New Locally Decodable Codes and Private Information Retrieval Schemes

Sergey Yekhanin
MIT
yekhanin@mit.edu

Abstract

A q query Locally Decodable Code (LDC) encodes an n-bit message x as an N-bit codeword $C(x)$, such that one can probabilistically recover any bit x_i of the message by querying only q bits of the codeword $C(x)$, even after some constant fraction of codeword bits has been corrupted.

We give new constructions of three query LDCs of vastly shorter length than that of previous constructions. Specifically, given any Mersenne prime $p = 2^t - 1$, we design three query LDCs of length $N = \exp\left(n^{1/t}\right)$, for every n. Based on the largest known Mersenne prime, this translates to a length of less than $\exp\left(n^{10^{-7}}\right)$, compared to $\exp\left(n^{1/2}\right)$ in the previous constructions. It has often been conjectured that there are infinitely many Mersenne primes. Under this conjecture, our constructions yield three query locally decodable codes of length $N = \exp\left(n^{O\left(\frac{1}{\log\log n}\right)}\right)$ for infinitely many n.

We also obtain analogous improvements for Private Information Retrieval (PIR) schemes. We give 3-server PIR schemes with communication complexity of $O\left(n^{10^{-7}}\right)$ to access an n-bit database, compared to the previous best scheme with complexity $O(n^{1/5.25})$. Assuming again that there are infinitely many Mersenne primes, we get 3-server PIR schemes of communication complexity $n^{O\left(\frac{1}{\log\log n}\right)}$ for infinitely many n.

Previous families of LDCs and PIR schemes were based on the properties of low-degree multivariate polynomials over finite fields. Our constructions are completely different and are obtained by constructing a large number of vectors in a small dimensional vector space whose inner products are restricted to lie in an algebraically nice set.

1 Introduction

Classical error-correcting codes allow one to encode an n bit string x into in N bit codeword $C(x)$, in such a way that x can still be recovered even if $C(x)$ gets corrupted in a number of coordinates. For instance, codewords $C(x)$ of length $N = O(n)$ already suffice to correct errors in up to δN locations of $C(x)$ for any constant $\delta < 1/2$. The disadvantage of classical error-correction is that one needs to consider all or most of the (corrupted) codeword to recover anything about x. Now suppose that one is only interested in recovering one or a few bits of x. In such case more efficient schemes are possible. Such schemes are known as locally decodable codes (LDCs). Locally decodable codes allow reconstruction of an arbitrary bit x_i, from looking only at q randomly chosen coordinates of $C(x)$, where q can be as small as 2. Locally decodable codes have found numerous applications in complexity theory and cryptography. See [22], [9] for a survey. Below is a slightly informal definition of LDCs:

A (q, δ, ϵ)-locally decodable code encodes n bit strings to N-bit codewords $C(x)$, such that for every $i \in [n]$,

the bit x_i can be recovered with probability $1-\epsilon$, by a randomized decoding procedure that makes only q queries, even if the codeword $C(x)$ is corrupted in up to δN locations.

One should think of $\delta > 0$ and $\epsilon < 1/2$ as constants. The main parameters of interest in LDCs are the length N and the query complexity q. Ideally we would like to have both of them as small as possible. The notion of locally decodable codes was explicitly discussed in various places in the early 1990s, most notably in [2, 21, 18]. Katz and Trevisan [14] were the first to provide a formal definition of LDCs and prove lower bounds on their length. Further work on locally decodable codes includes [4, 7, 17, 5, 13, 23]. The length of optimal 2 query LDCs was settled by Kerenidis and de Wolf in [13] and is $\exp(n)$. The length of optimal 3 query LDCs is unknown. The best upper bound prior to our work was $\exp\left(n^{1/2}\right)$ due to Beimel and Ishai [4], and the best lower bound is $\tilde{\Omega}(n^2)$ [13, 27]. For general (constant) q the best upper bound was $\exp\left(n^{O(\log\log q/(q\log q))}\right)$ due to Beimel et al. [5] and the best lower bound is $\tilde{\Omega}\left(n^{1+1/(\lceil q/2\rceil-1)}\right)$ [13, 27].

The current state of knowledge raises a natural question: Is the poor rate of known constructions an inherent property of locally decodable codes? Indeed, Gasarch [9, section 9] and Goldreich [10, conjecture 4.4] conjecture that the exponential dependence on n, i.e. the dependence of the form $N = \exp\left(n^{\Omega(1)}\right)$, is unavoidable for any constant number of queries. As our results suggest, such behavior may well not be inherent.

Our results

We give new families of locally decodable codes whose length is vastly shorter than that of previous constructions. We show that every Mersenne prime p (i.e. a prime of the form $p = 2^t - 1$) yields a family of three query locally decodable codes of length $\exp\left(n^{1/t}\right)$. The largest Mersenne prime known currently has $t = 32582657 > 10^7$. Substituting this prime into our theorem we conclude that for every n there exists a three query locally decodable code of length $\exp\left(n^{1/32582657}\right)$. It has often been conjectured that the number of Mersenne primes is infinite. In fact a much stronger conjecture regarding the density of Mersenne primes has been made by Lenstra, Pomerance and Wagstaff [24, 19, 25]. Using only the assumption that the number of Mersenne primes is infinite, our constructions yield three query locally decodable codes of length $N = \exp\left(n^{O\left(\frac{1}{\log\log n}\right)}\right)$ for infinitely many n.

1.1 Application to Private Information Retrieval

A q server *private information retrieval* (PIR) scheme allows a user to retrieve the i-th bit of an n-bit string x replicated between q servers while each server individually learns no information about i. The main parameter of interest in a PIR scheme is its communication complexity $C_q(n)$, namely the number of bits exchanged by the user and the servers. Below is a brief summary of known bounds for $C_q(n)$.

