List-decodability with large radius for Reed–Solomon codes

Asaf Ferber* Matthew Kwan† Lisa Sauermann‡

June 4, 2021

Abstract

List-decodability of Reed–Solomon codes has received a lot of attention, but the best-possible dependence between the parameters is still not well-understood. In this work, we focus on the case where the list-decoding radius is of the form $r = 1 - \varepsilon$ for ε tending to zero. Our main result states that there exist Reed–Solomon codes with rate $\Omega(\varepsilon)$ which are $(1-\varepsilon, O(1/\varepsilon))$-list-decodable, meaning that any Hamming ball of radius $1-\varepsilon$ contains at most $O(1/\varepsilon)$ codewords. This trade-off between rate and list-decoding radius is best-possible for any code with list size less than exponential in the block length.

By achieving this trade-off between rate and list-decoding radius we improve a recent result of Guo, Li, Shangguan, Tamo, and Wootters, and resolve the main motivating question of their work. Moreover, while their result requires the field to be exponentially large in the block length, we only need the field size to be polynomially large (and in fact, almost-linear suffices). We deduce our main result from a more general theorem, in which we prove good list-decodability properties of random puncturings of any given code with very large distance.

1 Introduction

Reed–Solomon codes are a family of error-correcting codes that have been studied intensively in many different contexts since they were introduced in [11]. As the parameters of the code, consider a prime power q and integers $1 \leq k < n \leq q$. Then, for n distinct *evaluation points* $\alpha_1, \ldots, \alpha_n \in \mathbb{F}_q$, the $[n,k]$-Reed–Solomon code with these evaluation points is defined to be the set of *codewords*

$$\mathcal{C}^{(k)}_{\alpha_1,\ldots,\alpha_n} := \{(f(\alpha_1), \ldots, f(\alpha_n)) \mid f \in \mathbb{F}_q[x], \deg f < k\} \subseteq \mathbb{F}_q^n.$$

One reason for the great interest in Reed–Solomon codes is that they behave optimally with respect to the classical *unique decoding* problem, having an optimal trade-off between rate and distance. For an alphabet Σ of size $|\Sigma| = q$, and a code $\mathcal{C} \subseteq \Sigma^n$, the *rate* of $\mathcal{C}$ is defined to be $\log_q |\mathcal{C}|/n$, and the *distance* of $\mathcal{C}$ is defined to be the minimum Hamming distance between a pair of distinct codewords $\gamma, \gamma' \in \mathcal{C}$ (recall that the Hamming distance between γ and γ' is the number of positions in which γ and γ' disagree). Every $[n,k]$-Reed–Solomon code has rate k/n and distance $n-k+1$. By the Singleton bound [15], this is the highest possible rate for any code of this distance. In addition, due to their simple structure, Reed–Solomon codes allow for efficient algorithms[1].

An important generalization of the unique decoding problem is the problem of *list-decoding*, and properties of Reed–Solomon codes with respect to list-decodability are much less understood. Roughly speaking, while the unique encoding problem demands that the original codeword can be uniquely reconstructed from a noisy

*Department of Mathematics, University of California, Irvine. Email: asaff@uci.edu. Research supported in part by NSF Awards DMS-1954395 and DMS-1953799.

†Department of Mathematics, Stanford University, Stanford, CA. Email: mattkwan@stanford.edu. Research supported by NSF Award DMS-1953990.

‡School of Mathematics, Institute for Advanced Study, Princeton, NJ. Email: lsauerma@mit.edu. Research supported by NSF Grant CCF-1900460 and NSF Award DMS-2100157.

[1] In this paper, we are not concerned with algorithmic questions, and only study the combinatorial properties of Reed–Solomon codes.

signal, for the list-decoding problem we are satisfied with a short list of candidate codewords for a noisy signal. List-decodability was first introduced by Elias and Wozencraft [4, 18] in the 1950s, and has since been used in several different areas of theoretical computer science. Regarding list-decodability of Reed–Solomon codes specifically, there are applications in complexity theory and the theory of pseudorandomness [3, 10, 16]. The problem of understanding the (combinatorial) list-decodability of Reed–Solomon codes has been raised by many researchers over the last two decades (see for example [6, p. 111], [12, p. 120], and [17, Problem 5.20]), and there has been a lot of recent work investigating this problem [5, 13, 14]. Still, a lot of questions remain open.

In order to formally define what it means for a code to be list-decodable, we need to introduce some more definitions. Given $r \in (0,1)$, an alphabet Σ, and $\beta \in \Sigma^n$, the *Hamming ball* of (relative) radius r centered at β is defined as

$$B_r(\beta) := \{\gamma \in \Sigma^n \mid \gamma[i] = \beta[i] \text{ for at least } (1-r)n \text{ positions } 1 \leq i \leq n\}$$

(here, by $\gamma[i]$ we denote the symbol in the i-th position of $\gamma \in \Sigma^n$). In other words, this Hamming ball consists of all points $\gamma \in \Sigma^n$ that differ in most rn coordinates from β.

