Reed Solomon Codes Against Adversarial Insertions and Deletions

Roni Con[*] Amir Shpilka [†] Itzhak Tamo [‡]

Abstract

In this work, we study the performance of Reed–Solomon codes against adversarial insertion-deletion (insdel) errors.

We prove that over fields of size $n^{O(k)}$ there are $[n, k]$ Reed-Solomon codes that can decode from $n-2k+1$ insdel errors and hence attain the half-Singleton bound. We also give a deterministic construction of such codes over much larger fields (of size $n^{k^{O(k)}}$). Nevertheless, for $k = O(\log n/\log\log n)$ our construction runs in polynomial time. For the special case $k = 2$, which received a lot of attention in the literature, we construct an $[n, 2]$ Reed-Solomon code over a field of size $O(n^4)$ that can decode from $n-3$ insdel errors. Earlier constructions required an exponential field size. Lastly, we prove that any such construction requires a field of size $\Omega(n^3)$.

[*] Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel. Email: roni.con93@gmail.com

[†] Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel. Email: shpilka@tauex.tau.ac.il. The research leading to these results has received funding from the Israel Science Foundation (grant number 514/20) and from the Len Blavatnik and the Blavatnik Family Foundation.

[‡] Department of EE-Systems, Tel Aviv University, Tel Aviv, Israel. Email: zactamo@gmail.com.

The work of Itzhak Tamo and Roni Con was partially supported by the European Research Council (ERC grant number 852953) and by the Israel Science Foundation (ISF grant number 1030/15).

1 Introduction

Error-correcting codes are among the most widely used tools and objects of study in information theory and theoretical computer science. The most common model of corruption that is studied in the TCS literature is that of errors or erasures. The model in which each symbol of the transmitted word is either replaced with a different symbol from the alphabet (an error) or with a '?' (an erasure). The theory of such codes began with the seminal work of Shannon, [Sha48], who studied random errors and erasures and the work of Hamming [Ham50] who studied the adversarial model for errors and erasures. These models are mostly well understood, and today we know efficiently encodable and decodable codes that are optimal for Shannon's model of random errors. For adversarial errors, we have optimal codes over large alphabets and good codes (codes of constant relative rate and relative distance) for every constant sized alphabet.

Another important model that has been considered ever since Shannon's work is that of *synchronization* errors. These are errors that affect the length of the received word. The most common model for studying synchronization errors is the insertion-deletion model (insdel for short): an insertion error is when a new symbol is inserted between two symbols of the transmitted word. A deletion is when a symbol is removed from the transmitted word. For example, over the binary alphabet, when 100110 is transmitted, we may receive the word 1101100, which is obtained from two insertions (1 at the beginning and 0 at the end) and one deletion (one of the 0's at the beginning of the transmitted word). Observe that compared to the more common error model, if an adversary wishes to *change* a symbol, then the cost is that of two operations - first deleting the symbol and then inserting a new one instead.

Insdel errors appear in diverse settings such as optical recording, semiconductor devices, integrated circuits, and synchronous digital communication networks. Another important example is the trace reconstruction problem, which has applications in computational biology and DNA-based storage systems [BLC$^+$16, YGM17, HMG19]. See the surveys [Mit09, MBT10] for a good picture of the problems and applications of error-correcting codes for the insdel model (insdel codes for short).

Reed-Solomon codes are the most widely used family of codes in theory and practice. Indeed, they have found many applications both in theory and in practice (their applications include QR codes [Soo08], secret sharing schemes [MS81], space transmission [WB99], encoding data on CDs [WB99] and more. The ubiquity of these codes can be attributed to their simplicity as well as to their efficient encoding and decoding algorithms. As such, it is an important problem to understand whether they can also decode from insdel errors. This problem received a lot of attention recently [SNW02, WMSN04, TSN07, DLTX19, LT21, CZ21, LX21], but besides very few constructions (i.e., evaluation points for Reed-Solomon codes), not much was known before our work. We discuss this line of work in more detail in Section 1.2.

In this paper, we first prove that there are Reed-Solomon codes that achieve the half-Singleton bound. In other words, there are optimal Reed-Solomon codes also against insdel errors. We also give a set of evaluation points that define a Reed-Solomon code that achieves this bound. As the field size that we get grows very fast, our construction runs in polynomial time only for very small values of δ. We also explicitly construct 2-dimensional RS codes over a field size smaller than the previous known constructions.

Unfortunately, we could not develop efficient decoding algorithms for our Reed-Solomon constructions, and we leave this as an open problem for future research.

1.1 Basic definitions and notation

For an integer k, we denote $[k] = \{1, 2, \dots, k\}$. Throughout this paper, $\log(x)$ refers to the base-2 logarithm. For a prime power q, we denote with $\mathbb{F}_q$ the field of size q.

We denote the ith symbol of a string s (or of a vector v) as s_i (equivalently v_i). Throughout this paper, we shall move freely between representations of vectors as strings and vice versa. Namely, we shall view each vector $v = (v_1, \dots, v_n) \in \mathbb{F}_q^n$ also as a string by concatenating all the symbols of the vector into one string, i.e., $(v_1, \dots, v_n) \leftrightarrow v_1 \circ v_2 \circ \dots \circ v_n$. Thus, if we say that s is a subsequence of some vector v, we mean that we view v as a string and s is a subsequence of that string.

An error correcting code of block length n over an alphabet Σ is a subset $\mathcal{C} \subseteq \Sigma^n$. The rate of $\mathcal{C}$ is $\frac{\log|\mathcal{C}|}{n \log|\Sigma|}$, which captures the amount of information encoded in every symbol of a codeword. A linear code over a field $\mathbb{F}$ is a linear subspace $\mathcal{C} \subseteq \mathbb{F}^n$. The rate of a linear code $\mathcal{C}$ of block length n is $\mathcal{R} = \dim(\mathcal{C})/n$. Every linear code of dimension k can be described as the image of a linear map, which, abusing notation, we also denote with $\mathcal{C}$, i.e., $\mathcal{C} : \mathbb{F}^k \to \mathbb{F}^n$. Equivalently, a linear code $\mathcal{C}$ can be defined by a *parity check matrix* H such that $x \in \mathcal{C}$ if and only if $Hx = 0$. When $\mathcal{C} \subseteq \mathbb{F}_q^n$ has dimension k we say that it is an $[n, k]_q$ code. The minimal distance of $\mathcal{C}$ with respect to a metric $d(\cdot, \cdot)$ is defined as $\mathrm{dist}_{\mathcal{C}} := \min_{v \neq u \in \mathcal{C}} d(v, u)$. Naturally, we would like the rate to be as large as possible, but there is an inherent tension between the rate of the code and the minimal distance (or the number of errors that a code can decode from). In this work, we focus on codes against insertions and deletions.

Definition 1.1. *Let s be a string over the alphabet Σ. The operation in which we remove a symbol from s is called a* deletion *and the operation in which we place a new symbol from Σ between two consecutive symbols in s, in the beginning, or at the end of s, is called an* insertion*.*

A substring *of s is a string obtained by taking consecutive symbols from s. A* subsequence *of s is a string obtained by removing some (possibly none) of the symbols in s.*

The relevant metric for such codes is the edit-distance that we define next.

