Electronic Colloquium on Computational Complexity, Report No. 70 (2020)

On the list recoverability of randomly punctured codes

Ben Lund [*] Aditya Potukuchi [†]

May 3, 2020

Abstract

We show that a random puncturing of a code with good distance is list recoverable beyond the Johnson bound. In particular, this implies that there are Reed-Solomon codes that are list recoverable beyond the Johnson bound. It was previously known that there are Reed-Solomon codes that do not have this property. As an immediate corollary to our main theorem, we obtain better degree bounds on unbalanced expanders that come from Reed-Solomon codes.

1 Introduction

List recoverable codes were defined by Guruswami and Rudra [GR06] to demonstrate a barrier to improving known algorithms for list decoding. Here, we study list recoverable codes in their own right, showing that random puncturings of codes over a sufficiently large alphabet are list recoverable. Our result is analogous to earlier work by Rudra and Wooters [RW14, RW15] on the list decodability of randomly punctured codes.

We use q to denote the alphabet size, and n to denote the block length of an arbitrary code $\mathcal{C} \subset [q]^n$. Given two codewords $c_1, c_2 \in [q]^n$, denote the Hamming distance between c_1 and c_2 by $\Delta(c_1, c_2)$. Denote the minimum distance between a codeword $c \in [q]^n$ and a set $\mathcal{L} \subseteq [q]^n$ by $\Delta(c, \mathcal{L})$.

Definition 1.1 (List recoverability)**.** *Let q, n, k be positive integers, and let $\delta > 0$ and $0 \leq \rho < 1$ be real numbers. A code $\mathcal{C} \subset [q]^n$ is (ℓ, δ, ρ) list recoverable if, for every collection of sets $\{L_i \subseteq [q]\}_{i \in [n]}$ with $|L_i| \leq \ell$ for each i, we have*

$$|\{c \in \mathcal{C} \mid \Delta(c, L_1 \times \cdots \times L_n) \leq \rho n\}| \leq \ell(1 + \delta).$$

[*] Department of Mathematics, Princeton University. lund.ben@gmail.com. Research supported by NSF grant DMS-1802787.

[†] Department of Computer Science, Rutgers University. aditya.potukuchi@cs.rutgers.edu.

ISSN 1433-8092

In the above definition, ℓ is called the *list size* from which the code can be recovered. The case $\rho = 0$ is already interesting, and called *zero-error* list recoverability. We say that a code $\mathcal{C}$ is (ℓ, δ) zero-error list recoverable if it is $(\ell, \delta, 0)$ list recoverable.

A *puncturing* of a code $\mathcal{C} \subset [q]^n$ to a set $S \subset [n]$ is the code $\mathcal{C}_S \subset [q]^S$ defined by $\mathcal{C}_S[i] = \mathcal{C}[i]$ for each $i \in S$. A punctured code will typically have higher rate, but lower distance, than the unpunctured version. Our main result is that every code over a large enough alphabet $[q]$ can be punctured to a code of rate $R > q^{-1/2}$ while being list recoverable with list size roughly R^{-2}. It may be helpful to first consider the case $\rho = 0$ of the following theorem.

Theorem 1.2. *There are positive constants c, n_0, and q_0 so that the following holds. Let $0 < \delta \leq 1$ and $0 \leq \rho < 1 - (1 + \delta)^{-1/2}$ be real numbers. Denote $\gamma = (1 + \delta)(1 - \rho)^2 - 1$ and $\sigma = (1 - \rho)(2 - \rho)^{-1}$. Let $n > n_0$ and $q > q_0$ be integers. Let $q^{-1/2} < \epsilon < \min(c, 2^{-1}\gamma\sigma)$. Then, every code $\mathcal{C} \subset [q]^n$ with distance at least $n(1 - q^{-1} - \epsilon^2)$ can be punctured to rate $\Omega\left(\frac{\epsilon}{\log q}\right)$ so that it is $(\epsilon^{-2}\sigma^2\gamma, \delta, \rho)$-list recoverable.*

Note that (ℓ, δ, ρ)-list recoverability implies (ℓ, δ', ρ)-list recoverability for any $\delta' \geq \delta$. The hypothesis $\delta \leq 1$ in Theorem 1.2 is needed only because we bound the list size as a function of δ.

