Efficient and Error-Correcting Data Structures for Membership and Polynomial Evaluation

Victor Chen* Elena Grigorescu† Ronald de Wolf‡

Abstract

We construct efficient data structures that are resilient against a constant fraction of adversarial noise. Our model requires that the decoder answers *most* queries correctly with high probability and for the remaining queries, the decoder with high probability either answers correctly or declares "don't know." Furthermore, if there is no noise on the data structure, it answers *all* queries correctly with high probability. Our model is the common generalization of an error-correcting data structure model proposed recently by de Wolf, and the notion of "relaxed locally decodable codes" developed in the PCP literature.

We measure the efficiency of a data structure in terms of its *length*, (the number of bits in its representation), and query-answering time, measured by the number of *bit-probes* to the (possibly corrupted) representation. We obtain results for the following two data structure problems:

- (Membership) Store a subset S of size at most s from a universe of size n such that membership queries can be answered efficiently, i.e., decide if a given element from the universe is in S.
  We construct an error-correcting data structure for this problem with length nearly linear in $s \log n$ that answers membership queries with $O(1)$ bit-probes. This nearly matches the asymptotically optimal parameters for the noiseless case: length $O(s \log n)$ and one bit-probe, due to Buhrman, Miltersen, Radhakrishnan, and Venkatesh.
- (Univariate polynomial evaluation) Store a univariate polynomial g of degree $\deg(g) \leq s$ over the integers modulo n such that evaluation queries can be answered efficiently, i.e., evaluate the output of g on a given integer modulo n.
  We construct an error-correcting data structure for this problem with length nearly linear in $s \log n$ that answers evaluation queries with $\mathrm{polylog}\, s \cdot \log^{1+o(1)} n$ bit-probes. This nearly matches the parameters of the best-known noiseless construction, due to Kedlaya and Umans.

*MIT CSAIL, victor@csail.mit.edu. Supported by NSF award CCF-0829672.

†MIT CSAIL, elena_g@mit.edu. This work started when this author was visiting CWI in Summer 2008. Supported by NSF award CCF-0829672.

‡CWI Amsterdam, rdewolf@cwi.nl. Supported by a Vidi grant from the Netherlands Organization for Scientific Research (NWO).

1 Introduction

The area of data structures is one of the oldest and most fundamental parts of computer science, in theory as well as in practice. The underlying question is a time-space tradeoff: we are given a piece of data, and we would like to store it in a short, space-efficient data structure that allows us to quickly answer specific queries about the stored data. On one extreme, we can store the data as just a list of the correct answers to all possible queries. This is extremely time-efficient (one can immediately look up the correct answer without doing any computation) but usually takes significantly more space than the information-theoretic minimum. At the other extreme, we can store a maximally compressed version of the data. This method is extremely space-efficient but not very time-efficient since one usually has to undo the whole compression first. A good data structure sits somewhere in the middle: it does not use much more space than the information-theoretic minimum, but it also stores the data in a structured way that enables efficient query-answering.

It is reasonable to assume that most practical implementations of data storage are susceptible to *noise*: over time some of the information in the data structure may be corrupted or erased by various accidental or malicious causes. This buildup of errors may cause the data structure to deteriorate so that most queries are not answered correctly anymore. Accordingly, it is a natural task to design data structures that are not only efficient in space and time but also resilient against a certain amount of *adversarial* noise, where the noise can be placed in positions that make decoding as difficult as possible.

Ways to protect information and computation against noise have been well studied in the theory of error-correcting codes and of fault-tolerant computation. In the data structure literature, constructions under often incomparable models have been designed to cope with noise, and we examine a few of these models. Aumann and Bender [2] studied pointer-based data structures such as linked lists, stacks, and binary search trees. In this model, errors (adversarial but detectable) occur whenever all the pointers from a node are lost. They measure the dependency between the number of errors and the number of nodes that become irretrievable, and designed a number of efficient data structures where this dependency is reasonable.

Another model for studying data structures with noise is the faulty-memory RAM model, introduced by Finocchi and Italiano [10]. In a faulty-memory RAM, there are $O(1)$ memory cells that cannot be corrupted by noise. Elsewhere, errors (adversarial and undetectable) may occur at any time, even during the decoding procedure. Many data structure problems have been examined in this model, such as sorting [8], searching [9], priority queues [13] and dictionaries [5]. However, the number of errors that can be tolerated is typically less than a linear portion of the size of the input. Furthermore, correctness can only be guaranteed for keys that are not affected by noise. For instance, for the problem of comparison-sorting on n keys, the authors of [8] designed a resilient sorting algorithm that tolerates $\sqrt{n \log n}$ keys being corrupted and ensures that the set of uncorrupted keys remains sorted.

Recently, de Wolf [20] considered another model of resilient data structures. The representation of the data structure is viewed as a bit-string, from which a decoding procedure can read any particular set of bits to answer a data query. The representation must be able to tolerate a constant fraction δ of adversarial noise in the bit-string[1] (but not inside the decoding procedure). His model generalizes the usual noise-free data structures (where $\delta = 0$) as well as the so-called "locally decodable codes" (LDCs) [14]. Informally, an LDC is an encoding that is tolerant of noise and allows fast decoding so that each message symbol can be retrieved correctly with high probability. Using LDCs as building blocks, de Wolf constructed data structures for several problems.

Unfortunately, de Wolf's model has the drawback that the optimal time-space tradeoffs are much worse than in the noise-free model. The reason is that all known constructions of LDCs that make $O(1)$ bit-probes [22, 7] have very poor encoding length (super-polynomial in the message length). In fact, the en-

[1] We only consider bit-flip-errors here, not erasures. Since erasures are easier to deal with than bit-flips, it suffices to design a data structure dealing with bit-flip-errors.

coding length provably must be super-linear in the message length [14, 16, 21]. As his model is a generalization of LDCs, data structures cannot have a succinct representation that has length proportional to the information-theoretic bound.

We thus ask: what is a clean model of data structures that allows efficient representations *and* has error-correcting capabilities? Compared with the pointer-based model and the faulty-memory RAM, de Wolf's model imposes a rather stringent requirement on decoding: *every* query must be answered correctly with high probability from the possibly corrupted encoding. While this requirement is crucial in the definition of LDCs due to their connection to complexity theory and cryptography, for data structures it seems somewhat restrictive.

In this paper, we consider a broader, more relaxed notion of error-correcting for data structures. In our model, for most queries, the decoder has to return the correct answer with high probability. However, for the few remaining queries, the decoder may claim ignorance, i.e., declare the data item unrecoverable from the (corrupted) data structure. Still, for *every* query, the answer is incorrect only with small probability. In fact, just as de Wolf's model is a generalization of LDCs, our model in this paper is a generalization of the "relaxed" locally decodable codes (RLDCs) introduced by Ben-Sasson, Goldreich, Harsha, Sudan, and Vadhan [4]. They relax the usual definition of an LDC by requiring the decoder to return the correct answer on *most* rather than all queries. For the remaining queries it is allowed to claim ignorance, i.e., to output a special symbol '$\perp$' interpreted as "don't know" or "unrecoverable." As shown in [4], relaxing the LDC-definition like this allows for constructions of RLDCs with $O(1)$ bit-probes of *nearly linear* length.

Using RLDCs as building blocks, we construct error-correcting data structures that are very efficient in terms of time as well as space. Before we describe our results, let us define our model formally. First, a *data structure problem* is specified by a set D of *data items*, a set Q of *queries*, a set A of *answers*, and a function $f : D \times Q \to A$ which specifies the correct answer $f(x, q)$ of query q to data item x. A data structure for f is specified by four parameters: t the number bit-probes, δ the fraction of noise, ε an upper bound on the error probability for each query, and λ an upper bound on the fraction of queries in Q that are not answered correctly with high probability (the 'λ' stands for "lost").

Definition 1. Let $f : D \times Q \to A$ be a data structure problem. Let $t > 0$ be an integer, $\delta \in [0, 1]$, $\varepsilon \in [0, 1/2]$, and $\lambda \in [0, 1]$. We say that f has a $(t, \delta, \varepsilon, \lambda)$-*data structure* of length N if there exist an encoder $\mathcal{E} : D \to \{0, 1\}^N$ and a (randomized) decoder $\mathcal{D}$ with the following properties: for every $x \in D$ and every $w \in \{0, 1\}^N$ at Hamming distance $\Delta(w, \mathcal{E}(x)) \leq \delta N$,

1. $\mathcal{D}$ makes at most t bit-probes to w,
2. $\Pr[\mathcal{D}^w(q) \in \{f(x, q), \perp\}] \geq 1 - \varepsilon$ for every $q \in Q$,
3. the set $G = \{q : \Pr[\mathcal{D}^w(q) = f(x, q)] \geq 1 - \varepsilon\}$ has size at least $(1 - \lambda)|Q|$ ('G' stands for "good"),
4. if $w = \mathcal{E}(x)$, then $G = Q$.

Here $\mathcal{D}^w(q)$ denotes the random variable which is the decoder's output on inputs w and q. The notation indicates that it accesses the two inputs in different ways: while it has full access to the query q, it only has bit-probe access (or "oracle access") to the string w.

We say that a $(t, \delta, \varepsilon, \lambda)$-data structure is *error-correcting*, or an *error-correcting data structure*, if $\delta > 0$. Setting $\lambda = 0$ recovers the original notion of error-correction in de Wolf's model [20]. A $(t, \delta, \varepsilon, \lambda)$-*relaxed locally decodable code (RLDC)*, defined in [4], is an error-correcting data structure for the membership function $f : \{0, 1\}^n \times [n] \to \{0, 1\}$, where $f(x, i) = x_i$. A (t, δ, ε)-*locally decodable code (LDC)*, defined by Katz and Trevisan [14], is an RLDC with $\lambda = 0$.