Definition 1.2. *Let s, s' be strings over the alphabet Σ. A* longest common subsequence *between s and s', is a subsequence s_{sub} of both s and s', of maximal length. We denote by* $\mathrm{LCS}(s, s')$ *the length of a longest common subsequence.*[1]

The edit distance *between s and s', denoted by $ED(s, s')$, is the minimal number of insertions and deletions needed in order to turn s into s'. One can verify that this measure indeed defines a metric (distance function).*

Lemma 1.3 (See e.g. Lemma 12.1 in [CR03]). *It holds that* $\mathrm{ED}(s, s') = |s| + |s'| - 2\mathrm{LCS}(s, s')$.

[1] Note that a longest common subsequence may not be unique as there can be a number of subsequences of maximal length. For example in the strings $s = (1, 0)$ and $s' = (0, 1)$.

We next define Reed-Solomon codes (RS-codes from now on).

Definition 1.4 (Reed-Solomon codes). *Let $\alpha_1, \alpha_2, \dots, \alpha_n \in \mathbb{F}_q$ be distinct points in a finite field $\mathbb{F}_q$ of order $q \geq n$. For $k \leq n$ the $[n,k]_q$ RS-code defined by the evaluation set $\{\alpha_1, \dots, \alpha_n\}$ is the set of codewords*

$$\{c_f = (f(\alpha_1), \dots, f(\alpha_n)) \mid f \in \mathbb{F}_q[x], \deg f < k\} \ .$$

In words, a codeword of an $[n,k]_q$ RS-code is the evaluation vector of some polynomial of degree less than k at n predetermined distinct points. It is well known (and easy to see) that the rate of $[n,k]_q$ RS-code is k/n and the minimal distance, with respect to the Hamming metric, is $n-k+1$.

1.2 Previous results

Linear codes against worst-case insdel errors were recently studied by Cheng, Guruswami, Haeupler, and Li [CGHL21]. Correcting an error in a preceding work, they proved that there are good linear codes against insdel errors.

Theorem 1.5 (Theorem 4.2 in [CGHL21]). *For any $\delta > 0$ and prime power q, there exists a family of linear codes over $\mathbb{F}_q$ that can correct up to δn insertions and deletions, with rate $(1-\delta)/2 - h(\delta)/\log_2(q)$.*

The proof of Theorem 1.5 uses the probabilistic method, showing that, with high probability, a random linear map generates such code. Complementing their result, they proved that their construction is almost tight. Specifically, they provided the following upper bound, which they call "half-Singleton bound," that holds over any field.

Theorem 1.6 (Half-Singleton bound: Corollary 5.1 in [CGHL21]). *Every linear insdel code which is capable of correcting a δ fraction of deletions has rate at most $(1-\delta)/2+o(1)$.*

The performance of RS-codes against insdel errors was studied much earlier than the recent work of Cheng et al. [CGHL21]. To the best of our knowledge, Safavi-Naini and Wang [SNW02] were the first to study the performance of RS-codes against insdel errors. They gave an algebraic condition that is sufficient for an RS-code to correct from insdel errors, yet they did not provide any construction. In fact, in our work, we consider an almost identical algebraic condition, and by simply using the Schwartz-Zippel-Demillo-Lipton lemma, we prove that there are RS-codes that meet this condition and, in addition, achieve the half-Singleton bound. In particular, RS-codes are optimal for insdel errors (see discussion in Section 2). Wang, McAven, and Safavi-Naini [WMSN04] constructed a $[5,2]$ RS-code capable of correcting a single deletion. Then, in [TSN07], Tonien and Safavi-Naini constructed an $[n,k]$ generalized-RS-codes capable of correcting from $\log_{k+1} n - 1$ insdel errors. Similar to our results, they did not provide an efficient decoding algorithm.

In another line of work Duc, Liu, Tjuawinata, and Xing [DLTX19], Liu and Tjuawinata [LT21], Chen and Zhang [CZ21], and Liu and Xing [LX21] studied the specific case of 2-dimensional RS-codes.

In [DLTX19, LT21], the authors presented constructions of $[n,2]$ RS-codes that for every $\varepsilon > 0$ can correct from $(1-\varepsilon)\cdot n$ insdel errors, for codes of length $n = \mathrm{poly}(1/\varepsilon)$

over fields of size $\Omega(\exp((\log n)^{1/\varepsilon}))$ and $\Omega(\exp(n^{1/\varepsilon}))$, respectively. In [DLTX19, CZ21], the authors present constructions of two-dimensional RS-codes that can correct from $n-3$ insdel errors where the field size is exponential in n. After a draft of this work appeared online, Liu and Xing [LX21] constructed, using different approach than us, a two dimensional RS-codes over that can correct from $n-3$ insdel errors, over a field size $O(n^5)$. Specifically, they prove the following.

Theorem 1.7. *[LX21, Theorem 4.8] Let $n \geq 4$. If $q > \frac{n(n-1)^2(n-2)^2}{4}$, then there is an $[n,2]_q$ RS-code, constructed in polynomial time, that can decode from $n-3$ insdel errors.*

1.3 Our results

First, we prove that there are RS-codes that achieve the half-Singleton bound. Namely, they are optimal linear codes for insdel errors.

Theorem 1.8. *Let k and n be positive integers such that $2k-1 \leq n$. For $q = O(n^{4k-2})$ there exists an $[n,k]_q$ RS-code defined by n distinct evaluation points $\alpha_1, \ldots, \alpha_n \in \mathbb{F}_q$, that can recover from $n-2k+1$ adversarial insdel errors.*

Observe that the constructed code achieves the half Singleton bound: its rate is $\mathcal{R} = k/n = (1-\delta)/2 + o(1)$ and $\delta = (n-2k+1)/n$.

Theorem 1.8 is an existential result and does not give an explicit construction. Using ideas from number theory and algebra, we construct RS-codes that can decode from $n-2k+1$ adversarial insdel errors, in particular, they achieve the half-Singleton bound. Specifically,

Theorem 1.9. *Let k and n be positive integers, where $2k-1 \leq n$. There is a deterministic construction of an $[n,k]_q$ RS-code that can correct from $n-2k+1$ insdel errors where $q = O\left(n^{k^2 \cdot ((2k)!)^2}\right)$. The construction runs in polynomial time for $k = O(\log(n)/\log(\log(n)))$.*

We note that for $k = \omega(\log(n)/\log\log(n))$ the field size is $\exp(n^{\omega(1)})$ and in particular, there is no efficient way to represent arbitrary elements of $\mathbb{F}_q$ in this case.

As discussed before, special attention was given in the literature to the case of two dimensional RS-codes. By using Sidon spaces that were constructed in [RRT17], we explicitly construct a family of $[n,2]_q$ RS-codes that can decode from $n-3$ insdel errors for $q = O(n^4)$. Besides improving on all previous constructions in terms of field size, our construction also requires a smaller field size than the one guaranteed by the randomized argument in Theorem 1.8. Such phenomena, where a deterministic algebraic construction outperforms the parameters obtained by a randomized construction, are scarce in coding theory and combinatorics. Well-known examples are AG codes that outperform the GV-bound [TVZ82] and constructions of extremal graphs with “many” edges that do not contain cycles of length 4, 6 or 10 (see [Con21]).

Theorem 1.10. *For any $n \geq 4$, there exists an explicit $[n,2]_q$ RS-code that can correct from $n-3$ insdel errors, where $q = O(n^4)$.*

We also prove a (very) weak lower bound on the field size.