The best upper bound for $C_2(n)$ is $O(n^{1/3})$ due to Chor et al. [6]. The best upper bounds for larger values of q are $C_q(n) \leq n^{O(\log\log q/(q\log q))}$ due to Beimel et al. [5]. In particular [5] show that $C_3(n) \leq O\left(n^{1/5.25}\right)$, $C_4(n) \leq O\left(n^{1/7.87}\right)$ and $C_5(n) \leq O\left(n^{1/10.83}\right)$. On the lower bounds side the progress has been scarce. We list the known results for the two server case. The first nontrivial lower bound of $4\log n$ is due to Mann [16]. Later it was improved to $4.4\log n$ by Kerenidis and de Wolf [13]. The current record of $5\log n$ is due to Wehner and de Wolf [23].

Private information retrieval schemes are closely related to locally decodable codes. In particular, our constructions of LDCs yield three server private information retrieval schemes with small communication complexity. We show that every Mersenne prime $p = 2^t - 1$ yields $C_3(n) \leq O(n^{1/(t+1)})$. Instantiating this with the largest known Mersenne prime we get $C_3(n) \leq O(n^{1/32582658})$. Assuming that the number of Mersenne primes is infinite our bound goes further down to $n^{O\left(\frac{1}{\log\log n}\right)}$ for infinitely many values of n.

1.2 Our technique

All previously known constructions of locally decodable codes and private information retrieval schemes are (implicitly or explicitly) centered around the idea of representing message x by an evaluation of a certain low degree polynomial over a finite field. Our constructions take a completely different approach. We start by reducing the problem of constructing locally decodable codes to the problem of designing certain families of sets with restricted intersections. We use elementary algebra over finite fields to design such families.

The heart of our construction is design of a set $S \subseteq \mathbb{F}_p^*$ for a prime p that simultaneously satisfies two properties: (1) There exist two large sequences of vectors $u_1, \ldots, u_n$, $v_1, \ldots, v_n$ in some low dimensional space $\mathbb{F}_p^m$, such that the dot products $(u_i, v_i) = 0$ for all i, and the dot products $(u_j, v_i) \in S$ for all $i \neq j$. We refer to this property as the combinatorial niceness of S; (2) For a small integer q there exists a q sparse polynomial $\phi(x) \in \mathbb{F}_2[x]$ such that the common GCD of all polynomials of the form $\phi(x^\beta),\ \beta \in S$ and the polynomial $x^p - 1$ is non-trivial. We refer to this property as the algebraic niceness of S. Our notion of combinatorial niceness is related to the notion of set families with restricted intersections in [3].

Our construction of locally decodable codes thus comes in three steps: First we show that a set S exhibiting both combinatorial and algebraic niceness leads to good locally decodable codes. In particular the length n of the sequences $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$ corresponds to the number of message bits we can encode, while the length of the codewords we build is $N = p^m$. So the longer the sequence and the smaller the dimension the better. The query complexity of our codes is given by the parameter q from the definition of algebraic niceness of S. This step of our construction is quite general and applies to vectors $u_1, \ldots, v_n$ and subsets S over any field. It leads us to the task of identifying good sets that are both combinatorially and algebraically nice, and these tasks narrow our choice of fields. As our second step we focus on combinatorial niceness. In general big sets tend to be "nicer" (allow longer sequences) than small ones. We show that every multiplicative subgroup of a prime field is combinatorially as nice as its cardinality would allow. This still leaves us with a variety of fields and subsets to work with. Finally as the last step we attempt to understand the algebraic niceness of sets. We focus on the very narrow case of Mersenne primes p and the subgroup generated by the element 2 in $\mathbb{F}_p^*$. We manage to show that this subgroup is nice enough to get 3-query locally decodable codes, leading to our final result.

1.3 Outline

In section 3 we formally define locally decodable codes and introduce certain combinatorial objects that we call regular intersecting families. Those objects later serve as our tool to construct LDCs. In section 4 we present a linear algebraic construction of a regular intersecting family that yields locally decodable codes with good (although, not the best known) parameters. The notions of combinatorial and algebraic niceness of sets are used implicitly in this section. Our main construction in section 5 builds upon the construction of section 4. We formally introduce combinatorial and algebraic niceness and show how the interplay between these two notions yields new LDCs. The last subsection of section 5 and section 6 contain our main results for LDCs and private information retrieval schemes.

2 Notation

We use the following standard mathematical notation:

- $[s] = \{1, \ldots, s\}$;
- $\mathbb{F}_q$ is a finite field of q elements;
- $\mathbb{F}_q^*$ is the multiplicative group of $\mathbb{F}_q$;

- $d_H(x, y)$ denotes the Hamming distance between binary vectors x and y;
- (u, v) stands for the dot product of vectors u and v.
- For a linear space $L \subseteq \mathbb{F}_2^m$, $L^\perp$ denotes the *dual* space. That is, $L^\perp = \{u \in \mathbb{F}_2^m \mid \forall v \in L, (u, v) = 0\}$.

3 A combinatorial approach to locally decodable codes

In this section we formally define locally decodable codes and introduce certain combinatorial objects that we call *regular intersecting families* of sets. We show that regular intersecting families of sets yield LDCs.

Definition 1 *A binary code $C : \{0,1\}^n \to \{0,1\}^N$ is said to be (q, δ, ϵ)-locally decodable if there exists a randomized decoding algorithm $\mathcal{A}$ such that*

1. *For all $x \in \{0,1\}^n$, $i \in [n]$ and $y \in \{0,1\}^N$ such that $d_H(C(x), y) \leq \delta N$: Pr$[A^y(i) = x_i] \geq 1 - \epsilon$,*[1] *where the probability is taken over the random coin tosses of the algorithm $\mathcal{A}$.*
2. *$\mathcal{A}$ makes at most q queries to y.*

A locally decodable code is called linear if C is a linear transformation over $\mathbb{F}_2$. Our constructions of locally decodable codes are linear. They are obtained by viewing the basis elements of the code and the decoding sets of the code as specifying a set system (where a vector corresponds to the set of coordinates on which it is non-zero), with some special intersection properties. We define these properties next. Let N, R and n be positive integers. Consider the set $[N]$. For $i \in [n]$, $r \in [R]$ let T_i and Q_{ir}, be subsets of $[N]$.