Remark. For the data structure problems considered in this paper, our decoding procedures make only *non-adaptive* probes, i.e., the positions of the probes are determined all at once and sent simultaneously to the oracle. For other data structure problems it may be natural for decoding procedures to be adaptive. Thus, we do not require $\mathcal{D}$ to be non-adaptive in Condition 1 of Definition 1.

1.1 Our results

We obtain efficient error-correcting data structures for the following two data structure problems.

MEMBERSHIP: Consider a universe $[n] = \{1, \ldots, n\}$ and some nonnegative integer $s \leq n$. Given a set $S \subseteq [n]$ with at most s elements, one would like to store S in a compact representation that can answer "membership queries" efficiently, i.e., given an index $i \in [n]$, determine whether or not $i \in S$. Formally $D = \{S : S \subseteq [n], |S| \leq s\}$, $Q = [n]$, and $A = \{0, 1\}$. The function $\text{MEM}_{n,s}(S, i)$ is 1 if $i \in S$ and 0 otherwise.

Since there are at least $\binom{n}{s}$ subsets of the universe of size at most s, each subset requiring a different instantiation of the data structure, the information-theoretic lower bound on the space of any data structure is at least $\log \binom{n}{s} \approx s \log n$ bits.[2] An easy way to achieve this is to store S in sorted order. If each number is stored in its own $\log n$-bit "cell," this data structure takes s cells, which is $s \log n$ bits. To answer a membership query, one can do a binary search on the list to determine whether $i \in S$ using about $\log s$ "cell-probes," or $\log s \cdot \log n$ bit-probes. The length of this data structure is essentially optimal, but its number of probes is not. Fredman, Komlós, and Szemerédi [11] developed a famous hashing-based data structure that has length $O(s)$ cells (which is $O(s \log n)$ bits) and only needs a *constant* number of cell-probes (which is $O(\log n)$ bit-probes). Buhrman, Miltersen, Radhakrishnan, and Venkatesh [6] improved upon this by designing a data structure of length $O(s \log n)$ bits that answers queries with *only one bit-probe* and a small error probability. This is simultaneously optimal in terms of time (clearly one bit-probe cannot be improved upon) and space (up to a constant factor).

None of the aforementioned data structures can tolerate a constant fraction of noise. To protect against noise for this problem, de Wolf [20] constructed an error-correcting data structure with $\lambda = 0$ using a locally decodable code (LDC). That construction answers membership queries in t bit-probes and has length roughly $L(s, t) \log n$, where $L(s, t)$ is the shortest length of an LDC encoding s bits with bit-probe complexity t. Currently, all known LDCs with $t = O(1)$ have $L(s, t)$ super-polynomial in s [3, 22, 7]. In fact, $L(s, t)$ must be super-linear for all constant t, see e.g. [14, 16, 21].

Under our present model of error-correction, we can construct much more efficient data structures with error-correcting capability. First, it is not hard to show that by composing the BMRV data structure [6] with the error-correcting data structure for $\text{MEM}_{n,n}$ (equivalently, an RLDC) [4], one can already obtain an error-correcting data structure of length $O((s \log n)^{1+\eta})$, where η is an arbitrarily small constant. However, following an approach taken in [20], we obtain a data structure of length $O(s^{1+\eta} \log n)$, which is much shorter than the aforementioned construction if $s = o(\log n)$.

Theorem 1. *For every $\varepsilon, \eta \in (0, 1)$, there exist an integer $t > 0$ and real $\tau > 0$, such that for all s and n, and every $\delta \leq \tau$, $\text{MEM}_{n,s}$ has a $(t, \delta, \varepsilon, \frac{s}{2n})$-data structure of length $O(s^{1+\eta} \log n)$.*

We will prove Theorem 1 in Section 2. Note that the size of the good set G is at least $n - \frac{s}{2}$. Hence corrupting a δ-fraction of the bits of the data structure may cause a decoding failure for at most half of the queries $i \in S$ but not all. One may replace this factor $\frac{1}{2}$ easily by another constant (though the parameters t and τ will then change).

[2] Our logs are always to base 2.

POLYNOMIAL EVALUATION: Let $\mathbb{Z}_n$ denote the set of integers modulo n and $s \leq n$ be some nonnegative integer. Given a univariate polynomial $g \in \mathbb{Z}_n[X]$ of degree at most s, we would like to store g in a compact representation so that for each evaluation query $a \in \mathbb{Z}_n$, $g(a)$ can be computed efficiently. Formally, $D = \{g : g \in \mathbb{Z}_n[X], \deg(g) \leq s\}$, $Q = \mathbb{Z}_n$, and $A = \mathbb{Z}_n$, and the function is $\text{POLYEVAL}_{n,s}(g, a) = g(a)$.

Since there are n^{s+1} polynomials of degree at most s, with each polynomial requiring a different instantiation of the data structure, the information-theoretic lower bound on the space of any data structure for this problem is at least $\log(n^{s+1}) \approx s \log n$ bits. Since each answer is an element of $\mathbb{Z}_n$ and must be represented by $\lfloor \log n \rfloor + 1$ bits, $\lfloor \log n \rfloor + 1$ is the information-theoretic lower bound on the bit-probe complexity.

Consider the following two naive solutions. On one hand, one can simply record the evaluations of g in a table with n entries, each with $\lfloor \log n \rfloor + 1$ bits. The length of this data structure is $O(n \log n)$ and each query requires reading only $\lfloor \log n \rfloor + 1$ bits. On the other hand, g can be stored as a table of its $s + 1$ coefficients. This gives a data structure of length and bit-probe complexity $(s + 1)(\lfloor \log n \rfloor + 1)$.

A natural question is whether one can construct a data structure that is optimal both in terms of space and time, i.e., has length $O(s \log n)$ and answers queries with $O(\log n)$ bit-probes. No such constructions are known to exist. However, some lower bounds are known in the weaker cell-probe model, where each cell is a sequence of $\lfloor \log n \rfloor + 1$ bits. For instance, as noted in [18], any data structure for POLYNOMIAL EVALUATION that stores $O(s^2)$ cells ($O(s^2 \log n)$ bits) requires reading at least $\Omega(s)$ cells ($\Omega(s \log n)$ bits). Moreover, by [17], if $\log n \gg s \log s$ and the data structure is constrained to store $s^{O(1)}$ cells, then its query complexity is $\Omega(s)$ cells. This implies that the second trivial construction described above is essentially optimal in the cell-probe model.

Recently, Kedlaya and Umans [15] obtained a data structure of length $s^{1+\eta} \log^{1+o(1)} n$ (where η is an arbitrarily small constant) and answers evaluation queries with $O(\text{polylog}\, s \cdot \log^{1+o(1)} n)$ bit-probes. These parameters exhibit the best tradeoff between s and n so far. When $s = n^\eta$ for some $0 < \eta < 1$, the data structure of Kedlaya and Umans [15] is much superior to the trivial solution: its length is nearly optimal, and the query complexity drops from $\text{poly}\, n$ to only $\text{polylog}\, n$ bit-probes.

Here we construct an error-correcting data structure for the polynomial evaluation problem that works even in the presence of adversarial noise, with length nearly linear in $s \log n$ and bit-probe complexity $O(\text{polylog}\, s \cdot \log^{1+o(1)} n)$. Formally:

Theorem 2. *For every $\varepsilon, \lambda, \eta \in (0, 1)$, there exists $\tau \in (0, 1)$ such that for all positive integers $s \leq n$, for all $\delta \leq \tau$, the data structure problem* $\text{POLYEVAL}_{n,s}$ *has a $(O(\text{polylog}\, s \cdot \log^{1+o(1)} n), \delta, \varepsilon, \lambda)$-data structure of length $O((s \log n)^{1+\eta})$.*

Remark. We note that Theorem 2 easily holds when $s = (\log n)^{o(1)}$. As we discussed previously, one can just store a table of the $s + 1$ coefficients of g. To make this error-correcting, encode the entire table by a standard error-correcting code. This has length and bit-probe complexity $O(s \log n) = O(\log^{1+o(1)} n)$.

1.2 Our techniques

At a high level, for both data structure problems we build our constructions by composing a relaxed locally decodable code with an appropriate noiseless data structure. If the underlying probe-accessing scheme in a noiseless data structure is "pseudorandom," then the noiseless data structure can be made error-correcting by appropriate compositions with other data structures. By pseudorandom, we mean that if a query is chosen uniformly at random from Q, then the positions of the probes selected also "behave" as if they are chosen uniformly at random. Such property allows us to analyze the error-tolerance of our constructions.

More specifically, for the MEMBERSHIP problem we build upon the noiseless data structure of Buhrman et al. [6]. While de Wolf [20] combined this with LDCs to get a rather long data structure with $\lambda = 0$, we will combine it here with RLDCs to get nearly optimal length with small (but non-zero) λ. In order to bound

λ in our new construction, we make use of the fact that the [6]-construction is a bipartite *expander graph*, as explained below after Theorem 4. This property wasn't needed in [20]. The left side of the expander represents the set of queries, and a neighborhood of a query (a left node) represents the set of possible bit-probes that can be chosen to answer this query. The expansion property of the graph essentially implies that for a random query, the distribution of a bit-probe chosen to answer this query is close to uniform.[3] This property allows us to construct an efficient, error-correcting data structure for this problem.