Proposition 1.11. *Any $[n,k]_q$ RS-code that can correct from $n-2k+1$ worst case insdel errors must satisfy*

$$q \geq \frac{1}{2} \cdot \left(\frac{n}{(2k-1)(k-1)} \right)^{\frac{2k-1}{k-1}} .$$

While for large values of k, this bound is meaningless, it implies that when $k=2$, the field size must be $\Omega(n^3)$. Thus, the construction given in Theorem 1.10 is nearly optimal. The gap between the field size in our construction and the one implied by the lower bound raises an interesting question: what is the minimal field size q for which an optimal $[n,2]_q$ RS-code exists?

1.4 Proof idea

To show that RS-codes can be used against insdel errors, we first prove an algebraic condition that is sufficient for n evaluation points to define an RS-code that can decode from insdel errors. This condition requires that a certain set of $n^{O(k)}$ matrices, determined by the evaluation points, must all have full rank. Then, a simple application of the Schwartz-Zippel-DeMilo-Lipton lemma [Sch80, Zip79, DL78] implies the existence of good evaluation points over fields of size $n^{O(k)}$. To obtain a deterministic construction, we show that by going to much larger field size, one can find evaluation points satisfying the full-rank condition. While the field size needs to be of size roughly $\Omega(n^{k^k})$, we note that, for not too large values of k, it is of exponential size, and in this case, our construction runs in polynomial time. A key ingredient in the analysis of this construction is our use of the 'abc theorem' for polynomials over finite fields [VW03].

For the case of $k=2$, we use a different idea that gives a better field size than the one implied by the probabilistic argument above. We do so by noting that in this case the full-rank condition can be expressed as the requirement that no two different triples of evaluation points (x_1,x_2,x_3) and (y_1,y_2,y_3) satisfy

$$\frac{y_1-y_2}{x_1-x_2} = \frac{y_2-y_3}{x_2-x_3} .$$

This condition is reminiscent of the condition behind the construction of Sidon spaces of [RRT17], and indeed, we build on their construction of Sidon spaces to define good evaluation points in a field of size $O(n^4)$.

1.5 Organization

The paper is organized as follows. In Section 2, we prove Theorem 1.8. In Section 3, we prove Theorem 1.9. Finally, in Section 4, we prove Theorem 1.10 and Proposition 1.11. Section 5 is devoted to conclusion and open questions.

2 Reed-Solomon codes achieving the half-Singleton bound

In this section, we prove our results concerning RS-codes. Specifically, we prove that RS-codes achieve the half-Singleton bound and give some explicit constructions. The proofs will follow by standard analysis of the LCS between any two distinct codewords.

We begin by reformulating the condition on the maximum length of an LCS as an algebraic condition (invertibility of certain matrices). Then we show that an RS-code that satisfies this condition would have the maximum possible edit distance and hence would be able to decode from the maximum number of insdel errors. We remark that a similar approach already appeared in [SNW02, Section 2.2] and we shall repeat some of the details here.

2.1 An algebraic condition

The following proposition is the main result of this section as it provides a sufficient condition for an RS-code to recover from the maximum number of insdel errors. We first make the following definitions: We say that a vector of indices $I \in [n]^s$ is an *increasing* vector if its coordinates are monotonically increasing, i.e., for any $1 \leq i < j \leq s$, $I_i < I_j$, where I_i is the ith coordinate of I. For a codeword c of length n and an increasing vector I, let c_I be the restriction of c to the coordinates with indices in I, i.e., $c_I = (c_{I_1}, \ldots, c_{I_s})$. For two vectors $I, J \in [n]^{2k-1}$ with distinct coordinates we define the following (variant of a) vandermonde matrix of order $(2k-1) \times (2k-1)$ in the formal variables $\mathbf{X} = (X_1, \ldots, X_n)$:

$$V_{I,J}(\mathbf{X}) = \begin{pmatrix} 1 & X_{I_1} & \ldots & X_{I_1}^{k-1} & X_{J_1} & \ldots & X_{J_1}^{k-1} \\ 1 & X_{I_2} & \ldots & X_{I_2}^{k-1} & X_{J_2} & \ldots & X_{J_2}^{k-1} \\ \vdots & \vdots & \ldots & \vdots & \vdots & \ldots & \vdots \\ 1 & X_{I_{2k-1}} & \ldots & X_{I_{2k-1}}^{k-1} & X_{J_{2k-1}} & \ldots & X_{J_{2k-1}}^{k-1} \end{pmatrix} . \tag{1}$$

Proposition 2.1. *Consider the $[n,k]_q$ RS-code defined by an evaluation vector $\alpha = (\alpha_1, \ldots, \alpha_n)$. If for every two increasing vectors $I, J \in [n]^{2k-1}$ that agree on at most $k-1$ coordinates, it holds that $\det(V_{I,J}(\alpha)) \neq 0$, then the code can correct any $n-2k+1$ insdel errors. Moreover, if the code can correct any $n-2k+1$ insdel errors, then the only possible vectors in Kernel$(V_{I,J}(\alpha))$ are of the form $(0, f_1, \ldots, f_{k-1}, -f_1, \ldots, -f_{k-1})$.*

Proof. Assume that the claim does not hold; therefore, there exist two distinct codewords $c \neq c'$ whose LCS is at least $2k-1$, i.e., $c_I = c'_J$ for two increasing vectors $I, J \in [n]^{2k-1}$. Assume further that c and c' are the encodings of the degree $k-1$ polynomials $f = \sum_i f_i x^i$ and $g = \sum_i g_i x^i$, respectively. If $I_\ell = J_\ell$ for at least k coordinates, then for every such ℓ

$$f(\alpha_{I_\ell}) = c_{I_\ell} = c'_{J_\ell} = g(\alpha_{I_\ell}) \; .$$

Hence $f \equiv g$, in contradiction to the fact that $c \neq c'$. Thus, we can assume that I, J agree on at most $k-1$ coordinates. In this case, $V_{I,J}(\alpha)$ is singular, since the vector $(f_0 - g_0, f_1, \ldots, f_{k-1}, -g_1, \ldots, -g_{k-1})^t$ is in its right kernel, which contradicts our assumption. From Lemma 1.3 it follows that the code can correct $n-2k+1$ insdel errors.

To prove the moreover part note that the argument above implies that if the code can correct any $n-2k+1$ insdel errors and $f \neq g$ then the vector $(f_0 - g_0, f_1, \ldots, f_{k-1}, -g_1, \ldots, g_{k-1})$ is not in the kernel. □

In [SNW02] Safavi-Naini and Wang identified (almost) the same condition (see Remark 2.2 below) and used it in their construction of traitor tracing schemes. Interestingly, the later work of [TSN07], which gave a construction of RS-codes capable of decoding from $\log_k(n+1)-1$ insdel errors, did not use this condition. In particular, as far as we know, prior to this work the condition in Proposition 2.1 was not used in order to show the existence of optimal RS-codes.

The following remark explains the difference between Proposition 2.1 and the condition in [SNW02].