Definition 2 *We say that subsets T_i and Q_{ir} form a (q, n, N, R, s) regular intersecting family if the following conditions are satisfied:*

1. *q is odd;*
2. *For all $i \in [n]$, $|T_i| = s$;*
3. *For all $i \in [n]$ and $r \in [R]$, $|Q_{ir}| = q$;*
4. *For all $i \in [n]$ and $r \in [R]$, $Q_{ir} \subseteq T_i$;*
5. *For all $i \in [n]$ and $w \in T_i$, $|\{r \in [R] \mid w \in Q_{ir}\}| = (Rq)/s$, (i.e. T_i is uniformly covered by the sets Q_{ir});*
6. *For all $i, j \in [n]$ and $r \in [R]$ such that $i \neq j$, $|Q_{ir} \cap T_j| \equiv 0 \mod (2)$.*

The following proposition shows that regular intersecting families imply locally decodable codes.

Proposition 3 *A (q, n, N, R, s) regular intersecting family yields a binary linear code encoding n bits to N bits that is $(q, \delta, \delta Nq/s)$ locally decodable for all δ.*

Proof: For a set $S \subseteq [N]$ let $I(S) \in \{0,1\}^N$ denote its incidence vector. Formally, for $w \in [N]$ we set $I(S)_w = 1$ if $w \in S$; and $I(S)_w = 0$ otherwise. We define linear code C via its generator matrix $G \in \{0,1\}^{n \times N}$. For $i \in [n]$, we set the i-th row of G to be the incidence vector of the set T_i. Below is the description of the decoding algorithm $\mathcal{A}$. Given oracle access to y and input $i \in [n]$, $\mathcal{A}$

[1] We remark that many earlier papers about LDCs used the parameter ϵ in a different way. They required Pr$[A^y(i) = x_i] \geq 1/2 + \epsilon$, rather then Pr$[A^y(i) = x_i] \geq 1 - \epsilon$. We choose to break with this tradition.

1. picks $r \in [R]$ uniformly at random;

2. outputs the dot product $(y, I(Q_{ir}))$ over $\mathbb{F}_2$.

Note that since $|Q_{ir}| = q$, $\mathcal{A}$ needs only q queries into y to compute the dot product. It is easy to verify that the decoding is correct if $\mathcal{A}$ picks $r \in [R]$ such that all bits of xG in locations $h \in Q_{ir}$ are not corrupted:

$$(xG, I(Q_{ir})) = \sum_{j=1}^{n} x_j (I(T_j), I(Q_{ir})) = x_i (I(T_i), I(Q_{ir})) = x_i. \quad (1)$$

The second equality in formula (1) follows from part 6 of definition 2 and the last equality follows from parts 1,3 and 4 of definition 2. Now assume that up to δN bits of the encoding xG have been corrupted. Part 5 of definition 2 implies that there are at most $(\delta NRq)/s$ sets Q_{ir} that contain at least one corrupted location. Thus with probability at least $1 - (\delta Nq)/s$ $\mathcal{A}$ outputs the correct value. ■

4 Basic construction

In this section we present our basic construction of regular intersecting families that yields q-query locally decodable codes of length $\exp\left(n^{1/(q-1)}\right)$ for prime values of $q \geq 3$. We choose sets T_i to be unions of cosets of certain hyperplanes and sets Q_{ir} to be lines. We argue the intersection properties based on elementary linear algebra. Let p be an odd prime and $m \geq p - 1$ be an integer.

Lemma 4 *Let $n = \binom{m}{p-1}$. There exist two families of vectors $\{u_1, \ldots, u_n\}$ and $\{v_1, \ldots, v_n\}$ in $\mathbb{F}_p^m$, such that*

- *For all $i \in [n]$, $(u_i, v_i) = 0$;*
- *For all $i, j \in [n]$ such that $i \neq j$, $(u_j, v_i) \neq 0$.*

Proof: Let $e \in \mathbb{F}_p^m$ be the vector that contains 1's in all the coordinates. We set vectors u_i to be incidence vectors of all possible $\binom{m}{p-1}$ subsets of $[m]$ of cardinality $(p-1)$. For every $i \in [n]$ we set $v_i = e - u_i$. It is straightforward to verify that this family satisfies the condition of the lemma. ■

Now we are ready to present our regular intersecting family. Set $N = p^m$ and $n = \binom{m}{p-1}$. Assume some bijection between the set $[N]$ and the space $\mathbb{F}_p^m$. For $i \in [n]$ set

$$T_i = \left\{x \in \mathbb{F}_p^m \mid (u_i, x) \in \mathbb{F}_p^*\right\}.$$

Set $R = s = (p-1) \cdot p^{m-1}$. For each $i \in [n]$ assume some bijection between points of T_i and elements of $[R]$. For $i \in [n]$ and $r \in [R]$ let w_{ir} be the r-th point of T_i. Set

$$Q_{ir} = \{w_{ir} + \lambda v_i \mid \lambda \in \mathbb{F}_p\}.^2$$

Lemma 5 *For $i \in [n]$ and $r \in [R]$ sets T_i and Q_{ir} form a (p, n, N, R, s) regular intersecting family.*

Proof: We simply need to verify that all 6 conditions listed in definition 2 are satisfied.

1. Condition 1 is trivial.

[2] Note that sets Q_{ir} are not all distinct.

2. Condition 2 is trivial.

3. Condition 3 is trivial.

4. Fix $i \in [n]$ and $r \in [R]$. Given that $(u_i, w_{ir}) \in \mathbb{F}_p^*$ let us show that $Q_{ir} \subseteq T_i$. By lemma 4 $(u_i, v_i) = 0$. Thus for every $\lambda \in \mathbb{F}_p : (u_i, w_{ir} + \lambda v_i) = (u_i, w_{ir})$. Condition 4 follows.

5. Fix $i \in [n]$ and $w \in T_i$. Note that

$$|\{r \in [R] \mid w \in Q_{ir}\}| = |\{w_{ir} \in T_i \mid \exists \lambda \in \mathbb{F}_p, w = w_{ir} + \lambda v_i\}| =$$

$$|\{w_{ir} \in T_i \mid \exists \lambda \in \mathbb{F}_p, w_{ir} = w - \lambda v_i\}| = p.$$

   It remains to notice that $Rp/s = p$. Condition 5 follows.