For the polynomial evaluation problem, we rely upon the noiseless data structure of Kedlaya and Umans [15], which has a decoding procedure that uses the reconstructive algorithm from the Chinese Remainder Theorem. The property that we need is the simple fact that if a is chosen uniformly at random from $\mathbb{Z}_n$, then for any $m \leq n$, a modulo m is uniformly distributed in $\mathbb{Z}_m$. This implies that for a random evaluation point a, the distribution of certain tuples of cell-probes used to answer this evaluation point is close to uniform. This observation allows us to construct an efficient, error-correcting data structure for polynomial evaluation. Our construction follows the non-error-correcting one of [15] fairly closely; the main new ingredient is to add redundancy to their Chinese Remainder-based reconstruction by using more primes, which gives us the error-correcting features we need.

Time-complexity of decoding and encoding. So far we have used the number of bit-probes as a proxy for the actual time the decoder needs for query-answering. This is fairly standard, and usually justified by the fact that the actual time complexity of decoding is not much worse than its number of bit-probes. This is also the case for our constructions. For MEMBERSHIP, it can be shown that the decoder uses $O(1)$ probes and $\mathrm{polylog}(n)$ time (as do the RLDCs of [4]). For POLYNOMIAL EVALUATION, the decoder uses $\mathrm{polylog}(s)\log^{1+o(1)}(n)$ probes and $\mathrm{polylog}(sn)$ time.

The efficiency of *encoding*, i.e., the "pre-processing" of the data into the form of a data structure, for both our error-correcting data structures MEMBERSHIP and POLYNOMIAL EVALUATION depends on the efficiency of encoding of the RLDC constructions in [4]. This is not addressed explicitly there, and needs further study.

2 The MEMBERSHIP problem

In this section we construct a data structure for the membership problem $\mathrm{MEM}_{n,s}$. First we describe some of the building blocks that we need to prove Theorem 1. Our first basic building block is the relaxed locally decodable code of Ben-Sasson et al. [4] with nearly linear length. Using our terminology, we can restate their result as follows:

Theorem 3 (BGHSV [4])**.** *For every $\varepsilon \in (0, 1/2)$ and $\eta > 0$, there exist an integer $t > 0$ and reals $c > 0$ and $\tau > 0$, such that for every n and every $\delta \leq \tau$, the membership problem $\mathrm{MEM}_{n,n}$ has a $(t, \delta, \varepsilon, c\delta)$-data structure for $\mathrm{MEM}_{n,n}$ of length $O(n^{1+\eta})$.*

Note that by picking the error-rate δ a sufficiently small constant, one can set $\lambda = c\delta$ (the fraction of unrecoverable queries) to be very close to 0.

The other building block that we need is the following one-probe data structure of Buhrman et al. [6].

Theorem 4 (BMRV [6])**.** *For every $\varepsilon \in (0, 1/2)$ and for every positive integers $s \leq n$, there is an $(1, 0, \varepsilon, 0)$-data structure for $\mathrm{MEM}_{n,s}$ of length $m = \frac{100}{\varepsilon^2} s \log n$ bits.*

[3] We remark that this is different from the notion of smooth decoding in the LDC literature, which requires that for every *fixed* query, each bit-probe by itself is chosen with probability close to uniform (though not independent of the other bit-probes).

Properties of the BMRV encoding: The encoding can be represented as a bipartite graph $\mathcal{G} = (L, R, E)$ with $|L| = n$ left vertices and $|R| = m$ right vertices, and regular left degree $d = \frac{\log n}{\varepsilon}$. $\mathcal{G}$ is an expander graph: for each set $S \subseteq L$ with $|S| \leq 2s$, its neighborhood $\Gamma(S)$ satisfies $|\Gamma(S)| \geq \left(1 - \frac{\varepsilon}{2}\right)|S|d$. For each assignment of bits to the left vertices with at most s ones, the encoding specifies an assignment of bits to the right vertices. In other words, each $x \in \{0,1\}^n$ of weight $|x| \leq s$ corresponds to an assignment to the left vertices, and the m-bit encoding of x corresponds to an assignment to the right vertices.

For each $i \in [n]$ we write $\Gamma_i := \Gamma(\{i\})$ to denote the set of neighbors of i. A crucial property of the encoding function $\mathcal{E}_{bmrv}$ is that for every x of weight $|x| \leq s$, for each $i \in [n]$, if $y = \mathcal{E}_{bmrv}(x) \in \{0,1\}^m$ then $\Pr_{j \in \Gamma_i}[x_i = y_j] \geq 1 - \varepsilon$. Hence the decoder for this data structure can just probe a random index $j \in \Gamma_i$ and return the resulting bit y_j. Note that this construction is not error-correcting at all, since $|\Gamma_i|$ errors in the data structure suffice to erase all information about the i-th bit of the encoded x. □

As we mentioned in the Section 1.1, by combining the BMRV encoding with the data structure for $\text{MEM}_{n,n}$ from Theorem 3, one easily obtains an $(O(1), \delta, \varepsilon, O(\delta))$-data structure for $\text{MEM}_{n,s}$ of length $O((s \log n)^{1+\eta})$. However, we can give an even more efficient, error-correcting data structure of length $O(s^{1+\eta} \log n)$. Our improvement follows an approach taken in de Wolf [20], which we now describe. For a vector $x \in \{0,1\}^n$ with $|x| \leq s$, consider a BMRV structure encoding $20n$ bits into m bits. Now, from Section 2.3 in [20], the following "balls and bins estimate" is known:

Proposition 5 (From [20])**.** *For every positive integers $s \leq n$, the BMRV bipartite graph $\mathcal{G} = ([20n], [m], E)$ for $\text{MEM}_{20n,s}$ with error parameter $\frac{1}{10}$ has the following property: there exists a partition of $[m]$ into $b = 10 \log(20n)$ disjoint sets $B_1, \ldots, B_b$ of $10^3 s$ vertices each, such that for each $i \in [n]$, there are at least $\frac{b}{4}$ sets B_k satisfying $|\Gamma_i \cap B_k| = 1$.*

Proposition 5 suggests the following encoding and decoding procedures. To encode x, we rearrange the m bits of $\mathcal{E}_{bmrv}(x)$ into $\Theta(\log n)$ disjoint blocks of $\Theta(s)$ bits each, according to the partition guaranteed by Proposition 5. Then for each block, encode these bits with the error-correcting data structure (RLDC) from Theorem 3. Given a received word w, to decode $i \in [n]$, pick a block B_k at random. With probability at least $\frac{1}{4}$, $\Gamma_i \cap B_k = \{j\}$ for some j. Run the RLDC decoder to decode the j-th bit of the k-th block of w. Since most blocks don't have much higher error-rate than the average (which is at most δ), with high probability we recover $\mathcal{E}_{bmrv}(x)_j$, which equals x_i with high probability. Finally, we will argue that most queries do not receive a blank symbol $\perp$ as an answer, using the expansion property of the BMRV encoding structure. We now proceed with a formal proof of Theorem 1.

Proof of Theorem 1. We only construct an error-correcting data structure with error probability 0.49. By a standard amplification technique we can reduce the error probability to any other positive constant (i.e., repeat the decoder $O(\log(1/\varepsilon))$ times).

By Theorem 4, there exists an encoder $\mathcal{E}_{bmrv}$ for an $(1, 0, \frac{1}{10}, 0)$-data structure for the membership problem $\text{MEM}_{20n,s}$ of length $m = 10^4 s \log(20n)$. Let $s' = 10^3 s$. By Theorem 3, for every $\eta > 0$, for some $t = O(1)$, and sufficiently small δ, $\text{MEM}_{s',s'}$ has an $(t, 10^5\delta, \frac{1}{100}, O(\delta))$-data structure of length $s'' = O(s'^{1+\eta})$. Let $\mathcal{E}_{bghsv}$ and $\mathcal{D}_{bghsv}$ be its encoder and decoder, respectively.

Encoding. Let $B_1, \ldots, B_b$ be a partition of $[m]$ as guaranteed by Proposition 5. For a string $w \in \{0,1\}^m$, we abuse notation and write $w = w_{B_1} \cdots w_{B_b}$ to denote the string obtained from w by applying the permutation on $[m]$ according to the partition $B_1, \ldots, B_b$. In other words, w_{B_k} is the concatenation of w_i where $i \in B_k$. We now describe the encoding process.

Encoder $\mathcal{E}$: on input $x \in \{0,1\}^n$, $|x| \leq s$,

1. Let $y = \mathcal{E}_{bmrv}\left(x0^{19n}\right)$ and write $y = y_{B_1} \ldots y_{B_b}$.

2. Output the concatenation $\mathcal{E}(x) = \mathcal{E}_{bghsv}(y_{B_1}) \dots \mathcal{E}_{bghsv}(y_{B_b})$.

The length of $\mathcal{E}(x)$ is $N = b \cdot O(s'^{1+\eta}) = O(s^{1+\eta} \log n)$.

Decoding. Given a string $w \in \{0,1\}^N$, we write $w = w^{(1)} \dots w^{(b)}$, where for $k \in [b]$, $w^{(k)}$ denotes the s''-bit string $w_{s''\cdot(k-1)+1} \dots w_{s''\cdot k}$.