Remark 2.2. *The main difference between the condition presented in [SNW02] and ours, is that they considered a $2k \times 2k$ matrix and a generalized RS-code. Given evaluation points $(\alpha_1, \ldots, \alpha_n)$ and a vector with nonzero coordinates $(v_1, \ldots, v_n) \in \mathbb{F}_q^n$, the generalized $[n,k]_q$ RS-code is defined as the set of all vectors $(v_1 \cdot f(\alpha_1), \ldots, v_n \cdot f(\alpha_n))$, such that* $\deg(f) < k$. *The matrix studied in [SNW02] is:*

$$V^v_{I,J}(\mathbf{X}) = \begin{pmatrix} v_{I_1} & v_{I_1}\cdot X_{I_1} & \ldots & v_{I_1}\cdot X_{I_1}^{k-1} & v_{J_1} & v_{J_1}\cdot X_{J_1} & \ldots & v_{J_1}\cdot X_{J_1}^{k-1} \\ v_{I_2} & v_{I_2}\cdot X_{I_2} & \ldots & v_{I_2}\cdot X_{I_2}^{k-1} & v_{J_2} & v_{J_2}\cdot X_{J_2} & \ldots & v_{J_2}\cdot X_{J_2}^{k-1} \\ \vdots & \vdots & \ldots & \ldots & \vdots & \vdots & \ldots & \vdots \\ v_{I_{2k}} & v_{I_{2k}}\cdot X_{I_{2k}} & \ldots & v_{I_{2k}}\cdot X_{I_{2k}}^{k-1} & v_{J_{2k}} & v_{J_{2k}}\cdot X_{J_{2k}} & \ldots & v_{J_{2k}}\cdot X_{J_{2k}}^{k-1} \end{pmatrix}. \quad (2)$$

In our matrix, we saved a coordinate (which leads to optimal codes) as we did not have two columns for the free terms of f and g (as defined in the proof). In contrast, the matrix (2) *has a column for the free term of f (the first) and a column for the free term of g (the column $(v_{J_1}, \ldots, v_{J_{2k}})$). This also leads to the requirement that I and J are of length $2k$ (they can still agree on at most $k-1$ indices).*

2.2 Optimal Reed-Solomon codes exist

In this section, we show that over large enough fields, there exist RS-codes that attain the half-Singleton bound. Specifically, we show that there exist RS-codes that can decode from a δ fraction of insdel errors and have rate $\mathcal{R} = (1-\delta)/2 + o(1)$. For convenience, we repeat the statement of Theorem 1.8.

Theorem 1.8. *Let k and n be positive integers such that $2k-1 \le n$. For $q = O(n^{4k-2})$ there exists an $[n,k]_q$ RS-code defined by n distinct evaluation points $\alpha_1, \ldots, \alpha_n \in \mathbb{F}_q$, that can recover from $n-2k+1$ adversarial insdel errors.*

For a vector I and an element a, we write $a \in I$ if a appears in one of the coordinates of I; otherwise, we write $a \notin I$.

Lemma 2.3. *Let $s \ge 2$ be an integer and $I, J \in [n]^s$ two increasing vectors that do not agree on* any *coordinate, i.e., $I_i \ne J_i$ for all $1 \le i \le s$. Then, there are two distinct indices $i \ne j \in [s]$ such that $I_i \notin J$ and $J_j \notin I$.*

Proof. W.l.o.g. assume that $I_1 < J_1$. Since J is an increasing vector, $I_1 \notin J$. In addition, some coordinate among $\{J_1, \ldots, J_s\}$ does not appear in $\{I_2, \ldots, I_s\}$, and any such coordinate is clearly different from I_1. □

Proposition 2.4. *Let $I, J \in [n]^{2k-1}$ be two increasing vectors that agree on at most $k-1$ coordinates. Then, in the expansion of* $\det(V_{I,J}(\mathbf{X}))$ *as a sum over permutations, there is a monomial that is obtained at exactly one of the $(2k-1)!$ different permutations. In particular, its coefficient is ± 1, depending on the sign of its corresponding permutation. Consequently,* $\det(V_{I,J}(\mathbf{X})) \ne 0$.

Proof. The result will follow by applying induction on k. For $k = 1$, $V_{I,J}(\mathbf{X}) = 1$ and the result follows. For the induction step, assume it holds for $k - 1$, and we prove it for $k \geq 2$. Consider two coordinates i, j, determined as follows. If I and J agree on some coordinate, say j, then we set i to be such that $I_i \notin J$. If they do not agree on any coordinate, then we let i, j be the two coordinates guaranteed by Lemma 2.3.

Next, in the determinant expansion of $V_{I,J}$ as a sum of $(2k-1)!$ monomials, collect all the monomials that are divisible by $X_{I_i}^{k-1} X_{J_j}^{k-1}$, and write them together as

$$X_{I_i}^{k-1} X_{J_j}^{k-1} f(\mathbf{X}),$$

for some polynomial f in the variables $(X_\ell : \ell \in (I \setminus \{I_i\}) \cup (J \setminus \{J_j\}))$. Note that the choice of i and j guarantees that such monomials exist. Observe that any monomial in the determinant expansion of $V_{I,J}$ that is divisible by $X_{I_i}^{k-1} X_{J_j}^{k-1}$ must be obtained by picking the (i, k) and the $(j, 2k-1)$ entries in the matrix (1). Hence, f equals the determinant of the submatrix $V'_{I,J}$ obtained by removing rows i, j and columns $k, 2k-1$ from $V_{I,J}$. Note that $V'_{I,J}$ is a matrix satisfying the conditions of the claim: it is a $(2k-3) \times (2k-3)$ matrix defined by two increasing vectors of length $2k-3$ that agree on at most $k-2$ coordinates. Indeed, i and j were chosen so that by removing them we remove one agreement, if such existed.

Hence, by the induction hypothesis $\det(V'_{I,J})$ has a monomial m (with a ± 1 coefficient) that is uniquely obtained among the $(2k-3)!$ different monomials. Therefore, $X_{I_i}^{k-1} X_{J_j}^{k-1} m$ is a monomial of $X_{I_i}^{k-1} X_{J_j}^{k-1} f$ with a ± 1 coefficient. Since there is no other way to obtain this monomial in the determinant expansion of $V_{I,J}$, this monomial is uniquely obtained in $\det(V_{I,J})$, and the result follows. □

We proceed to prove Theorem 1.8 by a standard application of the Schwartz-Zippel lemma.