6. Fix $i, j \in [n]$ and $r \in [R]$ such that $i \neq j$. Note that

$$|Q_{ir} \cap T_j| = \left|\left\{\lambda \in \mathbb{F}_p \mid \ (u_j, w_{ir} + \lambda v_i) \in \mathbb{F}_p^*\right\}\right| = \left|\left\{\lambda \in \mathbb{F}_p \mid \ ((u_j, w_{ir}) + \lambda (u_j, v_i)) \in \mathbb{F}_p^*\right\}\right| = p - 1.$$

   The last equality follows from the fact that $(u_j, v_i) \neq 0$, and therefore the univariate linear function $(u_j, w_{ir}) + \lambda (u_j, v_i)$ takes every value in $\mathbb{F}_p$ exactly once. It remains to notice that $p - 1$ is even. Condition 6 follows.

■

Combining lemma 5 and proposition 3 we get

Corollary 6 *Let p be an odd prime and $m \geq p - 1$ be an integer. There exists a binary linear code encoding $\binom{m}{p-1}$ bits to p^m bits that is $\left(p, \delta, \delta p^2/(p-1)\right)$ locally decodable for all δ.*

It is now easy to convert the above result into a dense family (i.e., one that has a code for every message length n, as opposed to infinitely many n's) of p-query LDCs of length $\exp\left(n^{1/(p-1)}\right)$.

Theorem 7 *Let p be a fixed odd prime. For every positive integer n there exists a code of length $\exp\left(n^{1/(p-1)}\right)$ that is $\left(p, \delta, \delta p^2/(p-1)\right)$ locally decodable for all δ.*

Proof: Given n, choose m to be the smallest integer such that $n \leq \binom{m}{p-1}$. Set $n' = \binom{m}{p-1}$. It is easy to verify that if n is sufficiently large we have $n' \leq 2n$. Given a message x of length n, we pad it with zeros to length n' and use the code from corollary 6 encoding x with a codeword of length $p^m = \exp\left(n^{1/(p-1)}\right)$. ■

5 Main construction

In the previous section we presented our basic linear algebraic construction of regular intersecting families. We chose sets T_i to be unions of cosets of certain hyperplanes. We chose sets Q_{ir} to be lines. The high-level idea behind our main construction, is to reduce the number of codeword locations queried by choosing sets Q_{ir} to be *proper subsets of lines* rather than whole lines. Before we proceed to our main construction we introduce two central technical concepts of our paper, that of *combinatorial* and *algebraic niceness.* Let p be an odd prime.

Definition 8 *A set $S \subseteq \mathbb{F}_p^*$ is called (m, n) combinatorially nice if there exist two families of vectors $\{u_1, \ldots, u_n\}$ and $\{v_1, \ldots, v_n\}$ in $\mathbb{F}_p^m$, such that*

- *For all* $i \in [n]$, $(u_i, v_i) = 0$;
- *For all* $i, j \in [n]$ *such that* $i \neq j$, $(u_j, v_i) \in S$.

Remark 9 Note that in lemma 4 we established that the set $S = \mathbb{F}_p^*$ is $\left(m, \binom{m}{p-1}\right)$ combinatorially nice for every integer $m \geq p-1$.

Definition 10 *A set* $S \subseteq \mathbb{F}_p^*$ *is called* q *algebraically nice if* q *is odd and there exist two sets* $S_0, S_1 \subseteq \mathbb{F}_p$ *such that*

- S_0 *is not empty;*
- $|S_1| = q$;
- *For all* $\alpha \in \mathbb{F}_p$ *and* $\beta \in S : |S_0 \cap (\alpha + \beta S_1)| \equiv 0 \mod (2)$.

Remark 11 It is easy to verify that the set $S = \mathbb{F}_p^*$ is p algebraically nice. Simply pick $S_1 = \mathbb{F}_p$ and $S_0 = \mathbb{F}_p^*$.

5.1 Removing points from lines

The next proposition shows how an interplay between combinatorial and algebraic niceness yields regular intersecting families. It is the core of our construction.

Proposition 12 *Assume* $S \subseteq \mathbb{F}_p^*$ *is simultaneously* (m, n) *combinatorially nice and* q *algebraically nice. Let* S_0 *and* S_1 *be the sets from the definition of algebraic niceness. The set* S *yields a* $\left(q, n, p^m, |S_0|p^{m-1}, |S_0|p^{m-1}\right)$ *regular intersecting family.*

Proof: For $i \in [n]$ let u_i, v_i be the vectors from the definition of combinatorial niceness. Set $N = p^m$ and $R = s = |S_0|p^{m-1}$. Assume a bijection between $[N]$ and $\mathbb{F}_p^m$. For all $i \in [n]$ set

$$T_i = \left\{x \in \mathbb{F}_p^m \mid (u_i, x) \in S_0\right\}.$$

For each $i \in [n]$ assume some bijection between $[R]$ and T_i. Let w_{ir} denote the r-th point of T_i. Set

$$Q_{ir} = \{w_{ir} + \lambda v_i \mid \lambda \in S_1\}.$$

It remains to verify that all 6 conditions listed in definition 2 are satisfied.

1. Condition 1 is trivial.
2. Condition 2 is trivial.
3. Condition 3 is trivial.
4. Fix $i \in [n]$ and $r \in [R]$. Given that $(u_i, w_{ir}) \in S_0$ let us show that $Q_{ir} \subseteq T_i$. Definition 8 implies that $(u_i, v_i) = 0$. Thus for every $\lambda \in S_1 : (u_i, w_{ir} + \lambda v_i) = (u_i, w_{ir})$. Condition 4 follows.
5. Fix $i \in [n]$ and $w \in T_i$. Note that

$$|\{r \in [R] \mid w \in Q_{ir}\}| = |\{w_{ir} \in T_i \mid \exists \lambda \in S_1, w = w_{ir} + \lambda v_i\}| =$$

$$|\{w_{ir} \in T_i \mid \exists \lambda \in S_1, w_{ir} = w - \lambda v_i\}| = |S_1| = q.$$

   It remains to notice that $Rq/s = q$. Condition 5 follows.