To attempt to make the parameters more transparent, we would like to draw the reader's focus to the list size, i.e., $\epsilon^{-2}\sigma^2\gamma$. The main point here is that this is as large as ϵ^{-2}, so one way to interpret the above theorem is that we get $(O_{\delta,\rho}(\epsilon^{-2}), \delta, \rho)$-list recoverability after the aforementioned puncturing. In fact, we show a random puncturing of $\mathcal{C}$ is list recoverable with the same list size with high probability; see Theorem 2.1 for a precise statement.

Theorem 1.2 is analogous to a theorem of Rudra and Wooters [RW14, RW15] on the *list decodability* of punctured codes over large alphabets. A code $\mathcal{C} \subset [q]^n$ is (ρ, ℓ)-list decodable if for each $x \in [q]^n$, there are at most ℓ codewords of $\mathcal{C}$ that differ from x in fewer than ρn coordinates.

Theorem 1.3 ([RW15])**.** *Let $\epsilon > q^{-1/2}$ be a real number, and q, n be sufficiently large integers. Every code $\mathcal{C} \subset [q]^n$ with distance $n(1 - q^{-1} - \epsilon^2)$ can be punctured to rate $\tilde{\Omega}\left(\frac{\epsilon}{\log q}\right)$ so that it is $(1 - O(\epsilon), O(\epsilon^{-1}))$-list decodable.*

Theorems 1.2 and 1.3 are most interesting in the case of Reed-Solomon codes. The codewords of the degree-d Reed-Solomon code over $\mathbb{F}_q$ with evaluation set $S \in \binom{[q]}{m}$ are the evaluations of all univariate polynomials of degree at most d on elements of S. In other words, suppose $S = \{s_1, \ldots, s_m\}$. The degree-d Reed-Solomon code on S is the set

$$\{(p(s_1), \ldots, p(s_m)) \mid \deg(p) \leq d\}.$$

The block length of this code is $m \leq q$. Since two distinct polynomials of degree at most d can agree on at most d locations, the distance of any degree-d Reed-Solomon code is at least $m - d$.

A fundamental result, which gives a lower bound on the list decodability of a code with given distance, is the *Johnson bound* (see, for example Corollary 3.2 in [Gur06]).

Theorem 1.4 (Johnson bound for list decoding)**.** *Every code $\mathcal{C} \subset [q]^n$ of minimum distance at least $n(1-(1/q)-\epsilon^2)$ is $(n(1-q^{-1}-\epsilon), O(\epsilon^{-1}))$- list decodable.*

One of the main points of Theorem 1.3 is that it shows that there are Reed-Solomon codes that are list decodable beyond the Johnson bound.

A similar result as Theorem 1.4, using a similar argument, also known as the Johnson bound, is known for list recoverability (see for example, Lemma 5.2 in [GKdO$^+$18]).

Theorem 1.5 (Johnson bound for list recovery)**.** *Let $\mathcal{C} \subseteq [q]^n$ be a code of relative distance α. Then $\mathcal{C}$ is (ℓ, δ, ρ)-list recoverable for any $\rho \leq 1 - \sqrt{\ell(1-\alpha)}$ where $\delta = \frac{\alpha}{(1-\rho)^2 - \ell(1-\alpha)} - 1$.*

A result of Guruswami and Rudra [GR06]) shows that there are Reed-Solomon codes that are not list recoverable beyond the Johnson bound.

Theorem 1.6. *Let $q = p^m$ where p is a prime, and let $\mathcal{C}$ denote the degree-$\left(\frac{p^m-1}{p-1}\right)$ Reed-Solomon code over $\mathbb{F}_q$ with $\mathbb{F}_q$ as the evauation set. Then there are lists $S_1, \ldots, S_q$ each of size p such that*

$$|\mathcal{C} \cap (S_1 \times \cdots \times S_q)| = q^{2^m}$$

To understand this, recall that a degree-d Reed-Solomon code has relative distance $1 - \frac{1}{q} - \frac{d}{q}$. Setting $\ell = p-1$ and $\rho = 0$ in the Johnson bound tells us that such a code is $(p-1, O(q), 0)$-list recoverable. Setting the list size as p in the bound gives us nothing, and Theorem 1.6 says that the number of codewords grows superpolynomially in q. On the other hand, Theorem 1.2 immediately gives the following corollary.

