Let* $b_1, b_2, \ldots, b_k$ *be a sequence of elements from the universe* $[q]$*. Recall that a prefix of such a sequence has a plurality tie if there are at least two elements of* $[q]$ *that occur the same number of times in the prefix, and no other element occurs a strictly greater number of times in the prefix. Let* Y *be the random variable counting the number of prefixes with a plurality tie in a random permutation of the* b_i*'s. Then* $\mathbf{E}[Y] = O(q^{3/2}\sqrt{k})$*.*

Proof: Assume $k \geq q$, or else $Y \leq k < q$ and the claimed bound holds trivially. For $\alpha \in [q]$, let Y_α be a random variable (over the choice of the permutation π of the sequence) counting the number of $i \in [k]$ such that the prefix $(b_{\pi(1)}, \ldots, b_{\pi(i)})$ has a plurality tie, α achieves the plurality, and $b_{\pi(i)} \neq \alpha$. Then $Y \leq \sum_\alpha Y_\alpha$. (For every prefix with a plurality tie, at least one of the two symbols achieving the plurality must be different from the last symbol in the prefix.) Thus, it suffices to show that $\mathbf{E}[Y_\alpha] = O(\sqrt{qk})$ for every α.

Fix α. Let ℓ be the number of occurrences of α in the sequence $b_1, \ldots, b_k$. We can obtain a random permutation of $b_1, \ldots, b_k$ by randomly ordering the $m = k - \ell$ elements of the sequence other than α, and then randomly merging the ℓ occurrences of α into this sequence (uniformly out of all $\binom{\ell+m}{\ell}$ ways). In fact we will bound the expectation of Y_α for every fixed ordering $c_1, \ldots, c_m$ of the elements other than α, and thus the only randomness is over the merging.

For each $r = 1, \ldots, m$, let $u_r = \#\mathsf{plur}(c_1, \ldots, c_r)$. Notice that $r \geq u_r \geq r/(q-1)$. Let X_r be the indicator random variable for whether upon merging, α occurs exactly u_r times before c_r (equivalently, occurs $v_r = \ell - u_r$ times after c_r). Then $Y_\alpha = \sum_{r=1}^m X_r$.

Fix $r \in [m]$, let $s = m - r$, $u = u_r$, $v = v_r = \ell - u$. Our aim is to bound $\mathbf{Pr}[X_r = 1]$. Notice that the merging can be viewed as uniformly choosing a set S of ℓ out of $m + \ell$ locations to place the α's (and putting the c_i's in the remaining m locations). Observe that $X_r = 1$ only if S contains exactly u of the first $u+r$ locations (and thus exactly v of the last $v+s$ locations); let E_r denote this event. ($X_r = 1$ also implies that S does not contain location $u+r$, but we will not make use of that.)

We bound $\mathbf{Pr}[E_r]$ for $r \in \{q, q+1, \ldots, m-1\}$ by considering two cases. (For $r < q$ and $r = m$, we will use the trivial bound $\mathbf{Pr}[E_r] \leq 1$.) First, suppose that $u/r > 2v/s$. Intuitively, this means that, for E_r to occur, S must be disproportionately partitioned in the merging. Specifically, the expected number of elements of S among the first $u+r$ locations is

$$\frac{u+r}{u+r+v+s} \cdot (u+v) < \frac{u}{2}.$$

By Chernoff bounds, the probability that the first $u+r$ locations contain more than u elements of S is at most $2^{-\Omega(u)} \leq 2^{-\Omega(r/q)} \leq O(\sqrt{q/r})$ for $r \geq q$. (The indicators for whether each location contains an element of S satisfy "negative dependence", and thus Chernoff bounds apply [DS].)

The second case is that

$$\frac{v}{s} \geq \frac{u}{2r} \geq \frac{1}{2(q-1)} \ . \tag{6}$$

Then, by Lemma 16,

$$\begin{aligned} \mathbf{Pr}[E_r] &= \binom{u+r}{r}\binom{v+s}{s}\Big/\binom{u+v+r+s}{r+s} \\ &= O\left(\sqrt{\frac{(u+r)\cdot(v+s)\cdot(u+v)\cdot(r+s)}{r\cdot u\cdot s\cdot v\cdot(u+r+v+s)}}\right) \ . \end{aligned} \tag{7}$$

Now for positive integers x, y, we have

$$\frac{x\cdot y}{x+y} = \frac{\min\{x,y\}\cdot\max\{x,y\}}{x+y} = \Theta(\min\{x,y\}) \ . \tag{8}$$

From (7) and (8), we conclude

$$\begin{aligned} \mathbf{Pr}[E_r] &= O\left(\sqrt{\frac{\min\{u+r, v+s\}}{\min\{u,v\}\cdot\min\{r,s\}}}\right) \\ &= O\left(\sqrt{\frac{q}{\min\{r,s\}}}\right) \quad \text{(using (6)).} \end{aligned}$$

Thus, in both cases we have $\mathbf{Pr}[E_r] = O(\sqrt{q/\min\{r,s\}})$ for $r \in \{q, q+1, \ldots, m\}$. Therefore

$$\begin{aligned} \mathbf{E}[Y_\alpha] &\leq q + \sum_{r=q}^{m-1} \mathbf{Pr}[E_r] \\ &= q + \sum_{r=1}^{m-1} O\left(\sqrt{\frac{q}{\min\{r, m-r\}}}\right) \\ &= O(q + \sqrt{qm}) \\ &= O(\sqrt{qk}), \end{aligned}$$

since $m \leq k$ and $q \leq k$. This gives the desired bound on $\mathbf{E}[Y_\alpha]$ for each $\alpha \in [q]$. ∎