A code $\mathcal{C} \subseteq \Sigma^n$ is called (r, L)*-list-decodable* (for some radius $r \in (0,1)$ and some *list size* $L \in \mathbb{N}$) if we have $|\mathcal{C} \cap B_r(\beta)| \leq L$ for all $\beta \in \Sigma^n$. In other words, $\mathcal{C}$ is (r, L)-list-decodable if each Hamming ball of (relative) radius r in Σ^n contains at most L codewords from $\mathcal{C}$. Note that for list size $L = 1$, the setting of (r, L)-list-decodability precisely corresponds to the classical unique decoding setting. In this paper, we are primarily interested in list-decodability for Reed–Solomon codes.

For any radius $r \in (0,1)$ and any list size L, one can ask for the maximum possible rate of an (r, L)-list-decodable Reed–Solomon code. Shangguan and Tamo [14] posed a precise conjecture for this general question, and made some partial progress towards their conjecture. Here, we focus on the case of radius $r = 1 - \varepsilon$, for ε tending to zero. The main problem we investigate is how large the rate can be for a $(1-\varepsilon, L)$-list-decodable $[n, k]$-Reed–Solomon code (for growing n), when the list size L is not too large (say, not exponential in n).

There are some general results which immediately give upper and lower bounds for this problem. First, the *list-decoding capacity theorem* (see for example [8, Theorem 7.4.1]) implies that the rate of any $(1-\varepsilon, L)$-list-decodable code (where the list size L is less than exponential in the block length n) can be at most[2] $O(\varepsilon)$. Second, the *Johnson bound* [9] gives a general bound on the list-decodability of a code in terms of its distance, and implies that *every* $[n, k]$-Reed–Solomon code with rate $k/n = \varepsilon^2$ is $(1-\varepsilon, qn^2)$-list-decodable. In particular, there exist $(1-\varepsilon, L)$-list-decodable Reed–Solomon codes that have rate ε^2, where the list size L is polynomial in n. Thus, the highest possible rate for the above problem lies somewhere between $\Omega(\varepsilon^2)$ and $O(\varepsilon)$.

Recently, Guo, Li, Shangguan, Tamo, and Wootters [5] made major progress on closing this gap, improving the lower bound ε^2 obtained from the Johnson bound[3]. They proved that over very large fields there exist $(1-\varepsilon, O(1/\varepsilon))$-list-decodable Reed–Solomon codes with rate $\Omega(\varepsilon/\log(1/\varepsilon))$, matching the list-decoding capacity upper bound up to a logarithmic factor. They stated that their "motivating question is whether or not RS codes can be list-decoded up to radius $1-\varepsilon$ with rates $\Omega(\varepsilon)$", and this question remained open.

Our main result resolves this question in the affirmative, closing the gap to the list-decoding capacity upper bound (up to constant factors). We prove that over sufficiently large fields there exist $(1-\varepsilon, O(1/\varepsilon))$-list-decodable Reed–Solomon codes with rate $\Omega(\varepsilon)$. This means that, up to constant factors, Reed–Solomon codes achieve the highest possible rate among all $(1-\varepsilon, L)$-list-decodable codes where the list size L is smaller than exponential in the block length n. A more precise statement of our main result is as follows.

Theorem 1. *Fix a constant $c \geq 5$. Let $\varepsilon \in (0,1)$, let $n \in \mathbb{N}$ be sufficiently large with respect to ε and c, and let q be a prime power with $q \geq n^{c/(c-1)}$. Then there exist $(1-\varepsilon, \lceil 3/\varepsilon \rceil)$-list-decodable $[n, k]$-Reed–Solomon codes over $\mathbb{F}_q$ with rate at least $\varepsilon/(3c)$.*

As mentioned above, the rate $\varepsilon/(3c)$ here is tight up to the constant factor $3c$. Furthermore, Theorem 1 also improves the above-mentioned result of [5] in terms of the required field size q: in [5], the field size q needs to

[2] More precisely, as n grows, the rate of such a code cannot be bounded above ε.

[3] Rudra and Wootters [13], in an earlier work, also proved lower bounds that improve on the Johnson bound in certain regimes of q and ε. See also the comment further below.

be exponential in the block length n, whereas Theorem 1 only assumes the polynomial bound $q \geq n^{c/(c-1)}$. By choosing a large constant c, the exponent $c/(c-1)$ can be taken arbitrarily close to 1. In this sense, we can take the field size to be almost-linear in the block length n.

Similarly to the approach in [5], we actually show that one can obtain the desired Reed–Solomon codes in Theorem 1 via a random choice of the evaluation points $(\alpha_1, \ldots, \alpha_n)$: For suitable parameters $L = O(1/\varepsilon)$, n and $k = \Omega(\varepsilon n)$, and for sufficiently large q, we prove that for almost all choices of $(\alpha_1, \ldots, \alpha_n) \in \mathbb{F}_q^n$ the $[n,k]$-Reed–Solomon code $\mathcal{C}^{(k)}_{\alpha_1,\ldots,\alpha_n}$ is $(1-\varepsilon, L)$-list-decodable (and has rate $k/n = \Omega(\varepsilon)$).