Decoder $\mathcal{D}$: on input i and with oracle access to a string $w \in \{0,1\}^N$,

1. Pick a random $k \in [b]$.

2. If $|\Gamma_i \cap B_k| \neq 1$, then output a random bit.
   Else, let $\Gamma_i \cap B_k = \{j\}$. Run and output the answer given by the decoder $\mathcal{D}_{bghsv}(j)$, with oracle access to the s''-bit string $w^{(k)}$.

Analysis. Fix $x \in D$ and $w \in \{0,1\}^N$ such that $\Delta(w, \mathcal{E}(x)) \leq \delta N$, where δ is less than some small constant τ to be specified later. We now verify the four conditions of Definition 1. For Condition 1, note that the number of probes the decoder $\mathcal{D}$ makes is the number of probes the decoder $\mathcal{D}_{bghsv}$ makes, which is at most t, a fixed integer.

We now examine Condition 2. Fix $i \in [n]$. By Markov's inequality, for a random $k \in [b]$, the probability that the relative Hamming distance between $\mathcal{E}(y_{B_k})$ and $w^{(k)}$ is greater than $10^5\delta$ is at most 10^{-5}. If k is chosen such that the fraction of errors in $w^{(k)}$ is at most $10^5\delta$ and $\Gamma_i \cap B_k = \{j\}$, then with probability at least 0.99, $\mathcal{D}_{bghsv}$ outputs y_j or $\perp$. Let $\beta \geq \frac{1}{4}$ be the fraction of $k \in [b]$ such that $|\Gamma_i \cap B_k| = 1$. Then

$$\Pr[\mathcal{D}(i) \in \{x_i, \perp\}] \geq (1-\beta)\frac{1}{2} + \beta\frac{99}{100} - \frac{1}{10^5} > 0.624. \tag{1}$$

To prove Condition 3, we need the expansion property of the BMRV structure, as explained after Theorem 4. For $k \in [b]$, define $G_k \subseteq B_k$ so that $j \in G_k$ if $\Pr\left[\mathcal{D}^{w^{(k)}}_{bghsv}(j) = y_j\right] \geq 0.99$. In other words, G_k consists of indices in block B_k that are answered correctly by $\mathcal{D}_{bghsv}$ with high probability. By Theorem 3, if the fraction of errors in $w^{(k)}$ is at most $10^5\delta$, then $|G_k| \geq (1 - c\delta)|B_k|$ for some fixed constant c. Set $A = \cup_{k\in[b]} B_k \backslash G_k$, Since we showed above that for a $(1 - 10^{-5})$-fraction of $k \in [b]$, the fractional number of errors in $w^{(k)}$ is at most $10^5\delta$, we have $|A| \leq c\delta m + 10^{-5}m$.

Recall that the BMRV expander has left degree $d = 10\log(20n)$. Take δ small enough that $|A| < \frac{1}{40}sd$; this determines the value of τ of the theorem. We need to show that for any such small set A, most queries $i \in [n]$ are answered correctly with probability at least 0.51. It suffices to show that for most i, most of the set Γ_i falls outside of A. To this end, let $B(A) = \{i \in [n] : |\Gamma_i \cap A| \geq \frac{d}{10}\}$. We show that if A is small then $B(A)$ is small.

Claim 6. *For every $A \subseteq [m]$ with $|A| < \frac{sd}{40}$, it is the case that $|B(A)| < \frac{s}{2}$.*

Proof. Suppose, by way of contradiction, that $B(A)$ contains a set W of size $s/2$. W is a set of left vertices in the underlying expander graph $\mathcal{G}$, and since $|W| < 2s$, we must have

$$|\Gamma(W)| \geq \left(1 - \frac{1}{20}\right) d|W|.$$

By construction, each vertex in W has at most $\frac{9}{10}d$ neighbors outside A. Thus, we can bound the size of

$\Gamma(W)$ from above as follows

$$\begin{aligned} |\Gamma(W)| &\leq |A| + \frac{9}{10}d|W| \\ &< \frac{1}{40}ds + \frac{9}{10}d|W| \\ &= \frac{1}{20}d|W| + \frac{9}{10}d|W| \\ &= \left(1 - \frac{1}{20}\right) d|W|. \end{aligned}$$

This is a contradiction. Hence no such W exists and $|B(A)| < \frac{s}{2}$. □

Define $G = [n] \backslash B(A)$ and notice that $|G| > n - \frac{s}{2}$. It remains to show that each query $i \in G$ is answered correctly with probability > 0.51. To this end, we have

$$\begin{aligned} \Pr[\mathcal{D}(i) = \perp] &\leq \Pr[\mathcal{D} \text{ probes a block with noise-rate } > 10^5\delta] + \\ &\quad \Pr[\mathcal{D} \text{ probes a } j \in A] + \Pr[\mathcal{D}(i) = \perp : \mathcal{D} \text{ probes a } j \notin A] \\ &\leq \frac{1}{10^5} + \frac{1}{10} + \frac{1}{100} < 0.111. \end{aligned}$$

Combining with Eq. (1), for all $i \in G$ we have

$$\Pr[\mathcal{D}(i) = x_i] = \Pr[\mathcal{D}(i) \in \{x_i, \perp\}] - \Pr[\mathcal{D}(i) = \perp] \geq 0.51.$$

Finally, Condition 4 follows from the corresponding condition of the data structure for $\text{MEM}_{n,n}$. □

3 The POLYNOMIAL EVALUATION problem

In this section we prove Theorem 2. Given a polynomial g of degree s over $\mathbb{Z}_n$, our goal is to write down a data structure of length roughly linear in $s \log n$ so that for each $a \in \mathbb{Z}_n$, $g(a)$ can be computed with approximately $\text{polylog}\, s \cdot \log n$ bit-probes. Our data structure is built on the work of Kedlaya and Umans [15]. Since we cannot quite use their construction as a black-box, we first give a high-level overview of our proof, motivating each of the proof ingredients that we need.

Encoding based on reduced polynomials: The most naive construction, by recording $g(a)$ for each $a \in \mathbb{Z}_n$, has length $n \log n$ and answers an evaluation query with $\log n$ bit-probes. As explained in [15], one can reduce the length by using the Chinese Remainder Theorem (CRT): If P_1 is a collection of distinct primes, then a nonnegative integer $m < \prod_{p \in P_1} p$ is uniquely specified by (and can be reconstructed efficiently from) the values $[m]_p$ for each $p \in P_1$, where $[m]_p$ denotes $m \mod p$.

Consider the value $g(a)$ over $\mathbb{Z}$, which can be bounded above by n^{s+2}, for $a \in \mathbb{Z}_n$. Let P_1 consist of the first $\log(n^{s+2})$ primes. For each $p \in P_1$, compute the reduced polynomial $g_p := g \mod p$ and write down $g_p(b)$ for each $b \in \mathbb{Z}_p$. Consider the data structure that simply concatenates the evaluation table of every reduced polynomial. This data structure has length $|P_1|(\max_{p \in P_1} p)^{1+o(1)}$, which is $s^{2+o(1)} \log^{2+o(1)} n$ by the Prime Number Theorem (see Fact 12 in Appendix B). Note that $g(a) < \prod_{p \in P_1} p$. So to compute $[g(a)]_n$, it suffices to apply CRT to reconstruct $g(a)$ over $\mathbb{Z}$ from the values $[g(a)]_p = g_p([a]_p)$ for each $p \in P_1$. The number of bit-probes is $|P_1| \log(\max_{p \in P_1} p)$, which is $s^{1+o(1)} \log^{1+o(1)} n$.

Error-correction with reduced polynomials: The above CRT-based construction has terrible parameters, but it serves as an important building block from which we can obtain a data structure with better parameters. For now, we explain how the above CRT-based encoding can be made error-correcting. One can protect the bits of the evaluation tables of each reduced polynomial by an RLDC as provided by Theorem 3. However, the evaluation tables can have non-binary alphabets, and a bit-flip in just one "entry" of an evaluation table can destroy the decoding process. To remedy this, one can first encode each entry by a standard error-correcting code and then encode the concatenation of all the tables by an RLDC. This is encapsulated in Lemma 7, which can be viewed as a version of Theorem 3 over non-binary alphabet. We prove this in Appendix A.

Lemma 7. *Let $f : D \times Q \to \{0,1\}^\ell$ be a data structure problem. For every $\varepsilon, \eta, \lambda \in (0,1)$, there exists $\tau \in (0,1)$ such that for every $\delta \le \tau$, f has an $(O(1), \delta, \varepsilon, \lambda)$-data structure of length $O((\ell|Q|)^{1+\eta})$.*

To apply Lemma 7, let D be the set of degree-s polynomials over $\mathbb{Z}_n$, Q be the set of all evaluation points of all the reduced polynomials of g (each specified by a pair (a,p)), and the data structure problem f outputs evaluations of some reduced polynomial of g.

By itself, Lemma 7 cannot guarantee resilience against noise. In order to apply the CRT to reconstruct $g(a)$, all the values $\{[g(a)]_p : p \in P_1\}$ must be correct, which is not guaranteed by Lemma 7. To fix this, we add redundancy, taking a larger set of primes than necessary so that the reconstruction via CRT can be made error-correcting. Specifically, we apply a Chinese Remainder Code, or CRT code for short, to the encoding process.

Definition 2 (CRT code)**.** Let $p_1 < p_2 < \ldots < p_N$ be distinct primes, $K < N$, and $T = \prod_{i=1}^{K} p_i$. The *Chinese Remainder Code (CRT code)* with basis $p_1, \ldots, p_N$ and rate $\frac{K}{N}$ over message space $\mathbb{Z}_T$ encodes $m \in \mathbb{Z}_T$ as $\langle [m]_{p_1}, [m]_{p_2}, \ldots, [m]_{p_N} \rangle$.