Corollary 1.7. *For a prime power q and $\epsilon \geq q^{-1/2}$, there are Reed-Solomon codes of rate $\tilde{\Omega}\left(\frac{\epsilon}{\log q}\right)$ which are $(q/2, 1/2)$-list recoverable.*

Again, one can easily check that setting $k = q/2$ in the Johnson bound gives nothing.

1.1 Unbalanced expander graphs from codes

The zero-error case of Theorem 1.2 leads to some progress on a question of Guruswami regarding unbalanced expanders obtained from Reed-Solomon graphs. This was also the main motivation behind this theorem.

Informally, an expander graph is a graph where every small set of vertices has a relatively large neighborhood. In this case, we say that all small sets *expand.* One interesting type of expander graphs are *unbalanced expanders.* These are bipartite graphs where one side is much larger than the other side, and we want that all the small subsets of the *larger* side expand.

Definition 1.8 (Unbalanced expander)**.** *A (k, d, ϵ)-regular unbalanced expander is a bipartite graph on vertex set $L \sqcup R$, $|L| \geq |R|$ where the degree of every vertex in L is d, and for every $S \subseteq L$ such that $|S| = k$, we have that $|N(S)| \geq d|S|(1 - \epsilon)$.*

Note that in the above definition, $|N(S)| \leq d|S|$. We are typically interested in infinite families of unbalanced expanders for which $\epsilon = o(1)$, $d = o(|R|)$, and $k = \tilde{\Omega}(|R|/d)$.

For a q-ary error correcting code $\mathcal{C} \subset [q]^n$, and a subset $S := \{i_1, \ldots, i_{|S|}\} \subseteq [n]$ with $i_1 < \cdots < i_{|S|}$, we use $\mathcal{C}_S$ to denote the S-punctured code given by

$$\mathcal{C}_S := \{(c_{i_1}, \ldots, c_{i_{|S|}}) \mid (c_1, \ldots, c_n) \in \mathcal{C}\}.$$

Thus, $\mathcal{C}_S$ is just the set of codewords of $\mathcal{C}$ restricted to the coordinates in S.

Given a code $\mathcal{C} \subseteq [q]^n$, it is natural to look at the bipartite graph, which we will denote by $G(\mathcal{C})$ where the vertex sets are $\mathcal{C} \sqcup ([n] \times [q])$. For every $c = (c_1, \ldots, c_n) \in \mathcal{C}$ the set of neighbors is $\{(1, c_1), \ldots, (n, c_n)\}$. This graph is especially interesting when $\mathcal{C}$ is a low-degree Reed-Solomon code evaluated at an appropriate set.

The following is a open question in the study of pseudorandomness that is attributed to Guruswami [Gur], (also explicitly stated in [CZ18]): Fix an integer d. For a subset $S \in \binom{[q]}{m}$, define $\mathcal{C}_S$ to be the degree-d Reed-Solomon code with S as the evaluation set, where d is a constant.

Question: *What is the smallest m such that when S is chosen uniformly at random, $G(\mathcal{C}_S)$ is, with high probability, a $(o(q), o(1))$-unbalanced expander?*

There are examples of explicit constructions unbalanced expanders that come from other means (in fact, other codes) [GUV09]. However, the above "natural" geometric/combinatorial question is still interesting in its own right and so far, seems to evade known techniques.

It was probably well known that $m = \Omega(\log q)$, and we also give a proof of this (Theorem 3.1) since we could not find it in the literature. But for upper bounds, it seems nothing better than the almost trivial $m = O(q)$ was known [Che]. Since the zero-error list recoverability of $\mathcal{C}$ is equivalent to the expansion of $G(\mathcal{C})$, an immediate Corollary to Theorem 2.1 gives an improved upper bound.

Corollary 1.9. *Let q, n be sufficiently large integers and $\delta \in (0, 1)$, $\epsilon > q^{-1/2}$ be real numbers. For every code $\mathcal{C} \subset [q]^n$ with relative distance $1 - q^{-1} - \epsilon^2$, there is a subset $S \subset [n]$ such that $|S| = O(\epsilon n \log q)$ such that $G(\mathcal{C}_S)$ is a $(\delta\epsilon^{-2}, |S|, \delta)$-unbalanced expander.*

Instantiating the above theorem for degree-d Reed-Solomon codes, we have $n = q$ and $\epsilon = (d/q)^{-\frac{1}{2}}$. This gives, $m = \tilde{O}(\sqrt{q})$.