Theorem 2. *Fix a constant $c \geq 5$. Let $\varepsilon \in (0,1)$, let $n \in \mathbb{N}$ be sufficiently large with respect to ε and c, and let $k = \lceil \varepsilon n/(3c) \rceil$. Furthermore, let q be a prime power with $q \geq n^{c/(c-1)}$. Then for a uniformly random choice of an n-tuple $(\alpha_1, \ldots, \alpha_n) \in \mathbb{F}_q^n$ with distinct entries $\alpha_1, \ldots, \alpha_n$, the Reed–Solomon code $\mathcal{C}^{(k)}_{\alpha_1,\ldots,\alpha_n}$ with rate $k/n \geq \varepsilon/(3c)$ is $(1-\varepsilon, \lceil 3/\varepsilon \rceil)$-list-decodable with probability at least $1 - q^{-\varepsilon n/(13c)}$.*

Note that Theorem 2 immediately implies Theorem 1. Our proof approach for Theorem 2 is inspired by the approach of Guo, Li, Shangguan, Tamo, and Wootters in [5], which in turn builds on the ideas in earlier work of Shangguan and Tamo [14]. However, our proof is significantly simpler and much shorter.

In fact, we deduce Theorem 2 from a much more general result about random puncturings of arbitrary codes with very large distance. A *puncturing* of a code $\mathcal{C} \subseteq \Sigma^m$ to a set $S \subseteq [m]$ is defined to be the code $\mathcal{C}_S \subseteq \Sigma^S$ whose codewords are obtained by restricting all the codewords in $\mathcal{C}$ to only the positions in S. Formally, $\mathcal{C}_S = \{(\gamma[i])_{i \in S} \mid \gamma \in \mathcal{C}\}$. We will consider random puncturings of a given code $\mathcal{C}$ obtained by choosing a uniformly random subset $S \subseteq [m]$ of a given size n (then the puncturing $\mathcal{C}_S$ has block length n).

We prove the following general result concerning list-decodability of random puncturings of a given code with large distance. Roughly speaking, this result states that for a code with block length m and distance $m-h$ (for some small h), a random puncturing with block length n is likely to be list-decodable with radius $1 - O(h/n)$ and list size $O(n/h)$, provided that the alphabet is large enough and n is not too big. In order to deduce Theorem 2, we apply Theorem 3 to the "full" $[q,k]$-Reed–Solomon code $\mathcal{C} \subseteq \mathbb{F}_q^q$ where the evaluation points are all of the q points in $\mathbb{F}_q$.

Theorem 3. *Fix a constant $c \geq 5$, and let $q \in \mathbb{N}$ be sufficiently large with respect to c. Suppose that $h, m \in \mathbb{N}$ are such that $h \leq q^{-1/c} \cdot m$, and let $\mathcal{C} \subseteq \Sigma^m$ be a code over an alphabet Σ of size $|\Sigma| = q$ such that $\mathcal{C}$ has distance at least $m - h$. Then for any $n \in \mathbb{N}$ satisfying*

$$3c \cdot h < n \leq \min\left(\sqrt{\log_2 q} \cdot \sqrt{c/8} \cdot h \,,\, e^{h/(4c^3)} \cdot (c/2) \cdot h \right),$$

a random puncturing of $\mathcal{C}$ of block length n is $(1 - (3ch/n), \lfloor n/(ch) \rfloor)$-list-decodable with probability at least $1 - q^{-h/4}$. In particular, there exist $(1 - (3ch/n), \lfloor n/(ch) \rfloor)$-list-decodable puncturings of $\mathcal{C}$ of block length n.

We made no particular effort to optimize the constants in the theorems above.

We remark that Rudra and Wootters [13] previously also proved results concerning list-decodability of random puncturings of codes with large distance. However, the details of their results and our Theorem 3 differ significantly. In particular, while our theorem requires a much larger distance of the code $\mathcal{C}$, in their results the block length n of the puncturing needs to be larger. For this reason, with their results one cannot obtain Reed–Solomon codes with rates as large as in Theorem 2.

Let us also briefly comment on some other works related to our main result, Theorem 2 (a more detailed review of the relevant literature can be found in [5, Section 1.2]). Using their random puncturing results mentioned above, Rudra and Wootters [13] proved a result similar to Theorem 2, but only with rate $\Omega(\varepsilon/(\log^5(1/\varepsilon) \log q))$. This is weaker than our Theorem 2 and than the result of Guo, Li, Shangguan, Tamo, and Wootters [5], and in particular due to the factor $\log q$ in the denominator the rate bound in [13] always goes to zero as n grows (since $q \geq n$). Shangguan and Tamo [14] proved a result of a similar spirit as Theorem 2 for small list sizes $L = 2$ and $L = 3$ (which in particular means that the radius r is bounded away from 1), but with an optimal trade-off between radius, rate and list size (more precisely, for given rate and list size $L \in \{2,3\}$ their result gives the exact best-possible list-decoding radius). On a different note, while Theorem 2 shows that almost all choices of the evaluation points $(\alpha_1, \ldots, \alpha_n) \in \mathbb{F}_q^n$ lead to $(1-\varepsilon, \lfloor 10/\varepsilon \rfloor)$-list-decodable Reed–Solomon codes, it is plausible that some choices of $(\alpha_1, \ldots, \alpha_n)$ fail to have this property.

There are some related results of Guruswami and Rudra [7] and of Ben-Sasson, Kopparty, and Radhakrishnan [2] pointing in this direction ([7] shows that for some choices of $(\alpha_1, \ldots, \alpha_n)$ the code $\mathcal{C}^{(k)}_{\alpha_1,\ldots,\alpha_n}$ fails to satisfy a stronger property called list-recoverability, and [2] shows a negative result concerning the list-decodability of Reed–Solomon codes in the case $q = n$ where up to permutation there is only one choice of the evaluation points $(\alpha_1, \ldots, \alpha_n)$).