Remark. By CRT, for distinct $m_1, m_2 \in \mathbb{Z}_T$, their encodings agree on at most $K-1$ coordinates. Hence the Chinese Remainder Code with basis $p_1 < \ldots < p_N$ and rate $\frac{K}{N}$ has distance $N - K + 1$.

It is known that good families of CRT code exist and that unique decoding algorithms for CRT codes (see e.g., [12]) can correct up to almost half of the distance of the code. The following statement can be easily derived from known facts, and we include a proof in Appendix B.

Theorem 8. *For every positive integer T, there exists a set P consisting of distinct primes, with (1) $|P| = O(\log T)$, and (2) $\forall p \in P, \log T < p < 500 \log T$, such that a CRT code with basis P and message space $\mathbb{Z}_T$ has rate $\frac{1}{2}$, and can correct up to a $(\frac{1}{4} - O(\frac{1}{\log\log T}))$-fraction of errors.*

We apply Theorem 8 to a message space of size n^{s+2} to obtain a set of primes P_1 with the properties described above. Note that these primes are all within a constant factor of one another, and in particular, the evaluation table of each reduced polynomial has the same length, up to a constant factor. This fact and Lemma 7 will ensure that our CRT-based encoding is error-correcting.

Reducing the bit-probe complexity: We now explain how to reduce the bit-probe complexity of the CRT-based encoding, using an idea from [15]. Write $s = d^m$, where $d = \log^C s$, $m = \frac{\log s}{C \log\log s}$, and $C > 1$ is a sufficiently large constant. Consider the following multilinear extension map $\psi_{d,m} : \mathbb{Z}_n[X] \to \mathbb{Z}_n[X_0, \ldots, X_{m-1}]$ that sends a univariate polynomial of degree at most s to an m-variate polynomial of degree less than d in each variable. For every $i \in [s]$, write $i = \sum_{j=0}^{m-1} i_j d^j$ in base d. Define $\psi_{d,m}$ which sends X^i to $X_1^{i_0} \cdots X_m^{i_{m-1}}$ and extends multilinearly to $\mathbb{Z}_n[X]$.

To simplify our notation, we write $\tilde{g}$ to denote the multivariate polynomial $\psi_{d,m}(g)$. For every $a \in \mathbb{Z}_n$, define $\tilde{a} \in \mathbb{Z}_n^m$ to be $([a]_n, [a^d]_n, [a^{d^2}]_n, \ldots, [a^{d^{m-1}}]_n)$. Note that for every $a \in \mathbb{Z}_n$, $g(a) = \tilde{g}(\tilde{a}) \pmod{n}$. Now the trick is to observe that the total degree of the multilinear polynomial $\tilde{g}$ is less than the degree of the univariate polynomial g, and hence its maximal value over the integers is much reduced. In particular, for every $a \in \mathbb{Z}_n^m$, the value $\psi_{d,m}(g)(a)$ over the integers is bounded above by $d^m n^{dm+1}$.

We now work with the reduced polynomials of $\tilde{g}$ for our encoding. Let P_1 be the collection of primes guaranteed by Theorem 8 when $T_1 = d^m n^{dm+1}$. For $p \in P_1$, let $\tilde{g}_p$ denote $\tilde{g} \mod p$ and $\tilde{a}_p$ denote the point $([a]_p, [a^d]_p, \ldots, [a^{d^{m-1}}]_p)$. Consider the data structure that concatenates the evaluation table of $\tilde{g}_p$ for each $p \in P_1$. For each $a \in \mathbb{Z}_n$, to compute $g(a)$, it suffices to compute $\tilde{g}(\tilde{a})$ over $\mathbb{Z}$, which by Theorem 8 can be reconstructed (even with noise) from the set $\{\tilde{g}_p(\tilde{a}_p) : p \in P_1\}$.

Since the maximum value of $\tilde{g}$ is at most $T_1 = d^m n^{dm+1}$ (whereas the maximum value of g is at most $d^m n^{d^m+1}$), the number of primes we now use is significantly less. This effectively reduces the bit-probe complexity. In particular, each evaluation query can be answered with $|P_1| \cdot \max_{p \in P_1} \log p = (dm \log n)^{1+o(1)}$ bit-probes, which by our choice of d and m is equal to $\mathrm{polylog}\, s \cdot \log^{1+o(1)} n$. However, the *length* of this encoding is still far from the information-theoretically optimal $s \log n$ bits. We shall explain how to reduce the length, but since encoding with multilinear reduced polynomials introduces potential complications in error-correction, we first explain how to circumvent these complications.

Error-correction with reduced multivariate polynomials: There are two complications that arise from encoding with reduced multivariate polynomials. The first is that not all the points in the evaluation tables are used in the reconstructive CRT algorithm. Lemma 7 only guarantees that most of the entries of the table can be decoded, not all of them. So if the entries that are used in the reconstruction via CRT are not decoded by Lemma 7, then the whole decoding procedure fails.

More specifically, to reconstruct $\tilde{g}(\tilde{a})$ over $\mathbb{Z}_n$, it suffices to query the point $\tilde{a}_p$ in the evaluation table of $\tilde{g}_p$ for each $p \in P_1$. Typically the set $\{\tilde{a}_p : a \in \mathbb{Z}_n\}$ will be much smaller than $\mathbb{Z}_p^m$, so not all the points in $\mathbb{Z}_p^m$ are used. To circumvent this issue, we only store the query points that are used in the CRT reconstruction. Let $B^p = \{\tilde{a}_p : a \in \mathbb{Z}_n\}$. For each $p \in P_1$, the encoding only stores the evaluation of $\tilde{g}_p$ at the points B^p instead of the entire domain $\mathbb{Z}_p^m$. The disadvantage of computing the evaluation at the points in B^p is that the encoding stage takes time proportional to n. We thus give up on encoding efficiency (which was one of the main goals of Kedlaya and Umans) in order to guarantee error-correction.

The second complication is that the sizes of the evaluation tables may no longer be within a constant factor of each other. (This is true even if the evaluation points come from all of $\mathbb{Z}_p^m$.) If one of the tables has length significantly longer than the others, then a constant fraction of noise may completely corrupt the entries of all the other small tables, rendering decoding via CRT impossible. This potential problem is easy to fix; we apply a repetition code to each evaluation table so that all the tables have equal length.

Reducing the length: Now we explain how to reduce the length of the data structure to nearly $s \log n$, along the lines of Kedlaya and Umans [15]. To reduce the length, we need to reduce the magnitude of the primes used by the CRT reconstruction. We can effectively achieve that by applying the CRT twice. Instead of storing the evaluation table of $\tilde{g}_p$, we apply CRT again and store evaluation tables of the reduced polynomials of $\tilde{g}_p$ instead. Whenever an entry of $\tilde{g}_p$ is needed, we can apply the CRT reconstruction to the reduced polynomials of $\tilde{g}_p$.

Note that for $p_1 \in P_1$, the maximum value of $\tilde{g}_{p_1}$ (over the integers rather than mod n) is at most $T_2 = d^m p_1^{dm+1}$. Now apply Theorem 8 with T_2 the size of the message space to obtain a collection of primes P_2. Recall that each $p_1 \in P_1$ is at most $O(dm \log n)$. So each $p_2 \in P_2$ is at most $O((dm)^{1+o(1)} \log\log n)$, which also bounds the cardinality of P_2 from above.

For each query, the number of bit-probes made is at most $|P_1||P_2| \max_{p_2 \in P_2} \log p_2$, which is at most $(dm)^{2+o(1)} \log^{1+o(1)} n$. Recall that by our choice $d = \log^C s$ and $m = \frac{\log s}{C \log\log s}$, we have $dm = \frac{\log^{C+1} s}{C \log\log s}$. Thus, the bit-probe complexity is $\mathrm{polylog}\, s \cdot \log^{1+o(1)} n$.

Next we bound the length of the encoding. Recall that by the remark following Theorem 2, we may assume without loss of generality that $s = \Omega(\log^\zeta n)$ for some $0 < \zeta < 1$. This implies $\log\log n = O(\log s)$. Then for each $p_2 \in P_2$,

$$p_2^m \leq \left(O\left((dm)^{1+o(1)} \log\log n\right)\right)^m \leq (dm)^{(1+o(1))m} \cdot s^{\frac{1}{C}+o(1)} \leq s^{1+\frac{2}{C}+o(1)}.$$

Now, by Lemma 7, the length of the encoding is nearly linear in $|P_1||P_2| \max_{p_2 \in P_2} p_2^m \log p_2$, which is at most $\mathrm{polylog}\, s \cdot \log^{1+o(1)} n \cdot \max_{p_2 \in P_2} p_2^m$. Putting everything together, the length of the encoding is nearly linear in $s \log n$. We now proceed with a formal proof.

Proof of Theorem 2. We only construct an error-correcting data structure with error probability $\varepsilon = \frac{1}{4}$. By a standard amplification technique (i.e., $O(\log(1/\varepsilon))$ repetitions) we can reduce the error probability to any other positive constant. We now give a formal description of the encoding and decoding algorithms.

Encoding: Apply Theorem 8 with $T = d^m n^{dm+1}$ to obtain a collection of primes P_1. Apply Theorem 8 with $T = d^m (\max_{p \in P_1} p)^{dm+1}$ to obtain a collection of primes P_2. Set $p_{max} = \max_{p_2 \in P_2} p_2$.