Proof of Theorem 1.8. Define

$$F(\mathbf{X}) = \prod_{i<j} (X_i - X_j) \prod_{I,J} \det(V_{I,J}(\mathbf{X})),$$

where the second product runs over all possible pairs of increasing vectors that agree on no more than $k-1$ coordinates. Clearly, by Proposition 2.4, $F(\mathbf{X})$ is a nonzero polynomial in the ring $\mathbb{Z}[\mathbf{X}]$. Next, we make two observations regarding the polynomial F. First, since there are $\binom{n}{2k-1}$ increasing vectors, and the degree of each $\det(V_{I,J}(\mathbf{X}))$ is at most $k(k-1)$, it follows that

$$\deg(F) \leq n^2 + \binom{n}{2k-1}^2 \cdot k(k-1) < n^{4k-2} .$$

Second, as each $\det(V_{I,J}(\mathbf{X}))$ is a nonzero polynomial with nonzero coefficients bounded in absolute values by $(2k-1)!$, the absolute value of any nonzero coefficients of F is at most

$$((2k-1)!)^{\binom{n}{2k-1}^2} \leq ((2k-1)!)^{\frac{n^{4k-2}}{((2k-1)!)^2}} < e^{n^{4k-2}}.$$

We claim that there is a prime q in the range $[n^{4k-2}, 2n^{4k-2}]$ that does not divide at least one of the nonzero coefficients of the polynomial F. Indeed, consider a nonzero coefficient of F, and assume towards a contradiction that it is divisible by all such primes. Then, by the growth rate of the primorial function, the absolute value of the coefficient is $\Omega(e^{n^{4k-2}(1+o(1))})$, in contradiction. Now, it is easy to verify that F is also a nonzero polynomial in $\mathbb{F}_q[\mathbf{X}]$, since the monomial whose nonzero coefficient is not divisible by q does not vanish. Therefore, by the Schwarz-Zippel-Demillo-Lipton lemma, there is an assignment $\alpha = (\alpha_1, \ldots, \alpha_n)$ to $\mathbf{X}$ for which $F(\alpha) \neq 0 \mod q$. This assignment clearly corresponds to n *distinct* evaluation points, which by Proposition 2.1, define an $[n,k]_q$ RS-code that can correct any $n-2k+1$ worst-case insdel errors, as claimed. $\square$

We remark again that Theorem 1.8 merely shows the existence of $[n,k]_q$ RS-codes that can decode from the maximum number of insdel errors over a field of size $q = O\left(n^{4k-2}\right)$. Further, the above argument is a standard union bound over all variable assignments that make the matrix defined in (1) to be singular. This by no means implies that such a large finite field is necessary. For example, a similar union-bound argument that shows the existence of MDS codes would require an exponentially large field for codes with a constant rate. In contrast, it is well-known that MDS codes over linear field size exist (e.g., RS-codes). It would be interesting to explicitly construct codes with the same or even better parameters than the ones given in Theorem 1.8. Unfortunately, we could not construct such codes, and this is left as an open question for further research. Nonetheless, in the next section, we provide a deterministic construction of an RS-code for any admissible n, k, at the expense of a larger field size than the one guaranteed by Theorem 1.8.

3 Deterministic construction for any k

In this section, we give our main construction of an $[n,k]$ RS-code that can correct any $n-2k+1$ insdel errors. Specifically, we prove Theorem 1.9 which is restated for convenience

Theorem 1.9. *Let k and n be positive integers, where $2k-1 \leq n$. There is a deterministic construction of an $[n,k]_q$ RS-code that can correct from $n-2k+1$ insdel errors where $q = O\left(n^{k^2 \cdot ((2k)!)^2}\right)$. The construction runs in polynomial time for $k = O(\log(n)/\log(\log(n)))$.*

Remark 3.1. *The downside of this construction is the field size $q = n^{k^{O(k)}}$, which renders it to run in polynomial time only for $k = O(\log(n)/\log(\log(n)))$. For larger values of k, the representation of each field element requires a super polynomial number of bits.*

The Mason–Stothers theorem [Mas84, Sto81] is a result about polynomials that satisfy a non-trivial linear dependence, which is analogous to the well-known *abc conjecture* in number theory [Mas85, Oes88]. Our main tool is one of the many extensions in the literature to the Mason–Stothers theorem. For stating the theorem we need the following notation: For a polynomial $Y(x) \in \mathbb{F}[x]$ over a field with $\mathrm{char}(\mathbb{F}) = p \neq 0$, denote by $\nu(Y(x))$ the number of distinct roots of $Y(x)$ with multiplicity not divisible by p.

Theorem 3.2 ("Moreover part" of Proposition 5.2 in [VW03]). *Let $m \geq 2$ and $Y_0(x) = Y_1(x) + \ldots + Y_m(x)$ with $Y_j(x) \in \mathbb{F}_p[x]$. Suppose that $\gcd(Y_0(x), \ldots, Y_m(x)) = 1$, and that*

$Y_1(x), \ldots, Y_m(x)$ are linearly independent over $\mathbb{F}_p(x^p)$.[2] *Then,*

$$\deg(Y_0(x)) \leq -\binom{m}{2} + (m-1)\sum_{j=0}^{m} \nu(Y_j(x)) \ .$$

Construction 3.3. *Let k be a positive integer and set $\ell = ((2k)!)^2$. Fix a finite field $\mathbb{F}_p$ for a prime $p > k^2 \cdot \ell$ and let n be an integer such that $2k-1 < n \leq p$. Let $\mathbb{F}_q$ be a field extension of $\mathbb{F}_p$ of degree $k^2 \cdot \ell$ and let $\gamma \in \mathbb{F}_q$ be such that $\mathbb{F}_q = \mathbb{F}_p(\gamma)$. Hence, each element of $\mathbb{F}_q$ can be represented as a polynomial in γ, of degree less than $k^2\ell$, over $\mathbb{F}_p$. Define the $[n,k]_q$ RS-code by setting $\alpha_i := (\gamma - i)^\ell$ for $1 \leq i \leq n$.*

Proposition 3.4. *The $[n,k]_q$ RS-code defined in Construction 3.3 can correct any $n - 2k+1$ worst case insdel errors.*

Proof. Let $I, J \in [n]^{2k-1}$ be two increasing vectors that agree on at most $k-1$ coordinates. By Proposition 2.1 it is enough to show that $V_{I,J}(\alpha)$ is non-singular, for every such I, J. By the Leibniz formula, $\det(V_{I,J}(\alpha))$ is a sum of $(2k-1)!$ terms corresponding to the different permutations. Denote these terms as $P_i(\gamma)$ for $i = 0, \ldots, (2k-1)! - 1$. Each of the terms is a product of the sign of the corresponding permutation with some $2k-1$ elements of the form $(\gamma - s)^{\ell \cdot j}$, for some $s \in I \cup J$ and $0 \leq j \leq k-1$. Assume towards a contradiction that $\det(V_{I,J}(\alpha)) = 0$ in $\mathbb{F}_q$, i.e.,

$$\det(V_{I,J}(\alpha)) = P_0(\gamma) + \ldots + P_{(2k-1)!-1}(\gamma) = 0 \ , \tag{3}$$

in $\mathbb{F}_q$. By viewing every term in (3) as a univariate polynomial in γ over $\mathbb{F}_p$, one can verify that, for any j, $\deg(P_j) = \ell \cdot k(k-1) < k^2\ell$. As $\mathbb{F}_q$ is an extension of $\mathbb{F}_p$ of degree $k^2\ell$, it follows that (3) holds also in $\mathbb{F}_p[\gamma]$, the ring of polynomials in the variable γ over $\mathbb{F}_p$. By Proposition 2.4 the determinant of the variable matrix (1) has a monomial that is uniquely obtained and therefore has a ±1 coefficient. Assume, without loss of generality, that P_0 is the image of this monomial under the mapping defined by the assignment $X_i \mapsto (\gamma - i)^\ell$. Note that since this mapping is injective on the set of monomials, no other monomial is mapped to a scalar multiple of P_0. In other words, P_0 and P_i are linearly independent for any $i \geq 1$. Assume further that (without loss of generality) $P_1, \ldots, P_m$ is a minimal subset among $\{P_i\}_{i \geq 1}$ that spans P_0 over $\mathbb{F}_p$. The existence of such a set follows from (3). Hence, we can write

$$P_0 = \sum_{i=1}^{m} a_i P_i, \text{ where } a_i \in \mathbb{F}_p \backslash \{0\}. \tag{4}$$

Clearly, by minimality, $P_1, \ldots, P_m$ are linearly independent over $\mathbb{F}_p$. Further, $m \geq 2$, since otherwise there would be an $i > 0$ such that P_i is a multiple of P_0.