6. Fix $i, j \in [n]$ and $r \in [R]$ such that $i \neq j$. Note that

$$|Q_{ir} \cap T_j| = |\{\lambda \in S_1 \mid (u_j, w_{ir} + \lambda v_i) \in S_0\}| =$$

$$|\{\lambda \in S_1 \mid ((u_j, w_{ir}) + \lambda(u_j, v_i)) \in S_0\}| = |S_0 \cap ((u_j, w_{ir}) + (u_j, v_i)S_1)| \equiv 0 \mod (2).$$

The last equality follows from the fact that $(u_j, v_i) \in S$, and definition 10. Condition 6 follows.

■

Observe that one can derive a regular intersecting family with parameters from lemma 5 using proposition 12 in combination with remarks 9 and 11.

5.2 On combinatorially nice subsets of $\mathbb{F}_p^*$

For $w \in \mathbb{F}_p^m$ and a positive integer l, let $w^{\otimes l} \in \mathbb{F}_p^{ml}$ denote the l-th tensor power of w. Coordinates of $w^{\otimes l}$ are labelled by all possible sequences in $[m]^l$ and $w^{\otimes l}_{i_1,\ldots,i_l} = \prod_{j=1}^{l} w_{i_j}$. The goal of this section is to establish the following

Lemma 13 *Let p be an odd prime and $m \geq p-1$ be an integer. Suppose S is a subgroup of $\mathbb{F}_p^*$; then S is $\left(\binom{m-1+\frac{p-1}{|S|}}{\frac{p-1}{|S|}}, \binom{m}{p-1}\right)$ combinatorially nice.*

Proof: Let $n = \binom{m}{p-1}$. For $i \in [n]$ let vectors u_i'' and v_i'' in $\mathbb{F}_p^m$ be the same as vectors u_i, v_i in the proof of lemma 4, i.e. vectors u_i'' are incidence vectors of all possible subsets of $[m]$ of cardinality $(p-1)$ and vectors v_i'' are their complements. Recall that

- For all $i \in [n]$, $(u_i'', v_i'') = 0$;
- For all $i, j \in [n]$ such that $i \neq j$, $(u_j'', v_i'') \neq 0$.

Let l be a positive integer and u, v be vectors in $\mathbb{F}_p^m$. Observe that

$$\left(u^{\otimes l}, v^{\otimes l}\right) = \sum_{(i_1,\ldots,i_l)\in[m]^l} \left(\prod_{j=1}^{l} u_{i_j} \prod_{j=1}^{l} v_{i_j}\right) = \sum_{(i_1,\ldots,i_l)\in[m]^l} \left(\prod_{j=1}^{l} u_{i_j} v_{i_j}\right) = \left(\sum_{i_1\in[m]} u_{i_1} v_{i_1}\right) \ldots \left(\sum_{i_l\in[m]} u_{i_l} v_{i_l}\right) = (u, v)^l. \tag{2}$$

Let $l = (p-1)/|S|$. For $i \in [n]$ set $u_i' = u_i''^{\otimes l}$ and $v_i' = v_i''^{\otimes l}$. Formula (2) and cyclicity of $\mathbb{F}_p^*$ yield

- For all $i \in [n]$, $(u_i', v_i') = 0$;
- For all $i, j \in [n]$ such that $i \neq j$, $(u_j', v_i') \in S$.

Note that vectors u_i' and v_i' are $m^{\frac{p-1}{|S|}}$ long. Therefore at this point we have already shown that the set S is $\left(m^{\frac{p-1}{|S|}}, \binom{m}{p-1}\right)$ combinatorially nice.

Let w be an arbitrary vector in $\mathbb{F}_p^m$. Note that the value of $w^{\otimes l}_{i_1,\ldots,i_l}$ depends on the *multi-set* $\{i_1, \ldots, i_l\}$ rather than the sequence $i_1, \ldots, i_l$. Thus many coordinates of $w^{\otimes l}$ contain identical (and therefore redundant) values.

We are going to reduce the length of vectors u'_i and v'_i using this observation. Let $F(m,l)$ denote the family of all multi-subsets of $[m]$ of cardinality l. Note that $|F(m,l)| = \binom{m-1+l}{l}$. For a multi-set $\sigma \in F(m,l)$ let $c(\sigma)$ denote the number of sequences in $[m]^l$ that represent σ. Now we are ready to define vectors u_i and v_i in $\mathbb{F}_p^{|F(m,l)|}$. Coordinates of vectors u_i and v_i are labelled by multi-sets $\sigma \in F(m,l)$. For all $i \in [n]$ and $\sigma \in F(m,l)$ we set

$$(u_i)_\sigma = c(\sigma)(u'_i)_\sigma \text{ and } (v_i)_\sigma = (v'_i)_\sigma.$$

It is easy to verify that for all $i,j \in [n]$, $(u_j, v_i) = \left(u'_j, v'_i\right)$. Combining this observation with the properties of vectors u'_i and v'_i that were established earlier, we conclude that the set S is $\left(\binom{m-1+\frac{p-1}{|S|}}{\frac{p-1}{|S|}}, \binom{m}{p-1}\right)$ combinatorially nice. ■

5.3 On algebraically nice subsets of $\mathbb{F}_p^*$

In this section we construct 3-algebraically nice subsets of $\mathbb{F}_p^*$, for primes p that have the form $p = 2^t - 1$. Such primes are known as *Mersenne* primes. Consider a natural one to one correspondence between subsets S_1 of $\mathbb{F}_p$ and polynomials $\phi_{S_1}(x)$ in the ring $\mathbb{F}_2[x]/(x^p - 1)$:

$$\phi_{S_1}(x) = \sum_{s \in S_1} x^s.$$

It is immediate to verify that for all sets $S_1 \subseteq \mathbb{F}_p$ and all $\alpha, \beta \in \mathbb{F}_p$, such that $\beta \neq 0$:

$$\phi_{\alpha+\beta S_1}(x) = x^\alpha \phi_{S_1}(x^\beta). \tag{3}$$

Lemma 14 *Let $p = 2^t - 1$ be a Mersenne prime. The set $S = \{1, 2, 4, 8, \ldots, 2^{t-1}\} \subseteq \mathbb{F}_p^*$ is three algebraically nice.*

Proof: Observe that the polynomial $x^p - 1 = x^{2^t - 1} - 1$ splits into distinct linear factors in the finite field $\mathbb{F}_{2^t}$. Clearly, every non-zero element of $\mathbb{F}_{2^t}$ is a root of $x^p - 1$. Let g be a primitive element of $\mathbb{F}_{2^t}$. Fix γ such that $1 + g + g^\gamma = 0$. Set $S_1 = \{0, 1, \gamma\}$.