Notation. Let $\mathbb{N} = \{1, 2, 3, \ldots\}$, and for $n \in \mathbb{N}$ let $[n] = \{1, \ldots, n\}$.

2 Proofs

Recall that Theorem 2 implies Theorem 1. We now show how Theorem 2 follows from Theorem 3.

Proof of Theorem 2. Let $c \geq 5$, let $\varepsilon \in (0, 1)$, and let $n \in \mathbb{N}$ be sufficiently large with respect to ε and c. Let $k = \lceil \varepsilon n/(3c) \rceil$ and let q be a prime power with $q \geq n^{c/(c-1)}$.

In order to apply Theorem 3, let $m = q$ and $h = k - 1 \leq \varepsilon n/(3c)$. Note that we have $h \leq n \leq q^{(c-1)/c} = q^{-1/c} \cdot m$ and

$$3c \cdot \lceil \varepsilon n/(3c) \rceil < n \leq \min\left(\sqrt{\log_2 q} \cdot \sqrt{c/8} \cdot (\lceil \varepsilon n/(3c) \rceil - 1)\,,\, e^{(\lceil \varepsilon n/(3c) \rceil - 1)/(4c^3)} \cdot (c/2) \cdot (\lceil \varepsilon n/(3c) \rceil - 1) \right),$$

by the assumption that n (and therefore also q) is sufficiently large with respect to ε and c.

Let us consider the alphabet $\Sigma = \mathbb{F}_q$ and the "full" $[q, k]$-Reed–Solomon code $\mathcal{C} \subseteq \mathbb{F}_q^q$ where the evaluation points are all of the q points in $\mathbb{F}_q$. Note that $\mathcal{C}$ has distance $q - k + 1 = m - h$.

Hence all assumptions of Theorem 3 are satisfied, and we can conclude that a random puncturing of $\mathcal{C}$ of block length n is $(1 - (3ch/n), \lfloor n/(ch) \rfloor)$-list-decodable with probability at least $1 - q^{-h/4} \geq 1 - q^{-(\varepsilon n/(12c)) + (1/4)} \geq 1 - q^{-\varepsilon n/(13c)}$ (using that n is sufficiently large with respect to ε and c). Noting that $1 - (3ch/n) \geq 1 - \varepsilon$ and $n/(ch) \leq n/(\varepsilon n/3 - c) < (3/\varepsilon) + 1$ (again, as n is sufficiently large with respect to ε and c), this implies that such a random puncturing of $\mathcal{C}$ is $(1 - \varepsilon, \lceil 3/\varepsilon \rceil)$-list-decodable with probability at least $1 - q^{-\varepsilon n/(13c)}$. In other words, for a uniformly random choice of an n-tuple $(\alpha_1, \ldots, \alpha_n) \in \mathbb{F}_q^n$ with distinct entries $\alpha_1, \ldots, \alpha_n$, the Reed–Solomon code $\mathcal{C}^{(k)}_{\alpha_1,\ldots,\alpha_n}$ is $(1 - \varepsilon, \lceil 3/\varepsilon \rceil)$-list-decodable with probability at least $1 - q^{-\varepsilon n/(13c)}$. □

Our aim for the rest of this section is to prove Theorem 3. We deduce Theorem 3 from the following theorem. This approach is motivated by [5] and [14], even though the setting there is specific to Reed–Solomon codes.

Theorem 4. *Fix a constant $c \geq 5$, suppose that $q, h, m \in \mathbb{N}$ are such that $h \leq q^{-1/c} \cdot m$, and let $\mathcal{C} \subseteq \Sigma^m$ be a code over an alphabet Σ of size $|\Sigma| = q$ such that $\mathcal{C}$ has distance at least $m - h$.*

Suppose L is a non-negative integer satisfying $L < e^{h/(4c^3)} - 2$. Let $n \in \mathbb{N}$ and consider subsets $I_1, \ldots, I_{L+1} \subseteq [n]$ such that

$$\sum_{j=1}^{L+1} |I_j| - \left| \bigcup_{j=1}^{L+1} I_j \right| > 2chL. \tag{1}$$

Let us say that an n-tuple $(a_1, \ldots, a_n) \in [m]^n$ with distinct entries $a_1, \ldots, a_n$ is bad *if there exist a point $\beta \in \Sigma^m$ and distinct codewords $\gamma_1, \ldots, \gamma_{L+1} \in \mathcal{C}$ such that for all $j = 1, \ldots, L + 1$ and all $i \in I_j$ we have $\gamma_j[a_i] = \beta[a_i]$. Then there are at most $q^{-h/2} \cdot m^n$ bad n-tuples $(a_1, \ldots, a_n) \in [m]^n$.*

Guo, Li, Shangguan, Tamo, and Wootters [5, Theorem 6.3] proved a statement similar to Theorem 4 in the specific setting of Reed–Solomon codes. However, their statement gives a weaker bound for the number of bad n-tuples and requires a stronger version of the assumption (1), where the term on the right-hand side of (1) is larger by a factor of $\Theta(\log L)$. This additional logarithmic factor leads to the logarithmic loss in the rate $\Omega(\varepsilon/\log(1/\varepsilon))$ of the Reed–Solomon codes in their result (and the weaker bound for the number of bad n-tuples leads to them requiring the field size q to be exponential in the block length n).