Now, for each $p_1 \in P_1$, $p_2 \in P_2$, define a collection of evaluation points $B^{p_1,p_2} = \{\tilde{a}_{p_1,p_2} : a \in \mathbb{Z}_n\}$. Fix a univariate polynomial $g \in \mathbb{Z}_n[x]$ of degree at most s. For every $p_1 \in P_1$, $p_2 \in P_2$, view each evaluation of the reduced multivariate polynomial $\tilde{g}_{p_1,p_2}$ as a bit-string of length exactly $\lceil \log p_{max} \rceil$. Let $L = \max_{p_1 \in P_1, p_2 \in P_2} |B^{p_1,p_2}|$ and for each $p_1 \in P_1$, $p_2 \in P_2$, set $r^{p_1,p_2} = \left\lceil \frac{L}{|B^{p_1,p_2}|} \right\rceil$. Define f^{p_1,p_2} to be the concatenation of r^{p_1,p_2} copies of the string $\langle \tilde{g}(q) \rangle_{q \in B^{p_1,p_2}}$. Define the string $f = \langle f^{p_1,p_2} \rangle_{p_1 \in P_1, p_2 \in P_2}$.

We want to apply Lemma 7 to protect the string f, which we can since f may be viewed as a data structure problem, as follows. The set of data-items is the set of polynomials g as above. The set of queries Q is $\bigcup_{p_1 \in P_1, p_2 \in P_2} B^{p_1,p_2} \times [r^{p_1,p_2}]$. The answer to query $(q^{p_1,p_2}, i^{p_1,p_2})$ is the i^{p_1,p_2}-th copy of $\tilde{g}_{p_1,p_2}(q^{p_1,p_2})$.

Fix $\lambda \in (0,1)$. By Lemma 7, for every $\eta > 0$, there exists $\tau_0 \in (0,1)$ such that for every $\delta \leq \tau_0$, the data structure problem corresponding to f has a $(O(\log p_{max}), \delta, 2^{-10}, \lambda^3 2^{-36})$-data structure. Let $\mathcal{E}_0, \mathcal{D}_0$ be its encoder and decoder, respectively. Finally, the encoding of the polynomial g is simply

$$\mathcal{E}(g) = \mathcal{E}_0(f).$$

Note that the length of $\mathcal{E}(g)$ is at most $(|P_1||P_2| \max_{p_2 \in P_2} p_2^m \log p_2)^{1+\eta}$, which as we computed earlier is bounded above by $O((s \log n)^{1+\zeta})$ for some arbitrarily small constant ζ.

Decoding: We may assume, without loss of generality, that the CRT decoder $\mathcal{D}_{crt}$ from Theorem 7 outputs $\perp$ when more than a $\frac{1}{16}$-fraction of its inputs are erasures (i.e., $\perp$ symbols).

The decoder $\mathcal{D}$, with input $a \in \mathbb{Z}_n$ and oracle access to w, does the following:

1. Compute $\tilde{a} = (a, a^d, \ldots, a^{d^{m-1}}) \in \mathbb{Z}_n^m$, and for every $p_1 \in P_1$, $p_2 \in P_2$, compute the reduced evaluation points $\tilde{a}_{p_1,p_2}$.
2. For every $p_1 \in P_1$, $p_2 \in P_2$, pick $j \in [r^{p_1,p_2}]$ uniformly at random and run the decoder $\mathcal{D}_0$ with oracle access to w to obtain the answers $v^{(a)}_{p_1,p_2} = \mathcal{D}_0(\tilde{a}_{p_1,p_2}, j)$.
3. For every $p_1 \in P_1$ obtain $v^{(a)}_{p_1} = \mathcal{D}_{crt}\left(\left(v^{(a)}_{p_1,p_2}\right)_{p_2 \in P_2}\right)$.

4. Output $v^{(a)} = \mathcal{D}_{crt}\left(\left(v_{p_1}^{(a)}\right)_{p_1 \in P_1}\right)$.

Analysis: Fix a polynomial g with degree at most s. Fix a bit-string w at relative Hamming distance at most δ from $\mathcal{E}(g)$, where δ is at most τ_0. We proceed to verify that the above encoding and decoding satisfy the conditions of Definition 1.

Conditions 1 and 4 are easily verified. For Condition 1, observe that for each $p_1 \in P_1$, $p_2 \in P_2$, $\mathcal{D}_0$ makes at most $O(\log p_{max})$ bit-probes. So $\mathcal{D}$ makes at most $O(|P_1||P_2| \log p_{max})$ bit-probes, which as we calculated earlier is at most $\mathrm{polylog}\, s \cdot \log^{1+o(1)} n$.

For Condition 4, note that since $\mathcal{D}_0$ decodes correctly when no noise is present, $v_{p_1,p_2}^{(a)}$ is equal to $\tilde{g}_{p_1,p_2}(\tilde{a}_{p_1,p_2})$. By our choice of P_1 and P_2, after two applications of the Chinese Remainder Theorem, it is easy to see that $\mathcal{D}$ outputs $v = \tilde{g}(\tilde{a})$, which equals $g(a)$.

Now we verify Condition 2. Fix $a \in \mathbb{Z}_n$. We want to show that with oracle access to w, with probability at least $\frac{3}{4}$, the decoder $\mathcal{D}$ on input a outputs either $g(a)$ or $\perp$. For $\pi \in P_1 \cup (P_1 \times P_2)$, we say that a point $v_\pi^{(a)}$ is *incorrect* if $v_\pi^{(a)} \notin \{\tilde{g}_\pi(\tilde{a}_\pi), \perp\}$.

By Lemma 7, for each $p_1 \in P_1$ and $p_2 \in P_2$, $v_{p_1,p_2}^{(a)}$ is incorrect with probability at most 2^{-10}. Now fix $p_1 \in P_1$. On expectation (over the decoder's randomness), at most a 2^{-10}-fraction of the points in the set $\{v_{p_1,p_2}^{(a)} : p_2 \in P_2\}$ are incorrect. By Markov's inequality, with probability at least $1 - 2^{-6}$, the fraction of points in the set $\{v_{p_1,p_2}^{(a)} : p_2 \in P_2\}$ that are incorrect is at most $\frac{1}{16}$. If the fraction of blank symbols in the set $\{v_{p_1,p_2}^{(a)}\}_{p_2 \in P_2}$ is at least $\frac{1}{16}$, then $\mathcal{D}_{crt}$ outputs $\perp$, which is acceptable. Otherwise, the fraction of errors and erasures (i.e., $\perp$ symbols) in the set $\{v_{p_1,p_2}^{(a)} : p_2 \in P_2\}$ is at most $\frac{1}{8}$. By Theorem 8, the decoder $\mathcal{D}_{crt}$ will output an incorrect $v_{p_1}^{(a)}$ with probability at most 2^{-6}. Thus, on expectation, at most a 2^{-6}-fraction of the points in $\{v_{p_1}^{(a)} : p_1 \in P_1\}$ are incorrect. By Markov's inequality again, with probability at least $\frac{3}{4}$, at most a $\frac{1}{16}$-fraction of the points in $\{v_{p_1}^{(a)} : p_1 \in P_1\}$ are incorrect, which by Theorem 8 implies that $\mathcal{D}_a^w$ is either $\perp$ or $g(a)$. This establishes Condition 2.

We now proceed to prove Condition 3. We show the existence of a set $G \subseteq \mathbb{Z}_n$ such that $|G| \geq (1-\lambda)n$ and for each $a \in G$, we have $\Pr[\mathcal{D}(a) = g(a)] \geq \frac{3}{4}$. Our proof relies on the following observation: for any $p_1 \in P_1$ and $p_2 \in P_2$, if $a \in \mathbb{Z}_n$ is chosen uniformly at random, then the evaluation point $\tilde{a}_{p_1,p_2}$ is like a uniformly chosen element $q \in B^{p_1,p_2}$. This observation implies that if a few entries in the evaluation tables of the multivariate reduced polynomials are corrupted, then for most $a \in \mathbb{Z}_n$, the output of the decoder $\mathcal{D}$ on input a remains unaffected. We now formalize this observation.

Claim 9. *Fix $p_1 \in P_1$, $p_2 \in P_2$, and a point $q \in B^{p_1,p_2}$. Then*

$$\Pr_{a \in \mathbb{Z}_n}\left[\tilde{a}_{p_1,p_2} \equiv q\right] \leq \frac{4}{p_2}.$$

Proof. For any pair of positive integers $m \leq n$, the number of integers in $[n]$ congruent to a fixed integer mod m is at most $\left\lfloor \frac{n}{m} \right\rfloor + 1$ and at least $\left\lfloor \frac{n}{m} \right\rfloor - 1$. Note that if $a, b \in \mathbb{Z}_n$ with $a \equiv b \mod m$, then for any integer i, $a^i \equiv b^i \mod m$. Thus, $\tilde{a}_m \equiv \tilde{b}_m$.