2 Proof of Theorem 1.2

The bulk of this section is the statement and proof of Theorem 2.1. After the proof of Theorem 2.1, we show how to derive Theorem 1.2 from it.

2.1 A sketch of the proof

Here, we sketch the proof when $\rho = 0$, i.e., for *zero-error* list recovery. This contains most of the main ideas required for the general theorem. Let $S = \{x_1, \ldots, x_m\} \subset [n]$ be a randomly chosen evaluation set. The main observation is that if there are input lists $L_1, \ldots, L_m \subseteq [q]$, such that $(L_1 \times \cdots \times L_m)$ contains a large subset $\mathcal{D} \subseteq \mathcal{C}$ of codewords, then there is a small subset $\mathcal{C}' \subseteq \mathcal{D} \subseteq \mathcal{C}$ which agree on an unusually high number of coordinates. An appropriately sized random subset of $\mathcal{D}$ does this. Thus the event that a given puncturing is bad is contained witnessed by the event that there are few codewords that agree a lot on the coordinates chosen in S. The number of events of the latter kind are far fewer in number, leaving us with fewer bad events to overcome for the union bound.

2.2 Proof of Theorem 1.2

The calculations in the proof of Theorem 2.1 are all explicit, but we have not tried to optimize the constant terms.

Theorem 2.1. *Let $0 < \delta < 1$ and $0 \le \rho < 1 - (1+\delta)^{-1/2}$ be real numbers. Let $q, n, d, \ell,$ and m be positive integers. Let $\mathcal{C} \subset [q]^n$ be a code of distance at least $n - nq^{-1} - d$. Denote $\gamma = (1+\delta)(1-\rho)^2 - 1$ and $\sigma = (1-\rho)(2-\rho)^{-1}$. Suppose that the following inequalities are satisfied:*

$$\begin{aligned} d &\ge nq^{-1}, \\ 4\gamma^{-1} &\le \ell \le 800^{-1}\,\sigma\gamma n d^{-1}, \\ \sigma m &\ge 1280\sqrt{\ell\gamma^{-1}}\,\log|\mathcal{C}|, \\ m &< n. \end{aligned}$$

Then, for $S \in \binom{[n]}{m}$ chosen uniformly at random, the probability that $\mathcal{C}_S$ is (ℓ, δ, ρ)-list recoverable is at least $1 - e^{-\sigma m/64}$.

Proof. For any $\mathcal{C}' \subseteq \mathcal{C}$, denote by $T(\mathcal{C}')$ the set of coordinates $i \in [n]$ such that there is a pair $c_1, c_2 \in \mathcal{C}'$ with $c_1[i] = c_2[i]$.

The basic outline of the proof is to first show that, for any S such that $\mathcal{C}_S$ is not (ℓ, δ, ρ)-list recoverable, there is a pair $S', \mathcal{C}'$ such that S' is large and $|T(\mathcal{C}') \cap S'|$ is unusually large. Taking a union bound over all candidates for $\mathcal{C}'$ then shows that there cannot be too many pairs of this sort.

Let $S \in \binom{[n]}{m}$ so that $\mathcal{C}_S$ is not (ℓ, δ, ρ)-list recoverable. We will show that there is a set $\mathcal{C}' \subset \mathcal{C}_S$ such that

$$|\mathcal{C}'| \leq 10\sqrt{\ell/\gamma}, \text{ and} \tag{1}$$

$$|T(\mathcal{C}') \cap S| \geq \sigma m/4. \tag{2}$$

Since $\mathcal{C}_S$ is not (ℓ, δ, ρ)-list recoverable, there are subsets $L_i \subseteq [q]$ for each $i \in S$ such that each $|L_i| \leq \ell$ and $|\{c \in \mathcal{C}_S : \Delta(c, \prod_{i \in S} L_i) \leq \rho n\}| > k(1+\delta)$.