Let us now show the deduction of Theorem 3 from Theorem 4. This deduction is fairly standard (similar arguments appear in [5, 14]). Afterwards, at the end of this section, we will present the proof of Theorem 4.

Proof of Theorem 3. Let us define $L = \lfloor n/(ch) \rfloor \geq 3$, and note that by the assumptions on n we have

$$L + 2 \leq 2L \leq \frac{2n}{ch} \leq e^{h/(4c^3)}.$$

Also note that

$$\frac{n + 2chL}{L+1} < \frac{n}{L+1} + 2ch \leq \frac{n}{n/(ch)} + 2ch = 3ch. \tag{2}$$

Let us also remark that the assumptions in Theorem 3 (including the assumption that q is sufficiently large with respect to c) imply that

$$\frac{n}{m} \leq \frac{\sqrt{\log_2 q} \cdot \sqrt{c/8} \cdot h}{q^{1/c} \cdot h} = \sqrt{c/8} \cdot \frac{\sqrt{\log_2 q}}{q^{1/c}} < \frac{1}{2} \tag{3}$$

and

$$\frac{2n^2}{m} \leq \frac{2 \cdot \log_2 q \cdot (c/8) \cdot h^2}{q^{1/c} \cdot h} = \frac{1}{4} \cdot c \cdot h \cdot \log_2 q \cdot q^{-1/c} < \frac{1}{12} \cdot h \cdot \log_2 q. \tag{4}$$

We need to show that a (uniformly) random puncturing of $\mathcal{C}$ of block length n is $(1 - (3ch/n), L)$-list-decodable with probability at least $1 - q^{-h/4}$. We can model the choice of such a random puncturing by taking a uniformly random n-tuple $(a_1, \ldots, a_n) \in [m]^n$ with distinct entries $a_1, \ldots, a_n$ and considering the puncturing $\mathcal{C}_S$ for $S = \{a_1, \ldots, a_n\}$. Note that $\mathcal{C}_S$ fails to be $(1 - (3ch/n), L)$-list-decodable if and only if there exist a point $\beta \in \Sigma^m$ and distinct codewords $\gamma_1, \ldots, \gamma_{L+1} \in \mathcal{C}$ such that for each $j = 1, \ldots, L+1$ we have $\gamma_j[s] = \beta[s]$ for at least $3ch$ elements $s \in S$. Recalling that $S = \{a_1, \ldots, a_n\}$, this condition is equivalent to having $\gamma_j[a_i] = \beta[a_i]$ for at least $3ch$ indices $i \in [n]$.

Hence, if for our random choice of $(a_1, \ldots, a_n) \in [m]^n$ the puncturing $\mathcal{C}_S$ fails to be $(1 - (3ch/n), L)$-list-decodable, then for each $j = 1, \ldots, L+1$ we can find a set $I_j \subseteq [n]$ of size $|I_j| \geq 3ch$ such that we have $\gamma_j[a_i] = \beta[a_i]$ for all $i \in I_j$. With the notation in Theorem 4, this means that the n-tuple $(a_1, \ldots, a_n) \in [m]^n$ is bad with respect to the subsets $I_1, \ldots, I_{L+1}$.

Note that there are at most $(2^n)^{L+1}$ possibilities to choose subsets $I_1, \ldots, I_{L+1} \subseteq [n]$ with $|I_j| \geq 3ch$ for $j = 1, \ldots, L+1$. For any such choice of subsets, by (2) we have

$$\sum_{j=1}^{L+1} |I_j| - \left| \bigcup_{j=1}^{L+1} I_j \right| \geq (L+1) \cdot 3ch - n > (L+1) \cdot \frac{n + 2chL}{L+1} - n = 2chL.$$

Hence, by Theorem 4, for any fixed choice of $I_1, \ldots, I_{L+1}$, there are at most $q^{-h/2} \cdot m^n$ different n-tuples $(a_1, \ldots, a_n) \in [m]^n$ which are bad with respect to $I_1, \ldots, I_{L+1}$. Overall, this means that there are at most $2^{n(L+1)} \cdot q^{-h/2} \cdot m^n$ different n-tuples $(a_1, \ldots, a_n) \in [m]^n$ which are bad with respect to some choice of subsets $I_1, \ldots, I_{L+1} \subseteq [n]$ (with $|I_j| \geq 3ch$ for $j = 1, \ldots, L+1$). Thus, the number of n-tuples $(a_1, \ldots, a_n) \in [m]^n$ with distinct entries $a_1, \ldots, a_n$, such that the puncturing $\mathcal{C}_S$ for $S = \{a_1, \ldots, a_n\}$ is not $(1 - (3ch/n), L)$-list-decodable, is at most

$$2^{n(L+1)} \cdot q^{-h/2} \cdot m^n \leq 2^{(4/3)Ln} \cdot q^{-h/2} \cdot m^n \leq 2^{(4/3)n^2/(ch)} \cdot q^{-h/2} \cdot m^n \leq q^{h/6} \cdot q^{-h/2} \cdot m^n = q^{-h/3} \cdot m^n,$$

where for the third inequality we used the assumption that $n \leq \sqrt{\log_2 q} \cdot \sqrt{c/8} \cdot h$.