Since the P_i's are of degree $\ell k(k-1)$, and P_0 was obtained from a unique monomial in the determinant expansion, it follows that the greatest common divisor $Q := \gcd(P_0, \ldots, P_m)$ has degree at most $\ell(k(k-1)-1)$. By dividing (4) by Q we have

$$\overline{P_0} = \sum_{i=1}^{m} a_i \overline{P_i}, \tag{5}$$

[2] $\mathbb{F}_p(x^p)$ is the field of rational functions in x^p. Namely, its elements are $f(x^p)/g(x^p)$ where $f(x), g(x) \in \mathbb{F}_p[x]$ and $g(x) \not\equiv 0$.

where $\overline{P_i} = P_i/Q$. We will need the following claim, whose proof is deferred to the end of this section.

Claim 3.5. *The polynomials $\overline{P_1}, \ldots, \overline{P_m}$ are linearly independent over $\mathbb{F}_p(\gamma^p)$.*

The contradiction will follow by invoking Theorem 3.2. Towards this end note that (i) By Claim 3.5 the polynomials $\overline{P_1}, \ldots, \overline{P_m}$ are linearly independent of $\mathbb{F}_p(\gamma^p)$ (ii) $\gcd(\overline{P_0}, \ldots, \overline{P_m}) = 1$, and (iii) $\nu(\overline{P_j}) \leq 2k - 2$, as P_j is the multiplication of $2k - 2$ non-constant polynomials, each having a single root. Thus, by (5) and Theorem 3.2

$$\begin{aligned} \ell \leq \deg(P_0) - \deg(Q) = \deg(\overline{P_0}) &\leq -\binom{m}{2} + (m-1) \cdot \sum_{i=1}^{m} \nu(\overline{P_j}) \\ &< m^2(2k-2) \\ &\leq ((2k-1)!)^2 \cdot (2k-2) \ , \end{aligned}$$

which is a contradiction by the choice of ℓ. This completes the proof. □

It remains to prove Claim 3.5.

Proof of Claim 3.5. Assume towards a contradiction that there exist $\lambda_1, \ldots, \lambda_m \in \mathbb{F}_p(\gamma^p)$ not all zero, such that

$$\sum_{j=1}^{m} \lambda_j(\gamma^p)\overline{P_j}(\gamma) = 0 \ . \tag{6}$$

By clearing the denominators of the λ_j's and any common factor they might have, we can assume that the λ_j's are polynomials in the variable γ^p with no common factors. Since $\deg(\overline{P_j}) \leq \deg(P_j) < p$, we get by reducing (6) modulo γ^p that

$$\sum_{j=1}^{m} \lambda_j(0)\overline{P_j}(\gamma) = 0 \ .$$

Note that $\lambda_j(0) \neq 0$ for some j, since otherwise γ^p would be a common factor of the λ_i's. Hence, $\overline{P_1}, \ldots, \overline{P_m}$ are linearly dependent over $\mathbb{F}_p$, which contradicts the fact that $P_1, \ldots, P_m$ are linearly independent over $\mathbb{F}_p$. □

By setting $n = p$ in Construction 3.3 it follows that the field size of Construction 3.3 is roughly $n^{k^{O(k)}}$ which is much worse than the field size guaranteed by the existential result in Theorem 1.8. Note, however, that the construction runs in polynomial time for RS-codes with dimension $O(\log(n)/\log(\log(n)))$. The proof of Theorem 1.9 immediately follows.

4 Explicit construction for $k = 2$ with quartic field size

In this section we prove Theorem 1.10, which is restated for convenience.

Theorem 1.10. *For any $n \geq 4$, there exists an explicit $[n, 2]_q$ RS-code that can correct from $n - 3$ insdel errors, where $q = O(n^4)$.*

The proof of Theorem 1.10 requires the notion of Sidon spaces, which were introduced in a work of Bachoc, Serra and Zémor [BSZ17] in the study of an analogue of Vosper's theorem for finite fields. Later, Roth, Raviv and Tamo gave an explicit construction of Sidon spaces and used it to provide a construction of cyclic subspace codes [RRT17]. Our construction relies on the construction of Sidon spaces of [RRT17], which was also recently used in [RLT21] to construct a public-key cryptosystem. We believe that Sidon spaces in general, and specifically the construction of [RRT17], might find more applications in coding theory and cryptography in the future. We begin with a formal definition of a Sidon space.

Definition 4.1. *An $\mathbb{F}_q$ linear subspace $S \subseteq \mathbb{F}_{q^n}$ is called a* Sidon space *if for any nonzero elements $a, b, c, d \in S$ such that $ab = cd$, it holds that that*

$$\{a\mathbb{F}_q, b\mathbb{F}_q\} = \{c\mathbb{F}_q, d\mathbb{F}_q\},$$

where $x\mathbb{F}_q = \{x \cdot \alpha : \alpha \in \mathbb{F}_q\}$.

A Sidon space S has the following interesting property, from which it draws its name: Given the product $a \cdot b$ of two nonzero elements $a, b \in S$, one can uniquely factor it to its two factors a, b from S, up to a multiplication by a scalar from the base field. Clearly, this is the best one can hope for, since for any nonzero $\alpha \in \mathbb{F}_q$ the product of the elements $\alpha \cdot a, b/\alpha \in S$ also equals $a \cdot b$. A Sidon space can be viewed as a multiplicative analogue to the well-known notion of *Sidon sets*, which is a common object of study in combinatorics, see e.g. [ET41].

We proceed to present the construction of a Sidon space given in [RRT17].

Theorem 4.2 (Construction 15, Theorem 16 in [RRT17])**.** *Let $q \geq 3$ be a prime power, $m \in \mathbb{N}$, and $n = 2m$. Then, there exists an explicit $\gamma \in \mathbb{F}_{q^n}$ such that $S = \{u + u^q \cdot \gamma \mid u \in \mathbb{F}_{q^m}\}$ is an m-dimensional Sidon space over $\mathbb{F}_q$.*

Another component in our construction is the "long" ternary code with minimum distance of at least 5, given in [GS86]. We note that we could also use the codes given in [DD08].

Theorem 4.3. *[GS86] For every $m \geq 1$, there exits an explicit $[(3^m+1)/2, (3^m+1)/2 - 2m]_3$ linear code with minimum distance at least 5.*

We next combine the above two algebraic objects and construct an RS-code with the desired properties.

Construction 4.4. *For $q = 3$ and $m \in \mathbb{N}$. Let $S \subset \mathbb{F}_{3^{4m}}$ be a $2m$-dimensional Sidon space over $\mathbb{F}_3$ as guaranteed by Theorem 4.2. Let $s_1, \dots, s_{2m}$ be a basis of S. Let $H = (h_{i,j})$ be a $(2m) \times ((3^m + 1)/2)$ parity check matrix of the code given in Theorem 4.3. Our $[n, 2]_{3^{4m}}$ RS-code of length $n = (3^m + 1)/2$ is defined by the evaluation points*

$$\alpha_j = \sum_{i=1}^{2m} s_i h_{i,j} \text{ for } 1 \leq j \leq (3^m + 1)/2 .$$

In other words, we can think of our evaluation points as the n coordinates of the vector $\alpha = (s_1, \dots, s_{2m}) \cdot H$.