Let

$$\mathcal{D} = \{c \in \mathcal{C}_s : \Delta(c, \prod_{i \in S} L_i) \leq \rho n\}.$$

For $i \in S$, let

$$\mathcal{D}_i = \{c \in \mathcal{D} : c[i] \in L_i\}.$$

Let

$$I = \{(c,i) \in \mathcal{D} \times S : c \in \mathcal{D}_i\}.$$

From the definition of $\mathcal{D}$, we have

$$|I| \geq |\mathcal{D}|(1-\rho)m. \tag{3}$$

Note that the average cardinality of the $\mathcal{D}_i$ is $(1-\rho)|\mathcal{D}|$. Let

$$S' = \{i \in S : |\mathcal{D}_i| \geq (1-\rho)^2|\mathcal{D}|\}.$$

If $\rho = 0$, then $\mathcal{D}_i = \mathcal{D}$ for each i, and hence $|S'| = m$. Next we show that, if $\rho > 0$, then $|S'| \geq (1-\rho)(2-\rho)^{-1}m = \sigma m$. Since $|\mathcal{D}_i \leq |\mathcal{D}|$ for each i, we have

$$|S'|\,|\mathcal{D}| \geq \sum_{i \in S'} |D_i| = |I| - \sum_{i \in S \setminus S'} \mathcal{D}_i. \tag{4}$$

Since $|D_i| < (1-\rho)^2|\mathcal{D}|$ for each $i \in S \setminus S'$, we have

$$\sum_{i \in S \setminus S'} \leq (m - |S'|)(1-\rho)^2|\mathcal{D}|. \tag{5}$$

A straightforward rearrangement of (3), (4), and (5) using the assumption that $\rho > 0$ leads to the claimed lower bound on $|S'|$:

$$|S'| \geq \sigma m. \tag{6}$$

Since $\sigma < 1$, the bound $|S'| \geq \sigma m$ holds for the case $\rho = 0$ as well.

For each $i \in S'$, choose a set $P_i \subset \binom{\mathcal{D}}{2}$ of $|P_i| \geq \gamma k/2$ disjoint pairs of codewords in $\mathcal{D}_i$ such that for each $\{c_1, c_2\} \in \mathcal{P}_i$, we have $c_1[i] = c_2[i]$. This is always possible since $|L_i| \leq \ell$ and $|\mathcal{D}_i| \geq (1+\rho)^2|\mathcal{D}| \geq (1+\gamma)\ell$.

Now choose $\mathcal{C}'$ randomly by including each element of $\mathcal{D}$ with probability $p = (\gamma\ell/2)^{-1/2}\ell(1+\delta)|\mathcal{D}|^{-1}$. Since $\ell \geq 4\gamma^{-1}$ by hypothesis and $|\mathcal{D}| \geq \ell(1+\delta)$

by the assumption that $\mathcal{C}_S$ is not (ℓ, δ, ρ)-list recoverable, we have $p < 1$. The expected size of $\mathcal{C}'$ is

$$\mathbf{E}[|\mathcal{C}'|] = p|\mathcal{D}| \leq (\gamma/(2\ell))^{-1/2}(1+\delta) \leq (8\ell/\gamma)^{1/2}.$$

We remark that this is the only place where we use the assumption that $\delta < 1$.

For any fixed pair $c_1 \neq c_2$ of codewords in $\mathcal{D}$, the probability that both are included in $\mathcal{C}'$ is p^2. Since the pairs in P_i are disjoint, the events that two distinct pairs $\{c_1, c_2\}, \{c_3, c_4\} \in P_i$ are both included in $\mathcal{C}'$ are independent. Hence, the probability that no pair in P_i is included in $\mathcal{C}'$ is $(1-p^2)^{|P_i|} < e^{-p^2|P_i|} < 1/2$. Consequently, for each fixed $i \in S'$, the probability that $i \in T(\mathcal{C}')$ is greater than $1/2$. By linearity of expectation, $\mathbf{E}[|T(\mathcal{C}') \cap S'|] \geq |S'|/2 \geq \sigma m/2$.

Let

$$Y = |T(\mathcal{C}') \cap S'| - \frac{\sigma m}{4} \frac{|\mathcal{C}'|}{\mathbf{E}[|\mathcal{C}'|]}.$$

By linearity of expectation, $\mathbf{E}[Y] \geq \sigma m/4$, hence there is some specific choice of $\mathcal{C}'$ for which $Y \geq \sigma m/4$. This can hold only if $|T(\mathcal{C}') \cap S| \geq |T(\mathcal{C}') \cap S'| \geq m/4$ and $|\mathcal{C}'| \leq 3\mathbf{E}(|\mathcal{C}'|)$ simultaneously, which establishes (1) and (2).