Let α be a variable ranging over $\mathbb{F}_p$ and β be a variable ranging over S. We are going to argue the existence of a set S_0 that has even intersections with all sets of the form $\alpha + \beta S_1$, by showing that all polynomials $\phi_{\alpha+\beta S_1}$ belong to a certain linear space $L \in \mathbb{F}_2[x]/(x^p - 1)$ of dimension less than p. In this case any nonempty set $T \subseteq \mathbb{F}_p$ such that $\phi_T \in L^\perp$ can be used as the set S_0. Let $\tau(x) = GCD(x^p - 1, \phi_{S_1}(x))$. Note that $\tau(x) \neq 1$ since g is a common root of $x^p - 1$ and $1 + x + x^\gamma$. Let L be the space of polynomials in $\mathbb{F}_2[x]/(x^p - 1)$ that are multiples of $\tau(x)$. Clearly, $\dim L = p - \deg \tau$. Fix some $\alpha \in \mathbb{F}_p$ and $\beta \in S$. Let us prove that $\phi_{\alpha+\beta S_1}(x)$ is in L :

$$\phi_{\alpha+\beta S_1}(x) = x^\alpha \phi_{S_1}(x^\beta) = x^\alpha (\phi_{S_1}(x))^\beta.$$

The last identity above follows from the fact that for any polynomial $f \in \mathbb{F}_2[x]$ and any integer i : $f(x^{2^i}) = (f(x))^{2^i}$ and our choice of the set S. ■

Parameters of a regular intersecting family that one gets by applying proposition 12 to a certain (nice) set S depend on the size of the set S_0 from the definition of algebraic niceness of S. The next lemma shows that one can always pick the set S_0 to be large.

Lemma 15 *Let $S \subseteq \mathbb{F}_p^*$ be a q algebraically nice set. Let $S_0, S_1 \subseteq \mathbb{F}_p$ be sets from the definition of algebraic niceness of S. One can always redefine the set S_0 to satisfy $|S_0| \geq \lceil p/2 \rceil$.*

Proof: Let $L \subset \mathbb{F}_2[x]/(x^p - 1)$ be the linear space spanned by polynomials of the form $\phi_{\alpha+\beta S_1}(x)$, for $\alpha \in \mathbb{F}_p$ and $\beta \in S$. Clearly, the space L is closed under cyclic shifts. This implies that the space $L^\perp$ is also closed under cyclic shifts. Note that $L^\perp$ has positive dimension since $\phi_{S_0}(x) \in L^\perp$. The last two observations imply that $L^\perp$ has *full support*, i.e. for every coordinate i there exists a vector $\phi \in L^\perp$ such that $\phi_i \neq 0$. It is easy to verify that any linear subspace of $\mathbb{F}_2^p$ that has full support contains a vector of Hamming weight at least $\lceil p/2 \rceil$. Let $\phi_T(x) \in L^\perp$ be such a vector. Redefining the set S_0 to be the set T we conclude the proof. ■

5.4 Results

Let $p = 2^t - 1$ be a Mersenne prime. Note that the set $S = \{1, 2, 4, 8, \ldots, 2^{t-1}\}$ is a multiplicative subgroup of $\mathbb{F}_p^*$. Combining proposition 12 with lemmas 13, 14 and 15 we conclude

Lemma 16 *Let $p = 2^t - 1$ be a Mersenne prime and $m \geq p - 1$ be an integer. Let $m' = \binom{m-1+(p-1)/t}{(p-1)/t}$. For some integer $z \geq \lceil p/2 \rceil$ there exists a regular intersecting family with parameters*

$$\left(3, \binom{m}{p-1}, p^{m'}, zp^{m'-1}, zp^{m'-1}\right).$$

Combining lemma 16 with proposition 3 we obtain the key lemma of the paper

Lemma 17 *Let $p = 2^t - 1$ be a Mersenne prime and $m \geq p - 1$ be an integer. Let $m' = \binom{m-1+(p-1)/t}{(p-1)/t}$. There exists a binary linear code encoding $n = \binom{m}{p-1}$ bits to $p^{m'}$ bits that is $(3, \delta, 6\delta)$ locally decodable code for all δ.*

For every fixed Mersenne prime $p = 2^t - 1$ we get a family of 3-query LDCs of length $\exp\left(n^{1/t}\right)$. We omit the proof since its essentially identical to the proof of theorem 7.

Theorem 18 *Let $p = 2^t - 1$ be a fixed Mersenne prime. For every positive integer n there exists a code of length $\exp\left(n^{1/t}\right)$ that is $(3, \delta, 6\delta)$ locally decodable for all δ.*

Mersenne primes have been a popular object of study in number theory for the last few centuries. It is still unknown whether the number of Mersenne primes is infinite. There has been a large amount of effort and computational power invested in search for large Mersenne primes [26]. The largest currently known Mersenne prime is $p = 2^{32582657} - 1$. It was discovered by C. Cooper and S. Boone [8] on September 4, 2006. Plugging p into theorem 18 we get

Theorem 19 *For every positive integer n there exists a code of length $\exp\left(n^{1/32582657}\right)$ that is $(3, \delta, 6\delta)$ locally decodable for all δ.*

We are not aware of who was the first to conjecture that the number of Mersenne primes is infinite. Lenstra, Pomerance, and Wagstaff have made a much stronger conjecture [24], [19], [25]. Their conjecture claims that not only are there infinitely many Mersenne primes, but that the number of Mersenne primes with exponent less than t is asymptotically approximated by $e^\gamma \log_2(t)$, where γ is the Euler-Mascheroni constant. In case the number of Mersenne primes is infinite we get three query locally decodable codes of sub-exponential length.