The following property of the evaluation points α_j follows easily from Theorem 4.3.

Lemma 4.5. *Any four distinct α_j's are linearly independent over $\mathbb{F}_3$.*

Proof. Consider four distinct α_j's, say $\alpha_1, \alpha_2, \alpha_3, \alpha_4$, and assume towards a contradiction that there exist $\beta_1, \ldots, \beta_4 \in \mathbb{F}_3$ not all zero, such that $\sum_{i=1}^{4} \beta_i \alpha_i = 0$. Then

$$0 = \sum_{i=1}^{4} \beta_i \alpha_i = \sum_{i=1}^{4} \beta_i \sum_{j=1}^{2m} s_j h_{j,i} = \sum_{j=1}^{2m} s_j \sum_{i=1}^{4} \beta_i h_{j,i} \ .$$

Since the s_j's are linearly independent over $\mathbb{F}_3$ it follows that $\sum_i \beta_i h_{j,i} = 0$ for every $j = 1, \ldots, 2m$. Hence, the four columns h_1, h_2, h_3, h_4 of H are linearly dependent over $\mathbb{F}_3$, which contradicts the fact that the minimum distance of the code checked by H is at least 5. □

We proceed to prove that the constructed RS-code can decode from the maximum number of insdel errors.

Theorem 4.6. *The $[n, 2]_{3^{4m}}$ RS-code given in Construction 4.4 can correct any $n - 3$ worst case insdel errors.*

Proof. Assume towards a contradiction that this is not the case. Proposition 2.1 implies that there must exist two triples of distinct evaluation points $(x_1, x_2, x_3), (y_1, y_2, y_3)$, that agree on at most one coordinate, such that

$$\left| \begin{pmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{pmatrix} \right| = 0 \ .$$

Equivalently, $(y_1 - y_2)(x_2 - x_3) = (y_2 - y_3)(x_1 - x_2)$. Since the x_i's are distinct elements of the Sidon space S, $x_2 - x_3$ and $x_1 - x_2$ are *nonzero* elements in S. Similarly, $y_1 - y_2$ and $y_2 - y_3$ are nonzero elements in S. By definition of Sidon spaces, there exists a nonzero $\lambda \in \mathbb{F}_3$ such that

$$\lambda(y_1 - y_2) = y_2 - y_3 \text{ or } \lambda(y_1 - y_2) = x_1 - x_2,$$

which contradicts Lemma 4.5. Indeed, each of the equations implies a nontrivial linear dependence over $\mathbb{F}_3$ between at least three and at most four evaluation points (here we used the facts that the elements of each triple are distinct and that the two triples agree on at most one coordinate). □

We conclude this section with the proof of Theorem 1.10.

Proof of Theorem 1.10. By Theorem 4.6, the code given in Construction 4.4 is an RS-code of length $n = (3^m + 1)/2$, defined over the field $\mathbb{F}_{3^{4m}}$, which is of order $O(n^4)$, as claimed. □

4.1 A lower bound on the field size

In Section 2.2 we proved the existence of optimal $[n,k]_q$ RS-codes for worst-case insdel errors over fields of size $q = n^{O(k)}$. This section complements this result by providing a lower bound on the field size for such codes. Specifically, we ask how large must q be in an $[n,k]_q$ RS-code that can correct from $n-2k+1$ worst-case insdel errors. We prove the following.

Proposition 1.11. *Any $[n,k]_q$ RS-code that can correct from $n-2k+1$ worst case insdel errors must satisfy*

$$q \geq \frac{1}{2} \cdot \left(\frac{n}{(2k-1)(k-1)} \right)^{\frac{2k-1}{k-1}} .$$

Proof. Consider an $[n,k]_q$ RS-code, defined by evaluation points $\alpha_1, \ldots, \alpha_n$, that can correct any $n-2k+1$ insdel errors. For a *non-constant* polynomial f of degree less than k let $\mathcal{V}_f$ be the set of all subsequences of the codeword corresponding to f, of length $2k-1$:

$$\mathcal{V}_f = \{ (f(\alpha_{i_1}), \ldots, f(\alpha_{i_{2k-1}})) : 1 \leq i_1 < \ldots < i_{2k-1} \leq n \} \subseteq \mathbb{F}_q^{2k-1}.$$

By Lemma 1.3, since the code can decode from any $n-2k+1$ insdel errors, the sets $\mathcal{V}_f$ and $\mathcal{V}_g$ for two distinct polynomials f, g, are disjoint. Therefore,[3]

$$\sum_{1 \leq \deg(f) < k} |\mathcal{V}_f| \leq q^{2k-1}. \tag{7}$$

Next, we provide a lower bound on the size of $\mathcal{V}_f$. For any non-constant polynomial f, of degree less than k, and any $a \in \mathbb{F}_q$ there are at most $k-1$ indices i such that $f(\alpha_i) = a$. Thus, for a fixed vector $(a_1, \ldots, a_{2k-1}) \in \mathcal{V}_f$ there are at most $(k-1)^{2k-1}$ increasing vectors of indices $(i_1, \ldots, i_{2k-1})$ such that

$$(f(\alpha_{i_1}), \ldots, f(\alpha_{i_{2k-1}})) = (a_1, \ldots, a_{2k-1}).$$

Therefore $|\mathcal{V}_f| \geq \binom{n}{2k-1}(k-1)^{-(2k-1)}$. Combined with (7) we have

$$\left(\frac{1}{k-1} \right)^{2k-1} \cdot \binom{n}{2k-1} \cdot \left(q^k - q \right) \leq q^{2k-1} ,$$

By rearranging and the fact that $q^{2k-1}/(q^k - q) \leq 2q^{k-1}$ for $q, k \geq 2$, we have

$$\left(\frac{1}{2} \right)^{\frac{1}{k-1}} \left(\frac{n}{(2k-1)(k-1)} \right)^{\frac{2k-1}{k-1}} \leq q . \qquad \square$$

As one can easily verify, this bound is rather weak, as it provides an improvement over the trivial lower bound of $q \geq n$ only for the vanishing rate regime of $k = O(n^{1/4})$. For codes of dimension 2, the bound implies $q = \Omega(n^3)$, and it slowly degrades as one increases k. Nevertheless, it is always at least $\Omega(n^2)$ for any constant k. It is interesting to note that by combining Proposition 1.11 and Theorem 4.6 it follows that an $[n,2]_q$ RS-code that can decode from $n-3$ insdel errors requires that $\Omega(n^3) \leq q \leq O(n^4)$. Determining the minimum possible value of q for this case is an interesting open problem.

[3] This equation remains true also if we include the constant polynomials.

5 Open questions

This paper studies the performence of RS codes against insdel errors. We showed that there are RS-codes are optimal against insdel errors, i.e., they achieve the half-Singleton bound. We also construct explicit RS codes that achieve this bound and as far as we know, this is the first linear code that is shown to achieve this bound.