Finally, note that the total number of n-tuples $(a_1, \ldots, a_n) \in [m]^n$ with distinct entries $a_1, \ldots, a_n$ is

$$m(m-1)\cdots(m-n+1) \geq \left(1 - \frac{n}{m}\right)^n \cdot m^n \geq 2^{-2n^2/m} \cdot m^n \geq q^{-h/12} \cdot m^n.$$

Here, we used that $1 - x \geq 2^{-2x}$ for all $x \in (0, 1/2)$, as well as (3) and (4).

All in all, this means that for a random choice of an n-tuple $(a_1, \ldots, a_n) \in [m]^n$ with distinct entries $a_1, \ldots, a_n$, the probability that the puncturing $\mathcal{C}_S$ for $S = \{a_1, \ldots, a_n\}$ fails to be $(1 - (3ch/n), L)$-list-decodable is at most

$$\frac{q^{-h/3} \cdot m^n}{q^{-h/12} \cdot m^n} = q^{-h/4}.$$

Hence a random puncturing of $\mathcal{C}$ of block length n is $(1 - (3ch/n), L)$-list-decodable with probability at least $1 - q^{-h/4}$, as desired. □

It remains to prove Theorem 4. This is the part of this paper requiring new ideas. Roughly speaking, the proof strategy is as follows. Recall that an n-tuple $(a_1, \ldots, a_n) \in [m]^n$ is called bad if there are distinct codewords $\gamma_1, \ldots, \gamma_{L+1} \in \mathcal{C}$ and a point $\beta \in \Sigma^m$ such that $\gamma_j[a_i] = \beta[a_i]$ whenever $i \in I_j$. Our goal is to prove an upper bound on the number of bad n-tuples $(a_1, \ldots, a_n)$. The key idea of the proof is to find a relatively small set of indices $Z \subseteq [n]$, such that specifying a_i and $\beta[a_i]$ for all $i \in Z$ already uniquely determines all of the codewords $\gamma_1, \ldots, \gamma_{L+1}$ (via the condition that $\gamma_j[a_i] = \beta[a_i]$ whenever $i \in I_j$, and the assumption that $\mathcal{C}$ has large distance). Once the codewords $\gamma_1, \ldots, \gamma_{L+1}$ are determined, for any distinct $j, j' \in \{1, \ldots, L+1\}$ and any $i \in I_j \cap I_{j'}$, there are only a small number of choices for $a_i \in [m]$. Indeed, we must have $\gamma_j[a_i] = \beta[a_i] = \gamma_{j'}[a_i]$, so a_i must be one of the few positions in which the codewords γ_j and $\gamma_{j'}$ agree. Overall, we obtain the desired upper bound for the number of bad n-tuples $(a_1, \ldots, a_n)$ by a counting argument that takes all of these restricted choices into account.

Proof of Theorem 4. We prove the theorem by induction on L. First, note that the statement is vacuously true for $L = 0$, because it is impossible for the condition $|I_1| - |I_1| > 2ch \cdot 0$ in (1) to be satisfied.

Let us now assume that $L \geq 1$, and that we have already proved the theorem for $L - 1$. First, we consider the case that for some index $t \in \{1, \ldots, L+1\}$ we have

$$\left| I_t \cap \bigcup_{j \in \{1,\ldots,L+1\} \setminus \{t\}} I_j \right| < 2ch.$$

Let us assume without loss of generality that $t = L + 1$, then we have

$$\left| I_{L+1} \cap \bigcup_{j=1}^{L} I_j \right| < 2ch.$$

But now (1) implies that

$$\sum_{j=1}^{L} |I_j| - \left| \bigcup_{j=1}^{L} I_j \right| = \sum_{j=1}^{L+1} |I_j| - \left| \bigcup_{j=1}^{L+1} I_j \right| - \left| I_{L+1} \cap \bigcup_{j=1}^{L} I_j \right| > 2chL - 2ch = 2ch(L-1).$$

This means that we can apply the induction hypothesis to $L - 1$ and the sets $I_1, \ldots, I_L$. This shows that the number of bad n-tuples $(a_1, \ldots, a_n) \in [m]^n$ is at most $q^{-h/2} \cdot m^n$, since every n-tuple which is bad for the sets $I_1, \ldots, I_{L+1}$ must also be bad for the sets $I_1, \ldots, I_L$.

So we may from now on assume that for all $t = 1, \ldots, L+1$ we have

$$\left| I_t \cap \bigcup_{j \in \{1,\ldots,L+1\} \setminus \{t\}} I_j \right| \geq 2ch.$$

Now let $M \subseteq [n]$ be the set of those elements $i \in [n]$ that are contained in at least two of the sets $I_1, \ldots, I_{L+1}$. Note that for each $t = 1, \ldots, L+1$, we have

$$|M \cap I_t| = \left| I_t \cap \bigcup_{j \in \{1,\ldots,L+1\} \setminus \{t\}} I_j \right| \geq 2ch.$$

In particular, we have $|M| \geq 2ch$.