Next we bound the probability that, for a fixed choice of $\mathcal{C}'$ and random S, we have have $|T(C') \cap S|$ large. Let $\mathcal{C}' \subset \mathcal{C}$ be an arbitrary set of $|\mathcal{C}'| \leq 10\ell^{1/2}\gamma^{-1/2}$ codewords. Since the distance of $\mathcal{C}'$ is at least $n - nq^{-1} - d$ and $d \geq nq^{-1}$, we have

$$|T(\mathcal{C}')| \leq (nq^{-1} + d)\binom{|\mathcal{C}'|}{2} < d|\mathcal{C}'|^2. \tag{7}$$

For $S \in \binom{[n]}{m}$ chosen uniformly at random, $|T(\mathcal{C}') \cap S|$ follows a hypergeometric distribution. Specifically, we are making m draws from a population size of n of which $|T(\mathcal{C}')| \leq d|\mathcal{C}'|^2$ contribute to $|T(\mathcal{C}') \cap S|$. Using the assumption that $\ell \leq \gamma\sigma n(800d)^{-1}$, the expected value of $|T(\mathcal{C}') \cap S|$ is

$$\mathbf{E}\left[|T(\mathcal{C}') \cap S|\right] \leq d|\mathcal{C}'|^2 n^{-1} m \leq 100\frac{d\ell}{\gamma n} m \leq \frac{\sigma m}{8}. \tag{8}$$

Next we use the following large deviation inequality for hypergeometric random variables (see [DP09]). Let X be a hypergeometric random variable with mean μ. Then for any $\alpha \geq 1$,

$$\mathbb{P}(X \geq (1+\alpha)\mu) \leq \exp(-\alpha\mu/4). \tag{9}$$

Together with (8), this gives

$$\mathbb{P}(|T(\mathcal{C}') \cap S| \geq \sigma m/4) \leq \exp\left(-\frac{\sigma m}{32}\right). \tag{10}$$

Finally, we take a union over all candidates for $\mathcal{C}'$. Let X be the event that $\mathcal{C}_S$ is not (ℓ, δ, ρ) list recoverable, with $S \in \binom{[n]}{m}$ uniformly at random. Using the assumption that $\sigma m \geq 1280\sqrt{\ell/\gamma}\log|\mathcal{C}|$, we have

$$\begin{aligned}
\mathbb{P}(X) &\leq \sum_{\mathcal{C}'\subset\mathcal{C}_S:|\mathcal{C}'|\leq 10\sqrt{\ell/\gamma}} \mathbb{P}(|T(\mathcal{C}'\cap S)| \geq \sigma m/4) \\
&\leq \binom{|\mathcal{C}|}{\lceil 10\sqrt{\ell/\gamma}\rceil + 1} \exp\left(-\frac{m}{32}\right) \\
&< \exp\left(20\sqrt{\ell/\gamma}\log|\mathcal{C}| - \sigma m/32\right) \\
&\leq \exp(-\sigma m/64),
\end{aligned}$$

as claimed. □

We now show how to derive Theorem 1.2 from Theorem 2.1.

Proof of Theorem 1.2. Suppose we have δ, ρ, n, q, and ϵ as in the hypotheses of Theorem 1.2. Let $m = \lceil 1280\epsilon^{-1}\log|C|\rceil$. The singleton bound combined with the assumption that $\epsilon < c$ for a suitably chosen absolute constant c implies that $m < n$. Choose $S \in \binom{[n]}{m}$ uniformly at random. The rate of $\mathcal{C}_S$ is

$$R = \log|\mathcal{C}|(m\log q)^{-1} = \Omega(\epsilon(\log q)^{-1}).$$

It is straightforward to check that the hypotheses of Theorem 2.1 are satisfied if we take $\ell = \epsilon^{-2}\sigma^2\gamma$, and hence we have that $\mathcal{C}_S$ is $(\epsilon^{-2}\sigma^2\gamma, \delta, \rho)$-list recoverable with high probability. □

3 Upper bound

Here we show the aforementioned upper bound for the rate to which a degree-d Reed-Solomon code over $\mathcal{F}_q$ can be randomly punctured to be $(q/2, 1/2)$-zero-error list-recoverable.