As discussed, Construction 3.3 is not optimal in terms of the field size. It is a fascinating open question to find an RS-code with an optimal field size. Specifically, the challenge is to construct an RS-code that can correct from any $n - 2k + 1$ insdel errors, over a field of size $O(n^{O(k)})$ (Theorem 1.8 proves the existence of such codes).

The lower bound on the field size proved in Proposition 1.11 is far from giving a full picture of the tradeoff between dimension and field size. The natural open question is to significantly improve our lower bound or provide a better upper bound.

Finally, another interesting question is to provide an efficient decoding algorithm for our constructions of RS-codes.

Acknowledgement

We thank Shu Liu, Ivan Tjuawinata and Chaoping Xing for spotting an inaccuracy in the original statement of Proposition 2.1.

References

[BLC$^+$16] James Bornholt, Randolph Lopez, Douglas M Carmean, Luis Ceze, Georg Seelig, and Karin Strauss. A DNA-based archival storage system. *ACM SIGARCH Computer Architecture News*, 44(2):637–649, 2016.

[BSZ17] Christine Bachoc, Oriol Serra, and Gilles Zémor. An analogue of Vosper's theorem for extension fields. *Mathematical Proceedings of the Cambridge Philosophical Society*, 163(3):423–452, 2017.

[CGHL21] Kuan Cheng, Venkatesan Guruswami, Bernhard Haeupler, and Xin Li. Efficient linear and affine codes for correcting insertions/deletions. In Dániel Marx, editor, *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms, SODA 2021, Virtual Conference, January 10 - 13, 2021*, pages 1–20. SIAM, 2021.

[Con21] David Conlon. Extremal numbers of cycles revisited. *The American Mathematical Monthly*, 128(5):464–466, 2021.

[CR03] Maxime Crochemore and Wojciech Rytter. *Jewels of stringology: text algorithms.* World Scientific, 2003.

[CZ21] Bocong Chen and Guanghui Zhang. Improved Singleton bound on insertion-deletion codes and optimal constructions. *arXiv preprint arXiv:2105.02004*, 2021.

[DD08] Danyo Danev and Stefan Dodunekov. A family of ternary quasi-perfect BCH codes. *Designs, Codes and Cryptography*, 49(1-3):265–271, 2008.

[DL78] Richard A. DeMillo and Richard J. Lipton. A probabilistic remark on algebraic program testing. *Inf. Process. Lett.*, 7(4):193–195, 1978.

[DLTX19] Tai Do Duc, Shu Liu, Ivan Tjuawinata, and Chaoping Xing. Explicit Constructions of Two-Dimensional Reed-Solomon Codes in High Insertion and Deletion Noise Regime. *arXiv preprint arXiv:1909.03426*, 2019.

[ET41] Paul Erdös and Pál Turán. On a problem of Sidon in additive number theory, and on some related problems. *Journal of the London Mathematical Society*, 1(4):212–215, 1941.

[GS86] Igor Borisovich Gashkov and Vladimir Michilovich Sidel'nikov. Linear ternary quasi-perfect codes correcting double errors. *Problemy Peredachi Informatsii*, 22(4):43–48, 1986.

[Ham50] Richard W. Hamming. Error detecting and error correcting codes. *Bell System technical journal*, 29(2):147–160, 1950.

[HMG19] Reinhard Heckel, Gediminas Mikutis, and Robert N Grass. A characterization of the DNA data storage channel. *Scientific reports*, 9(1):1–12, 2019.

[LT21] Shu Liu and Ivan Tjuawinata. On 2-dimensional insertion-deletion Reed-Solomon codes with optimal asymptotic error-correcting capability. *Finite Fields and Their Applications*, 73:101841, 2021.

[LX21] Shu Liu and CHAOPING XING. Bounds and constructions for insertion and deletion codes. *arXiv preprint arXiv:2111.14026*, 2021.

[Mas84] Richard C. Mason. *Diophantine Equations over Function Fields*. London Mathematical Society Lecture Note Series. Cambridge University Press, 1984.

[Mas85] David W. Masser. Open problems. In *Chen, W.W.L. (ed.). Proceedings of the Symposium on Analytic Number Theory. Imperial College, London*, 1985.

[MBT10] Hugues Mercier, Vijay K Bhargava, and Vahid Tarokh. A survey of error-correcting codes for channels with symbol synchronization errors. *IEEE Communications Surveys & Tutorials*, 12(1):87–96, 2010.

[Mit09] Michael Mitzenmacher. A survey of results for deletion channels and related synchronization channels. *Probability Surveys*, 6:1–33, 2009.

[MS81] Robert J. McEliece and Dilip V. Sarwate. On sharing secrets and Reed-Solomon codes. *Communications of the ACM*, 24(9):583–584, 1981.

[Oes88] Joseph Oesterlé. Nouvelles approches du "théoreme" de Fermat. *Astérisque*, 161(162):165–186, 1988.

[RLT21] Netanel Raviv, Ben Langton, and Itzhak Tamo. Multivariate Public Key Cryptosystem from Sidon Spaces. In Juan A. Garay, editor, *Public-Key Cryptography - PKC 2021 - 24th IACR International Conference on Practice and Theory of Public Key Cryptography, Virtual Event, May 10-13, 2021, Proceedings, Part I*, volume 12710 of *Lecture Notes in Computer Science*, pages 242–265. Springer, 2021.

[RRT17] Ron M. Roth, Netanel Raviv, and Itzhak Tamo. Construction of Sidon spaces with applications to coding. *IEEE Transactions on Information Theory*, 64(6):4412–4422, 2017.

[Sch80] Jacob T. Schwartz. Fast probabilistic algorithms for verification of polynomial identities. *J. ACM*, 27(4):701–717, 1980.

[Sha48] Claude Elwood Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

[SNW02] Reihaneh Safavi-Naini and Yejing Wang. Traitor tracing for shortened and corrupted fingerprints. In *ACM workshop on Digital Rights Management*, pages 81–100. Springer, 2002.

[Soo08] Tan Jin Soon. QR code. *Synthesis Journal*, 2008:59–78, 2008.

[Sto81] Walter W. Stothers. Polynomial identities and Hauptmoduln. *The Quarterly Journal of Mathematics*, 32(3):349–370, 1981.

[TSN07] Dongvu Tonien and Reihaneh Safavi-Naini. Construction of deletion correcting codes using generalized Reed–Solomon codes and their subcodes. *Designs, Codes and Cryptography*, 42(2):227–237, 2007.

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

[VW03] Leonid N Vaserstein and Ethel R Wheland. Vanishing polynomial sums. *Communications in Algebra*, 31(2):751–772, 2003.

[WB99] Stephen B Wicker and Vijay K Bhargava. *Reed-Solomon codes and their applications.* John Wiley & Sons, 1999.

[WMSN04] Yejing Wang, Luke McAven, and Reihaneh Safavi-Naini. Deletion correcting using generalized Reed-Solomon codes. In *Coding, Cryptography and Combinatorics*, pages 345–358. Springer, 2004.

[YGM17] S.M. Hossein Tabatabaei Yazdi, Ryan Gabrys, and Olgica Milenkovic. Portable and error-free DNA-based data storage. *Scientific reports*, 7(1):1–6, 2017.

[Zip79] Richard Zippel. Probabilistic algorithms for sparse polynomials. In *EUROSAM*, pages 216–226, 1979.