Claim 5. *There exists a set $Z \subseteq M$ of size $|Z| \leq |M|/(2c-2)$ such that $|Z \cap I_t| > h$ for all $t = 1, \ldots, L+1$.*

Proof. Let us choose the set $Z \subseteq M$ randomly by including each element of M into the set Z independently with probability $1/(2c-1)$. By the Chernoff bound (see for example [1, Theorem A.1.4]), we have that

$$\Pr\left(|Z| > \frac{|M|}{2c-2} \right) < \exp\left(-\frac{2}{|M|} \cdot \left(\frac{|M|}{(2c-2)(2c-1)} \right)^2 \right) \leq e^{-|M|/(8c^4)} \leq e^{-h/(4c^3)}.$$

Furthermore, for each $t = 1, \ldots, L+1$, each element of $M \cap I_t$ is an element of the set Z independently with probability $1/(2c-1)$. Hence, again by the Chernoff bound, we have (recalling that $|M \cap I_t| \geq 2ch$)

$$\Pr(|Z \cap I_t| \leq h) \leq \exp\left(-\frac{2}{|M \cap I_t|} \cdot \left(\frac{|M \cap I_t|}{2c-1} - h\right)^2\right) = \exp\left(-2|M \cap I_t| \cdot \left(\frac{1}{2c-1} - \frac{h}{|M \cap I_t|}\right)^2\right)$$
$$\leq \exp\left(-2 \cdot 2ch \cdot \left(\frac{1}{2c-1} - \frac{1}{2c}\right)^2\right) = \exp\left(-\frac{4ch}{(2c)^2(2c-1)^2}\right) \leq e^{-h/(4c^3)}.$$

All in all, by a union bound, the probability of having $|Z| \leq |M|/(2c-2)$ and $|Z \cap I_t| > h$ for all $t = 1, \ldots, L+1$ is at least

$$1 - e^{-h/(4c^3)} - (L+1) \cdot e^{-h/(4c^3)} = 1 - (L+2) \cdot e^{-h/(4c^3)} > 0$$

(recalling our assumption that $L < e^{h/(4c^3)} - 2$). This means that the desired set $Z \subseteq M$ exists. □

Let us now fix a set $Z \subseteq M$ as in Claim 5. Now we can show the desired upper bound on the number of bad n-tuples $(a_1, \ldots, a_n) \in [m]^n$ in the following way. Recall that for a bad n-tuple $(a_1, \ldots, a_n) \in [m]^n$ there exist a point $\beta \in \Sigma^m$ and distinct codewords $\gamma_1, \ldots, \gamma_{L+1} \in \mathcal{C}$ such that for all $j = 1, \ldots, L+1$ and all $i \in I_j$ we have $\gamma_j[a_i] = \beta[a_i]$.

Note that we have at most $m^{|Z|}$ choices for the elements a_i for all $i \in Z$ (recall that the elements a_i need to all be distinct). Furthermore, there are $|\Sigma|^{|Z|} = q^{|Z|}$ possibilities for the values $\beta[a_i]$ for all $i \in Z$. Now, knowing a_i and $\beta[a_i]$ for all $i \in Z$ already determines the codewords $\gamma_1, \ldots, \gamma_{L+1}$. Indeed, for each $t = 1, \ldots, L+1$ we have $|Z \cap I_t| > h$ and $\gamma_t[a_i] = \beta[a_i]$ for all $i \in Z \cap I_t$ (and the coordinates $a_i \in [m]$ for $i \in Z \cap I_t$ are distinct). Since any two codewords in $\mathcal{C}$ agree in at most h positions (as $\mathcal{C}$ has distance at least $m-h$), for each $t = 1, \ldots, L+1$ there is at most one possible codeword γ_t satisfying $\gamma_t[a_i] = \beta[a_i]$ for all $i \in Z \cap I_t$. Thus, after choosing a_i and $\beta[a_i]$ for all $i \in Z$, there is at most one possibility for the codewords $\gamma_1, \ldots, \gamma_{L+1}$.

Furthermore, knowing the codewords $\gamma_1, \ldots, \gamma_{L+1}$ there are at most h possibilities for each $a_i \in [m]$ with $i \in M \setminus Z$. Indeed, for each $i \in M \setminus Z$ there exist two distinct indices $j, j' \in \{1, \ldots, L+1\}$ with $i \in I_j \cap I_{j'}$ and we must have $\gamma_j[a_i] = \beta[a_i] = \gamma_{j'}(a_i)$. Hence the codewords γ_j and $\gamma_{j'}$ must agree in position a_i. However, as $\mathcal{C}$ has distance at least $m-h$, the codewords γ_j and $\gamma_{j'}$ agree in at most h positions, and so there are at most h possible choices for a_i. Thus, for each $i \in M \setminus Z$, there are indeed at most h choices for a_i and altogether this gives at most $h^{|M|-|Z|}$ choices for determining all the the elements $a_i \in [m]$ with $i \in M \setminus Z$.