It is not hard to see that for a fixed $q_1 \in B^{p_1}$, the number of integers $a \in \mathbb{Z}_n$ such that $\tilde{a}_{p_1} \equiv q_1$ is at most $\left\lfloor \frac{n}{p_1} \right\rfloor + 1$. Furthermore, for a fixed $q_2 \in B^{p_1,p_2}$, the number of points in B^{p_1} that are congruent to q_2 mod p_2 is at most $\left\lfloor \frac{p_1}{p_2} \right\rfloor + 1$. Thus, for a fixed $q \in B^{p_1,p_2}$, the number of integers $a \in \mathbb{Z}_n$ such that $\tilde{a}_{p_1,p_2} \equiv q$ is at most $\left(\left\lfloor \frac{n}{p_1} \right\rfloor + 1\right)\left(\left\lfloor \frac{p_1}{p_2} \right\rfloor + 1\right)$, which is at most $4\frac{n}{p_2}$ since $n \geq p_1 \geq p_2$. ∎

Now, for every $p_1 \in P_1$ and $p_2 \in P_2$, we say that a query $(q, j) \in B^{p_1,p_2} \times [r^{p_1,p_2}]$ is *bad* if the probability that $\mathcal{D}_0^w(q, j) \neq \tilde{g}_{(p_1,p_2)}(q)$ is greater than 2^{-10}. By Lemma 7, the fraction of bad queries in $\cup_{p_1,p_2} B^{p_1,p_2} \times [r^{p_1,p_2}]$ is at most $\lambda_0 := \lambda^3 2^{-36}$. We say that a tuple of primes $(p_1, p_2) \in P_1 \times P_2$ is *bad* if more than a $2^{11}\lambda_0\lambda^{-1}$-fraction of queries in $B^{p_1,p_2} \times [r^{p_1,p_2}]$ are bad (below, *good* always denotes not bad.) By averaging, the fraction of bad tuples (p_1, p_2) is at most $2^{-11}\lambda$.

For a fixed good tuple (p_1, p_2), we say that an index i^{p_1,p_2} is *bad* if more than a $2^{-11}\lambda$-fraction of queries in the copy $B^{p_1,p_2} \times \{i^{p_1,p_2}\}$ are bad. Since (p_1, p_2) is good, by averaging, at most a $2^{22}\lambda_0\lambda^{-2}$-fraction of $[r^{p_1,p_2}]$ are bad. Recall that in Step 2 of the decoder $\mathcal{D}$, the indices $\{j^{p_1,p_2} : p_1 \in P_1, p_2 \in P_2\}$ are chosen uniformly at random. So on expectation, the set of indices $\{j^{p_1,p_2} : (p_1, p_2) \text{ is good}\}$ has at most a $2^{22}\lambda_0\lambda^{-2}$-fraction of bad indices. By Markov's inequality, with probability at least $\frac{7}{8}$, the fraction of bad indices in the set $\{j^{p_1,p_2} : (p_1, p_2) \text{ is good}\}$ is at most $2^{25}\lambda_0\lambda^{-2}$. We condition on this event occurring and fix the indices j^{p_1,p_2} for each $p_1 \in P_1$, $p_2 \in P_2$.

Fix a good tuple (p_1, p_2) and a good index j^{p_1,p_2}. By Claim 9, for a uniformly random $a \in \mathbb{Z}_n$, the query $(\tilde{a}_{p_1,p_2}, j^{p_1,p_2})$ is bad with probability at most $2^{-9}\lambda$. By linearity of expectation, for a random $a \in \mathbb{Z}_n$, the expected fraction of bad queries in the set $S^a = \{(\tilde{a}_{p_1,p_2}, j^{p_1,p_2}) : p_1 \in P_1, p_2 \in P_2\}$ is at most $2^{-11}\lambda + 2^{25}\lambda_0\lambda^{-2} + 2^{-9}\lambda$, which is at most $2^{-8}\lambda$ by definition of λ_0. Thus, by Markov's inequality, for a random $a \in \mathbb{Z}_n$, with probability at least $1 - \lambda$, the fraction of bad queries in the set S^a is at most 2^{-8}. By linearity of expectation, there exists some subset $G \subseteq \mathbb{Z}_n$ with $|G| \geq (1 - \lambda)n$ such that for every $a \in G$, the fraction of bad queries in S^a is at most 2^{-8}.

Now fix $a \in G$. By definition, the fraction of bad queries in S^a is at most 2^{-8}, and furthermore, each of the good queries in S^a is incorrect with probability at most 2^{-10}. So on expectation, the fraction of errors and erasures in S^a is at most $2^{-8} + 2^{-10}$. By Markov's inequality, with probability at least $\frac{7}{8}$, the fraction of errors and erasures in the set $\{v^{(a)}_{p_1,p_2} : p_1 \in P_1, p_2 \in P_2\}$ is at most $2^{-5} + 2^{-7}$, which is at most $\frac{1}{25}$. We condition on this event occurring. By averaging, for more than a $\frac{4}{5}$-fraction of the primes $p_1 \in P_1$, the set $\{v^{(a)}_{p_1,p_2} : p_2 \in P_2\}$ has at most $\frac{1}{5}$-fraction of errors and erasures, which can be corrected by the CRT decoder $\mathcal{D}_{crt}$. Thus, after Step 3 of the decoder $\mathcal{D}$, the set $\{v^{(a)}_{p_1}\}$ has at most a $\frac{1}{5}$-fraction of errors and erasures, which again will be corrected by the CRT decoder $\mathcal{D}_{crt}$. Hence, by the union bound, the two events that we conditioned on earlier occur simultaneously with probability at least $\frac{3}{4}$, and $\mathcal{D}(a)$ will output $g(a)$. □

4 Conclusion and future work

We presented a relaxation of the notion of error-correcting data structures recently proposed in [20]. While the earlier definition does not allow data structures that are both error-correcting and efficient in time and space (unless an unexpected breakthrough happens for constant-probe LDCs), our new definition allows us to construct efficient, error-correcting data structures for both the MEMBERSHIP and the POLYNOMIAL EVALUATION problems. This opens up many directions: what other data structures can be made error-correcting?

The problem of computing *rank* within a sparse ordered set is a good target. Suppose we are given a universe $[n]$, some nonnegative integer $s \leq n$, and a subset $S \subseteq [n]$ of size at most s. The rank problem is to store S compactly so that on input $i \in [n]$, the value $|\{j \in S : j \leq i\}|$ can be computed efficiently. For easy information-theoretic reasons, any data structure for this problem needs length at least $\Omega(s \log n)$ and makes $\Omega(\log s)$ bit-probes for each query. If $s = O(\log n)$, one can trivially obtain an error-correcting data structure of optimal length $O(s \log n)$ with $O(\log^2 n)$ bit-probes, which is only quadratically worse than optimal: write down S as a string of $s \log n$ bits, encode it with a good error-correcting code, and read the entire encoding when an index is queried. However, it may be possible to do something smarter and more involved. We leave the construction of near-optimal error-correcting data structures for rank with small s

(as well as for related problems such as *predecessor*) as challenging open problems.

Acknowledgments

We thank Madhu Sudan for helpful comments and suggestions on the presentation of this paper.

References

[1] T. M. Apostol. *Introduction to Analytic Number Theory*. Springer-Verlag, New York, 1979.

[2] Y. Aumann and M. Bender. Fault-tolerant data structures. In *Proceedings of 37th IEEE FOCS*, pages 580–589, 1996.

[3] A. Beimel, Y. Ishai, E. Kushilevitz, and J. Raymond. Breaking the $O(n^{1/(2k-1)})$ barrier for information-theoretic Private Information Retrieval. In *Proceedings of 43rd IEEE FOCS*, pages 261–270, 2002.

[4] E. Ben-Sasson, O. Goldreich, P. Harsha, M. Sudan, and S. Vadhan. Robust PCPs of proximity, shorter PCPs and applications to coding. *SIAM Journal on Computing*, 36(4):889–974, 2006. Earlier version in STOC'04.

[5] G. Brodal, R. Fagerberg, I. Finocchi, F. Grandoni, G. Italiano, A. Jørgenson, G. Moruz, and T. Mølhave. Optimal resilient dynamic dictionaries. In *Proceedings of 15th European Symposium on Algorithms (ESA)*, pages 347–358, 2007.

[6] H. Buhrman, P. B. Miltersen, J. Radhakrishnan, and S. Venkatesh. Are bitvectors optimal? *SIAM Journal on Computing*, 31(6):1723–1744, 2002. Earlier version in STOC'00.

[7] K. Efremenko. 3-query locally decodable codes of subexponential length. In *Proceedings of 41st ACM STOC*, 2009.

[8] I. Finocchi, F. Grandoni, and G. Italiano. Optimal resilient sorting and searching in the presence of memory faults. In *Proceedings of 33rd ICALP*, volume 4051 of *Lecture Notes in Computer Science*, pages 286–298, 2006.

[9] I. Finocchi, F. Grandoni, and G. Italiano. Resilient search trees. In *Proceedings of 18th ACM-SIAM SODA*, pages 547–553, 2007.

[10] I. Finocchi and G. Italiano. Sorting and searching in the presence of memory faults (without redundancy). In *Proceedings of 36th ACM STOC*, pages 101–110, 2004.

[11] M. Fredman, M. Komlós, and E. Szemerédi. Storing a sparse table with $O(1)$ worst case access time. *Journal of the ACM*, 31(3):538–544, 1984.

[12] O. Goldreich, D. Ron, and M. Sudan. Chinese remaindering with errors. *IEEE Transactions on Information Theory*, 46(4):1330–1338, 2000.

[13] A. G. Jørgenson, G. Moruz, and T. Mølhave. Resilient priority queues. In *Proceedings of 10th International Workshop on Algorithms and Data Structures (WADS)*, volume 4619 of *Lecture Notes in Computer Science*, 2007.

[14] J. Katz and L. Trevisan. On the efficiency of local decoding procedures for error-correcting codes. In *Proceedings of 32nd ACM STOC*, pages 80–86, 2000.