Theorem 20 *Suppose that the number of Mersenne primes is infinite; then for infinitely many values of n there exists a code of length $\exp\left(n^{O\left(\frac{1}{\log\log n}\right)}\right)$ that is $(3, \delta, 6\delta)$ locally decodable for all δ.*

Proof: Given a Mersenne prime p, set $m = 2^p$. Substituting m and p into lemma 17 and making some basic manipulations we conclude that there exists a $(3, \delta, 6\delta)$ locally decodable code encoding $n = m^{\Theta(\log m)}$ bits to $N = \exp\left(m^{O\left(\frac{\log m}{\log\log m}\right)}\right)$ bits. An observation that $\log\log n = \Theta(\log\log m)$ completes the proof. ■

6 Application to Private Information Retrieval

We start with a formal definition of a three server PIR protocol. Let $x \in \{0,1\}^n$ be the database.

Definition 21 *A three server PIR protocol is a triplet of non-uniform algorithms $\mathcal{P} = (\mathcal{Q}, \mathcal{A}, \mathcal{C})$. We assume that each algorithm is given n as an advice. At the beginning of the protocol, the user $\mathcal{U}$ tosses random coins and obtains a random string r. Next $\mathcal{U}$ invokes $\mathcal{Q}(i,r)$ to generate a triple of queries (que_1, que_2, que_3). For $i \in [3]$, $\mathcal{U}$ sends que_i to $\mathcal{S}_i$. Each server $\mathcal{S}_j$ responds with an answer $ans_j = \mathcal{A}(j, x, que_j)$. (We can assume without loss of generality that servers are deterministic; hence, each answer is a function of a query and a database.) Finally, $\mathcal{U}$ computes its output by applying the reconstruction algorithm $\mathcal{C}(ans_1, ans_2, ans_3, i, r)$. A protocol as above should satisfy the following requirements:*

- **Correctness :** *For any $n,\ x \in \{0,1\}^n$ and $i \in [n]$, the user outputs the correct value of x_i with probability 1 (where the probability is over the random strings r).*
- **Privacy :** *Each server individually learns no information about i. To formalize this let $\mathcal{Q}_j$ denote the j-th output of $\mathcal{Q}$. We require that for $j = 1,2,3$ and any $n,\ i_1, i_2 \in [n]$ the distributions $\mathcal{Q}_j(i_1, r)$ and $\mathcal{Q}_j(i_2, r)$ are identical.*

There are known generic procedures [14] to convert q query LDCs into q server PIR schemes. However a simple application of such a procedure to our LDCs will either yield a PIR protocol with perfect privacy, but small probability of error, or a PIR protocol with perfect correctness and some slight privacy leakage. Fortunately, it is possible to achieve both perfect privacy and perfect correctness simultaneously via a specially designed reduction. We sketch the idea behind such a reduction.

The servers encode the database x with a three query locally decodable code C from lemma 17. We are going to use the notation from that lemma. Recall the the coordinates of $C(x)$ are in one to one correspondence with points in $\mathbb{F}_p^{m'}$. In order to decode x_i the user has to query three locations $\{w + \lambda v_i \mid \lambda \in S_1\}$ for some $w \in T_i$, where T_i is the union of certain cosets of the hyperplane $\left\{ y \in \mathbb{F}_p^{m'} \mid (u_i, y) = 0 \right\}$. Unlike the LDC setup in the PIR setup the user can not pick $w \in T_i$ uniformly at random and then query locations $\{w + \lambda v_i \mid \lambda \in S_1\}$ from three different servers, since in such case the servers would observe the uniform distribution on T_i rather than the uniform distribution on $\mathbb{F}_p^{m'}$. Here is our way to go around this problem.

Let $e \in \mathbb{F}_p^{m'}$ be the all-ones vector. Assume $m \neq 0 \mod (p)$. The definition of vectors u_i in lemma 13 implies that in such a case $(e, u_i) \neq 0$ for all $i \in [n]$. Thus for every $i \in [n]$ and every $w \in \mathbb{F}_p^{m'}$ there is some $\gamma \in \mathbb{F}_p$ such that $w + \gamma e \in T_i$. The user picks $w \in \mathbb{F}_p^{m'}$ uniformly at random and (simultaneously) asks p triples of queries of the from $\{w + \gamma e + \lambda v_i \mid \lambda \in S_1\}$ for all $\gamma \in \mathbb{F}_p$. For every triple the first query always goes to server 1, the second to server 2 and the last to server 3. It is easy to verify that in such case each server individually observes a uniform distribution independent of i, while the user always successfully reconstructs x_i from one of the triples. Our argument yields

Lemma 22 *Let $p = 2^t - 1$ be a Mersenne prime and $m \geq p-1$ be an integer such that $m \neq 0 \mod (p)$. Let $n = \binom{m}{p-1}$ and $m' = \binom{m-1+(p-1)/t}{(p-1)/t}$. There exists a three server PIR protocol with questions of length $pm' \log p$ and answers of length p that allows private retrieval of bits from databases of length n.*

The next theorem captures the asymptotic behavior of our PIR schemes for a fixed Mersenne prime p.

Theorem 23 *Let $p = 2^t - 1$ be a fixed Mersenne prime. For every positive integer n there exists a three server PIR protocol with questions of length $O\left(n^{1/t}\right)$ and answers of length p.*

Combing theorem 23 with a standard balancing technique from [6] and plugging in the value of the largest known Mersenne prime, we conclude

Theorem 24 *For every positive integer n there exists a three server PIR protocol with communication complexity of $O\left(n^{1/32582658}\right)$.*

Finally, under the assumption that the number of Mersenne primes is infinite we get

Theorem 25 *Suppose that the number of Mersenne primes is infinite; then for infinitely many values of n there exists a three server PIR protocol with communication complexity of $n^{O\left(\frac{1}{\log\log n}\right)}$.*

7 Conclusion

We presented a novel approach to constructing locally decodable codes and substantially improved the known upper bounds. However the gap between the upper and lower bounds for LDCs still remains very large. It might be the case the technique proposed in this paper has not yet been pushed to its limit and further improvements will be obtained in this way. In particular, proposition 12 generalizes to arbitrary finite fields (rather than just prime fields), and even finite commutative rings R. It may happen that a clever choice of a ring R and a subset $S \subseteq R$ that is simultaneously combinatorially and algebraically nice will yield shorter LDCs.