Finally, there are at most $m^{n-|M|}$ choices for the elements $a_i \in [m]$ with $i \in [n] \setminus M$. All in all, this means that the number of possible choices for a bad n-tuple $(a_1, \ldots, a_n) \in [m]^n$ is at most

$$m^{|Z|} \cdot q^{|Z|} \cdot h^{|M|-|Z|} \cdot m^{n-|M|} = \left(\frac{h}{m}\right)^{|M|} \left(\frac{qm}{h}\right)^{|Z|} m^n$$
$$\leq \left(\frac{h}{m}\right)^{|M|} \left(\frac{qm}{h}\right)^{|M|/(2c-2)} m^n = \left(\frac{h}{m} \cdot q^{1/(2c-3)}\right)^{\frac{2c-3}{2c-2} \cdot |M|} m^n$$
$$\leq \left(q^{-1/c} \cdot q^{1/(2c-3)}\right)^{\frac{2c-3}{2c-2} \cdot |M|} m^n = q^{-\frac{c-3}{2c-2} \cdot \frac{1}{c} \cdot |M|} m^n \leq q^{-|M|/(4c)} m^n \leq q^{-h/2} m^n.$$

Here, we used the assumptions $h \leq q^{-1/c} \cdot m$ and $c \geq 5$ as well as $|Z| \leq |M|/(2c-2)$ and $|M| \geq 2ch$. □

3 Concluding remarks

We have proved that there exist Reed–Solomon codes which are list-decodable with radius $1 - \varepsilon$ (and polynomial list size) and have rate $\Omega(\varepsilon)$. Moreover, such codes exist with block length n and field size q whenever n is sufficiently large and $q \geq n^{1+\delta}$, for any constant $\delta > 0$. There are several interesting further directions of research.

First, our result uses the probabilistic method and is fundamentally non-constructive. It would be very interesting if, in the setting of Theorem 2, one could achieve the same bound with *explicit* choices of evaluation

points $\alpha_1, \ldots, \alpha_n \in \mathbb{F}_q$. In fact, it would be interesting if one could beat the Johnson bound at all with an explicit Reed–Solomon code (we remark that there are constructions in [5, 14] which are in a certain sense explicit, but they require an exponential field size and therefore do not lead to efficient algorithms).

Second, it would be interesting to further improve the bounds in Theorem 1. While our field size requirement $q \geq n^{1+\delta}$ is much weaker than the requirement in [5], it would still be interesting to sharpen this further: does it suffice to assume that $q \geq Cn$ for some constant C? Also, it would be nice to optimize the constant factors in the trade-off between the rate and the list-decoding radius. In particular, it seems likely that there should exist Reed–Solomon codes which are list-decodable with radius $1 - \varepsilon$ (and polynomial list size) and have rate $(1 - o(1))\varepsilon$. An exact conjecture for the best-possible relationship between rate, list-decoding radius and list size was made by Shangguan and Tamo [14].

Acknowledgements. We would like to thank Shachar Lovett for introducing us to list-decodability of Reed–Solomon codes, and Avi Wigderson for many very helpful suggestions.

References

[1] N. Alon and J. H. Spencer, *The Probabilistic Method*, 4th ed., Wiley, 2016.

[2] E. Ben-Sasson, S. Kopparty, and J. Radhakrishnan, *Subspace polynomials and limits to list decoding of Reed–Solomon codes*, IEEE Trans. Inform. Theory **56** (2010), 113–120.

[3] Jin-Yi Cai, Aduri Pavan, and D Sivakumar, *On the hardness of permanent*, In Annual Symposium on Theoretical Aspects of Computer Science (STACS 1999), pages 90–99, 1999.

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

[5] Zeyu Guo, Ray Li, Chong Shangguan, Itzhak Tamo, and Mary Wootters, *Improved List-Decodability of Reed–Solomon Codes via Tree Packings*, preprint, 2020, arXiv:2011.04453.

[6] Venkatesan Guruswami, *List Decoding of Error-Correcting Codes, Winning Thesis of the 2002 ACM Doctoral Dissertation Competition*, Lecture Notes in Computer Science, vol. 3282, Springer, 2004.

[7] V. Guruswami and A. Rudra, *Limits to list decoding Reed–Solomon codes*, IEEE Trans. Inform. Theory **52** (2006), 3642–3649.

[8] Venkatesan Guruswami, Atri Rudra, and Madhu Sudan, *Essential coding theory*, book draft (2019), available at `https://cse.buffalo.edu/faculty/atri/courses/coding-theory/book/`.

[9] Selmer Johnson, *A new upper bound for error-correcting codes*, IRE Transactions on Information Theory **8** (1962), 203–207.

[10] Ben Lund and Aditya Potukuchi, *On the list recoverability of randomly punctured codes*, In Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2020), article no. 30, 2020.

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

[12] Atri Rudra, *List decoding and property testing of error-correcting codes*, PhD thesis, University of Washington, 2007.

[13] Atri Rudra and Mary Wootters, *Every list-decodable code for high noise has abundant near-optimal rate puncturings*, In Proceedings of the 46th Annual ACM Symposium on Theory of Computing (STOC 2014), pages 764–773, 2014.

[14] Chong Shangguan, Itzhak Tamo, *Combinatorial list-decoding of Reed-Solomon codes beyond the Johnson radius*, preprint, 2019, arXiv:1911.01502.

[15] R. Singleton, *Maximum distance q-nary codes*, IEEE Trans. Inform. Theory **10** (1964), 116–118.

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

[17] Salil P. Vadhan, *Pseudorandomness*, Foundations and Trends in Theoretical Computer Science, vol. 7, 2012.

[18] John M. Wozencraft, *List decoding*, In Quarterly Progress Report, Research Laboratory of Electronics, MIT, vol. 48, pages 90–95, 1958.