[15] K. S. Kedlaya and C. Umans. Fast modular composition in any characteristic. In *Proceedings of 49th IEEE FOCS*, pages 146–155, 2008.

[16] I. Kerenidis and R. de Wolf. Exponential lower bound for 2-query locally decodable codes via a quantum argument. *Journal of Computer and System Sciences*, 69(3):395–420, 2004. Earlier version in STOC'03. quant-ph/0208062.

[17] P. B. Miltersen. On the cell probe complexity of polynomial evaluation. *Theor. Comput. Sci.*, 143(1):167–174, 1995.

[18] P. B. Miltersen. Cell probe complexity - a survey. Invited paper at *Advances in Data Structures* workshop. Available at Miltersen's homepage, 1999.

[19] V. S. Pless, W. C. Huffman, and R. A. Brualdi, editors. *Handbook of Coding Theory, Vol.1*. Elsevier Science, New York, NY, USA, 1998.

[20] R. de Wolf. Error-correcting data structures. In *Proceedings of 26th Annual Symposium on Theoretical Aspects of Computer Science (STACS'2009)*, pages 313–324, 2009. cs.DS/0802.1471.

[21] D. Woodruff. New lower bounds for general locally decodable codes. Technical report, ECCC Report TR07–006, 2006.

[22] S. Yekhanin. Towards 3-query locally decodable codes of subexponential length. *Journal of the ACM*, 55(1), 2008. Earlier version in STOC'07.

A Non-binary answer set

We prove Lemma 7, a version of Theorem 3 when the answer set A is non-binary. We first encode the $\ell|Q|$-bit string $\langle f(x,q)\rangle_{q\in Q}$ by an RLDC, and use the decoder of the RLDC to recover each of the ℓ bits of $f(x,q)$. Now it is possible that for each $q \in Q$, the decoder outputs some blank symbols $\perp$ for some of the bits of $f(x,q)$, and no query could be answered correctly. To circumvent this, we first encode each ℓ-bit string $f(x,q)$ with a good error-correcting code, then encode the entire string by the RLDC. Now if the decoder does not output too many errors or blank symbols among the bits of the error-correcting code for $f(x,q)$, we can recover it. We need a family of error-correcting codes with the following property, see e.g. page 668 in [19].

Fact 10. *For every $\delta \in (0,1/2)$ there exists $R \in (0,1)$ such that for all n, there exists a binary linear code of block length n, information length Rn, Hamming distance δn, such that the code can correct from e errors and s erasures, as long as $2e + s < \delta n$.*

Proof of Lemma 7. We only construct an error-correcting data structure with error probability $\varepsilon = \frac{1}{4}$. By a standard amplification technique (i.e., $O(\log(1/\varepsilon))$ repetitions) we can reduce the error probability to any other positive constant. Let $\mathcal{E}_{ecc} : \{0,1\}^{\ell} \to \{0,1\}^{\ell'}$ be an asymptotically good binary error-correcting code (from Fact 10), with $\ell' = O(\ell)$ and relative distance $\frac{3}{8}$, and decoder $\mathcal{D}_{ecc}$. By Theorem 3, there exist $c_0, \tau_0 > 0$ such that for every $\delta \le \tau_0$, there is a $(O(1), \delta, \frac{1}{32}, c_0\delta)$-relaxed locally decodable code (RLDC). Let $\mathcal{E}_0$ and $\mathcal{D}_0$ denote its encoder and decoder, respectively.

Encoding. We construct a data structure for f as follows. Define the encoder $\mathcal{E} : D \to \{0,1\}^N$, where $N = O((\ell' \cdot |Q|)^{1+\eta})$, as

$$\mathcal{E}(x) = \mathcal{E}_0\left(\langle\, \mathcal{E}_{ecc}(f(x,q))\,\rangle_{q\in Q}\right).$$

Decoding. Without loss of generality, we may impose an ordering on the set Q and identify each $q \in Q$ with an integer in $[Q]$.

The decoder $\mathcal{D}$, with input $q \in Q$ and oracle access to $w \in \{0,1\}^N$, does the following:

1. For each $j \in [\ell']$, let $r_j = \mathcal{D}_0^w\left((q-1)\ell' + j\right)$ and set $r = r_1 \ldots r_{\ell'} \in \{0,1,\perp\}^{\ell'}$.
2. If the number of blank symbols $\perp$ in r is at least $\frac{\ell'}{8}$, then output $\perp$. Else, output $\mathcal{D}_{ecc}(r)$.

Analysis. Fix $x \in D$ and $w \in \{0,1\}^N$ such that $\Delta(w, \mathcal{E}(x)) \le \delta N$, and $\delta \le \tau$, where τ is the minimum of τ_0 and $\lambda 2^{-6} c_0^{-1}$. We need to argue that the above encoding and decoding satisfies the four conditions of Definition 1. For Condition 1, since $\mathcal{D}_0$ makes $O(1)$ bit-probes and $\mathcal{D}$ runs this ℓ' times, $\mathcal{D}$ makes $O(\ell') = O(\ell)$ bit-probes into w.

We now show $\mathcal{D}$ satisfies Condition 2. Fix $q \in Q$. We want to show $\Pr[\mathcal{D}^w(q) \in \{f(x,q), \perp\}] \ge \frac{3}{4}$. By Theorem 3, for each $j \in [\ell']$, with probability at most $\frac{1}{32}$, $r_j = f(x,q)_j \oplus 1$. So on expectation, for at most a $\frac{1}{32}$-fraction of the indices j, $r_j = f(x,q)_j \oplus 1$. By Markov's inequality, with probability at least $\frac{3}{4}$, the number of indices j such that $r_j = f(x,q)_j \oplus 1$ is at most $\frac{\ell'}{8}$. If the number of $\perp$ symbols in r is at least $\frac{\ell'}{8}$ then $\mathcal{D}$ outputs $\perp$, so assume the number of $\perp$ symbols is less than $\frac{\ell'}{8}$. Those $\perp$'s are viewed as erasures in the codeword $\mathcal{E}_{ecc}(f(x,q))$. Since $\mathcal{E}_{ecc}$ has relative distance $\frac{3}{8}$, by Fact 10, $\mathcal{D}_{ecc}$ will correct these errors and erasures and output $f(x,q)$.

For Condition 3, we show there exists a large subset G of q's satisfying $\Pr[\mathcal{D}^w(q) = f(x,q)] \ge \frac{3}{4}$. Let $y = \langle\, \mathcal{E}_{ecc}(f(x,q))\,\rangle_{q\in Q}$, which is a $\ell'|Q|$-bit string. Call an index i in y *bad* if $\Pr[\mathcal{D}_0^w(i) = y_i] < \frac{3}{4}$. By Theorem 3, at most a $c_0\delta$-fraction of the indices in y are bad. We say that a query $q \in Q$ is *bad* if more than a $\frac{1}{64}$-fraction of the bits in $\mathcal{E}_{ecc}(f(x,q))$ are bad. By averaging, the fraction of bad queries in Q is at most $64c_0\delta$, which is at most λ by our choice of τ. We define G to be the set of $q \in Q$ that are not bad. Clearly $|G| \ge (1-\lambda)|Q|$.

Fix $q \in G$. On expectation (over the decoder's randomness), the fraction of indices in r such that $r_j \ne f(x,q)_j$ is at most $\frac{1}{64} + \frac{1}{32}$. Hence by Markov's inequality, with probability at least $\frac{3}{4}$, the fraction of indices in r such that $r_j \ne f(x,q)_j$ is at most $\frac{3}{16}$. Thus, by Fact 10, $\mathcal{D}_{ecc}(r)$ will recover from these errors and erasures and output $f(x,q)$.

Finally, Condition 4 follows since the pair $(\mathcal{E}_0, \mathcal{D}_0)$ satisfies Condition 4, finishing the proof. □

B CRT codes

In this section we explain how Theorem 8 follows from known facts. In [12], Goldreich, Ron, and Sudan designed a unique decoding algorithm for CRT code.

Theorem 11 (from [12])**.** *Given a CRT Code with basis $p_1 < \ldots < p_N$ and rate K/N, there exists a polynomial-time algorithm that can correct up to $\frac{\log p_1}{\log p_1 + \log p_N}(N-K)$ errors.*

By choosing the primes appropriately, we can establish Theorem 8. In particular, the following well-known estimate, essentially a consequence of the Prime Number Theorem, is useful. See for instance Theorem 4.7 in [1] for more details.

Fact 12. *For an integer $\ell > 0$, the ℓth prime (denoted q_ℓ) satisfies $\frac{1}{6}\ell \log \ell < q_\ell < 13\ell \log \ell$.*

Proof of Theorem 8. Let $K = \lfloor \frac{12 \log T}{\log \log T} \rfloor$ and q_ℓ denote the ℓ-th prime. By Fact 12, $q_K > \frac{1}{6} K \log K > \log T$ and $q_{3K-1} < 39K \log 3K < 500 \log T$. Also, notice that $\prod_{i=K}^{2K-1} q_i > q_K^K > (\log T)^{\frac{\log T}{\log \log T}} = T$. Thus, by Definition 2, the CRT code with basis $q_K < \ldots < q_{2K-1} < \ldots < q_{3K-1}$ and message space $\mathbb{Z}_T$, has rate at most $\frac{K}{2K} = \frac{1}{2}$. Lastly, by Theorem 11, the code can correct a fraction $\frac{1}{4} - O(\frac{1}{\log \log T})$ of errors. $\square$