First, we recall a bit of standard and relevant sumset notation. For a group G and subsets $A, B \subseteq G$, we denote the sumset $A+B = \{a+b \mid a \in A,\ b \in B\}$. Clearly, we have $|A+B| \leq |A|\cdot|B|$. If $G = \mathbb{Z}_p$, then for $n < p/2$, we have that $[n]+[n] = \{2, \ldots, 2n\}$. We are now ready to state and prove the upper bound.

Theorem 3.1. *Let $m = o(\log q)$, and $X = \{x_0, \ldots, x_m\}$ be a uniformly random subset of $\mathbb{F}_q$ where q is a prime. Then every $d \geq 1$, the degree-d Reed-Solomon code with the evaluation set at X is, with high probability, not $(q/2, 1/2)$-zero-error list-recoverable.*

Proof. Let $X = \{x_0, \ldots, x_m\}$. Let n be a large number such that $n^m = o(\sqrt{q})$. We are using the fact that $m = o(\log q)$ for the existence of such an n. W.L.O.G assume $x_0 = 0$ and $x_1 = 1$ (if $0, 1 \notin S$, then adding them to S only makes the lower bound stronger). Consider the two sets

$$X_0 = \frac{1}{1-x_2}[n] + \cdots \frac{1}{1-x_{m-1}}[n]$$

and

$$X_1 = \frac{1}{x_2}[n] + \cdots \frac{1}{x_{m-1}}[n].$$

Claim 3.2. *With high probability over the choice of X, we have that $|X_0|, |X_1| = \Omega(\binom{n^{m-2}}{})$.*

Proof. We do the proof for X_0, the case for X_1 follows analogously. Let P be the set of "collisions" in X_0. Formally:

$$P := \left\{ (a_2, \ldots, a_{m-2}, b_2, \ldots, b_{m-2}) \mid \sum_{i=2}^{m-2} a_i x_i = \sum_{i=2}^{m-2} b_i x_i \right\}.$$

So the number of distinct elements in X_0 is at least $n^{m-2} - |P|$. We observe that

$$\begin{aligned} \mathbf{E}[|P|] &= \sum_{\substack{a_2,\ldots,a_{m-2}\in[n] \\ b_2,\ldots,b_{m-2}\in[n]}} \mathbb{P}\left(\sum_{i=2}^{m-2} a_i x_i = \sum_{i=2}^{m-2} b_i x_i \right) \\ &\leq \frac{1}{p} n^{2m-4} \\ &= o(n^{m-2}). \end{aligned}$$

So by Markov's Inequality, with high probability, $|X_0| \sim n^{m-2}$. □

Consider $\mathcal{D}$, the set of degree-1 Reed-Solomon codes given by the lines

$$\{Y = aX + b\}_{b \in X_0, a \in X_1}.$$

First, we note that $|Y| = \Omega(n^{2m-4})$. Geometrically, $\mathcal{D}$ is just the set of all lines passing through some point of $\{0\} \times X_0$ and $\{1\} \times X_1$. Clearly, $\{c[0] \mid c \in \mathcal{C}\} = X_0$ and $\{c[1] \mid c \in \mathcal{D}\} = X_1$. For $i \neq 0, 1$, let us similarly define $X_i := \{c[x_i] \mid c \in \mathcal{D}\}$.We have that

$$\begin{aligned} X_i &= \{a(1-x_i) + bx_i\}_{b \in X_0, a \in X_1} \\ &= (1-x_i)\left(\frac{1}{1-x_2}[n] + \cdots \frac{1}{1-x_{m-1}}[n] \right) + x_i \left(\frac{1}{x_2}[n] + \cdots \frac{1}{x_{m-1}}[n] \right) \\ &= \left([n] + \sum_{2\leq j\leq m,\ j\neq i} \frac{1-x_i}{1-x_j}[n] \right) + \left([n] + \sum_{2\leq j\leq m,\ j\neq i} \frac{x_i}{x_j}[n] \right) \\ &= \{2, \ldots, 2n\} + \sum_{2\leq j\leq m,\ j\neq i} \frac{1-x_i}{1-x_j}[n] + \sum_{2\leq j\leq m,\ j\neq i} \frac{x_i}{x_j}[n]. \end{aligned}$$

Thus, $|X_i| \leq (2n) \times n^{2m-6} \leq 2n^{2m-5}$.