Acknowledgement

I am indebted to Madhu Sudan for expressing his optimism regarding viability of the approach taken in this paper at an early stage of the work. I would also like to thank him for many helpful in-depth technical discussions. Many thanks to Oded Goldreich, Nick Harvey, Kiran Kedlaya, Swastik Kopparty and David Woodruff for their valuable comments.

References

[1] A. Ambainis, "Upper bound on the communication complexity of private information retrieval," *Proc. of 32th ICALP, LNCS* 1256, pp. 401-407, 1997.

[2] L. Babai, L. Fortnow, L. Levin, and M. Szegedy, "Checking computations in polylogarithmic time,". In *Proc. of the 23th ACM Sym. on Theory of Computing (STOC)*, pp. 21-31, 1991.

[3] L. Babai, P. Frankl, "Linear Algebra Methods in Combinatorics." 1998.

[4] A. Beimel and Y. Ishai. "Information-Theoretic Private Information Retrieval: A Unified Construction," Technical Report TR01-15, *Electronic Colloquim on Computational Complexity* (ECCC), 2001. Extended abstract in: ICALP 2001, vol. 2076 of LNCS, pp. 89-98, 2001.

[5] A. Beimel, Y. Ishai, E. Kushilevitz, and J. F. Raymond. "Breaking the Barrier for Information-Theoretic Private Information Retrieval," In *Proc. of the 43rd IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 261-270, 2002.

[6] B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan. "Private information retrieval," In *Proc. of the 36rd IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 41-50, 1995. Also, in *Journal of the ACM*, 45, 1998.

[7] A. Deshpande, R. Jain, T. Kavitha, S. Lokam and J. Radhakrishnan, "Better lower bounds for locally decodable codes," In *Proc. of the 20th IEEE Computational Complexity Conference (CCC)*, pp. 184-193, 2002.

[8] Curtis Cooper, Steven Boone, http://www.mersenne.org/32582657.htm

[9] W. Gasarch, "A survey on private information retrieval," *The Bulletin of the EATCS,* 82:72-107, 2004.

[10] O. Goldreich, "Short Locally Decodable Codes and Proofs," Technical Report TR05-014, *Electronic Colloquim on Computational Complexity* (ECCC), 2005.

[11] O. Goldreich, H. Karloff, L. Schulman, L. Trevisan "Lower bounds for locally decodable codes and private information retrieval," In *Proc. of the 17th IEEE Computational Complexity Conference (CCC)*, pp. 175-183, 2002.

[12] Y. Ishai and E. Kushilevitz "Improved upper bounds on information-theoretic private information retrieval," In *Proc. of the 31th ACM Sym. on Theory of Computing (STOC)*, pp. 79-88, 1999.

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

[14] J. Katz and L. Trevisan, "On the efficeincy of local decoding procedures for error-correcting codes," In *Proc. of the 32th ACM Sym. on Theory of Computing (STOC)*, pp. 80-86, 2000.

[15] R. Lidl and H. Niederreiter, Finite Fields. Cambridge: Cambridge University Press, 1985.

[16] E. Mann, Private access to distributed information. Master's thesis, Technion - Israel Institute of Technology, Haifa, 1998.

[17] K. Obata, "Optimal lower bounds for 2-query locally decodable linear codes," In *Proc. of the 6th RANDOM*, vol. 2483 of Lecture Notes in Computer Science, pp. 39-50, 2002.

[18] A. Polishchuk and D. Spielman, "Nearly-linear size holographic proofs," In *Proc. of the 26th ACM Sym. on Theory of Computing (STOC)*, pp. 194-203, 1994.

[19] C. Pomerance, "Recent developments in primality testing," *Math. Intelligencer,* 3:3, pp. 97-105, (1980/81).

[20] A. Razborov and S. Yekhanin "An $\Omega(n^{1/3})$ Lower Bound for Bilinear Group Based Private Information Retrieval," In *Proc. of the 47rd IEEE Symposium on Foundations of Computer Science (FOCS)*, 2006.

[21] M. Sudan, Efficient Checking of Polynomials and Proofs and the Hardness of Approximation Problems. PhD thesis, University of California at Berkeley, 1992.

[22] L. Trevisan, "Some Applications of Coding Theory in Computational Complexity," *Quaderni di Matematica,* 13:347-424, 2004.

[23] S. Wehner and R. de Wolf, "Improved Lower Bounds for Locally Decodable Codes and Private Information Retrieval," In *Proc. of 32nd International Colloquium on Automata, Languages and Programming (ICALP'05)*, LNCS 3580, pp.1424-1436.

[24] Lenstra-Pomerance-Wagstaff conjecture. (2006, May 22). In Wikipedia, The Free Encyclopedia. Retrieved 00:18, October 3, 2006, from http://en.wikipedia.org/w/index.php?title=Lenstra-Pomerance-Wagstaff_conjecture&oldid=54506577

[25] S. Wagstaff, “Divisors of Mersenne numbers,” *Math. Comp.*, 40:161, pp. 385-397, 1983.

[26] Mersenne prime. (2006, October 1). In Wikipedia, The Free Encyclopedia. Retrieved 21:04, October 2, 2006, from http://en.wikipedia.org/w/index.php?title=Mersenne_prime&oldid=78877957

[27] D. Woodruff, “Some new lower bounds for general locally decodable codes,” Manuscript, 2006.

[28] D. Woodruff and S. Yekhanin, “A Geometric Approach to Information Theoretic Private Information Retrieval,” In *Proc. of the 20th IEEE Computational Complexity Conference (CCC)*, pp. 275-284, 2005.