This shows that there are lists $X_0, X_1, \ldots, X_m$ each of size at most $\ell := 2n^{2m-5}$ such that there are at least $\Omega(n^{2m-4}) = \ell^{1+\frac{1}{k}}$ codewords, namely $\mathcal{D}$, contained in $X_0 \times \cdots \times X_m$. □

For a fixed d, the above theorem rules out hope of randomly puncturing degree-d Reed-Solomon codes to rate $\omega\left(\frac{1}{\log q}\right)$ for the desired list recoverability. We believe that this is essentially the barrier. We state the concrete conjecture that we alluded to in Section 1.1.

Conjecture 3.3. *For any $\delta > 0$, the degree-d Reed-Solomon code with evaluation set $\mathbb{F}_q$ can be randomly punctured to rate $\Omega_d\left(\frac{1}{\log q}\right)$ so that is it $(\delta q, \delta)$-list recoverable with high probability.*

4 Discussion and open problems

The main open problem that we would like to showcase is Conjecture 3.3. This was probably believed to be true but we could not find it written down explicitly in the literature. List recoverable codes have connections to various other combinatorial objects (see [Vad07]) and if true, Conjecture 3.3 could lead to the construction of some other interesting combinatorial objects.

The second open problem is to derandomize Theorem 1.2, i.e., to find an *explicit* Reed-Solomon code which is list recoverable beyond the Johnson bound at least in the zero-error case. Understanding how these evaluation sets look like could lead to progress on Conjecture 3.3, or could be interesting in its own right.

Finally, the last open problem is that given a Reed-Solomon code $\mathcal{C} \subset [q]^m$ of rate R on a randomly chosen evaluation set S, find an efficient algorithm for list recovery, i.e., take input lists $L_1, \ldots, L_m$ of size $O(R^{-2}(\log q)^{-1})$, and output all the codewords contained in $L_1 \times \cdots \times L_m$ with high probability (over the choice of S and the randomness used by the algorithm). This would also likely require some understanding of the properties of the evaluation set.

References

[Che] Xue Chen. personal communication.

[CZ18] Xue Chen and David Zuckerman. Existence of simple extractors. *Electronic Colloquium on Computational Complexity (ECCC)*, 25:116, 2018.

[DP09] Devdatt Dubhashi and Alessandro Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms.* Cambridge University Press, New York, NY, USA, 1st edition, 2009.

[GKdO$^+$18] Sivakanth Gopi, Swastik Kopparty, Rafael Mendes de Oliveira, Noga Ron-Zewi, and Shubhangi Saraf. Locally testable and locally correctable codes approaching the Gilbert-Varshamov bound. *IEEE Trans. Information Theory*, 64(8):5813–5831, 2018.

[GR06] Venkatesan Guruswami and Atri Rudra. Limits to list decoding Reed-Solomon codes. *IEEE Trans. Information Theory*, 52(8):3642–3649, 2006.

[Gur] Venkatesan Guruswami. personal communication.

[Gur06] Venkatesan Guruswami. Algorithmic results in list decoding. *Foundations and Trends in Theoretical Computer Science*, 2(2), 2006.

[GUV09] Venkatesan Guruswami, Christopher Umans, and Salil P. Vadhan. Unbalanced expanders and randomness extractors from parvaresh-vardy codes. *J. ACM*, 56(4):20:1–20:34, 2009.

[RW14] Atri Rudra and Mary Wootters. Every list-decodable code for high noise has abundant near-optimal rate puncturings. In *Symposium on Theory of Computing, STOC 2014, New York, NY, USA, May 31 - June 03, 2014*, pages 764–773, 2014.

[RW15] Atri Rudra and Mary Wootters. It'll probably work out: Improved list-decoding through random operations. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science, ITCS 2015, Rehovot, Israel, January 11-13, 2015*, pages 287–296, 2015.

[Vad07] Salil P. Vadhan. The unified theory of pseudorandomness: guest column. *SIGACT News*, 38(3):39–54, 2007.