

# Optimal Lower Bounds for 2-Query Locally Decodable Linear Codes

Kenji Obata <sup>\*</sup>

Computer Science Division  
University of California, Berkeley

**Abstract.** This paper presents essentially optimal lower bounds on the size of linear codes

$$\mathbf{C} : \{0, 1\}^n \rightarrow \{0, 1\}^m$$

which have the property that, for constants  $\delta, \epsilon > 0$ , any bit of the message can be recovered with probability  $\frac{1}{2} + \epsilon$  by an algorithm reading only 2 bits of a codeword corrupted in up to  $\delta m$  positions. Such codes are known to be applicable to, among other things, the construction and analysis of information-theoretically secure private information retrieval schemes. In this work, we show that  $m$  must be at least  $2^{\Omega(\frac{\delta}{1-2\epsilon}n)}$ . Our results extend work by Goldreich, Karloff, Schulman, and Trevisan [GKST02], which is based heavily on methods developed by Katz and Trevisan [KT00]. The key to our improved bounds is an analysis which bypasses an intermediate reduction used in both prior works. The resulting improvement in the efficiency of the overall analysis is sufficient to achieve a lower bound optimal within a constant factor in the exponent. A construction of a locally decodable linear code matching this bound is presented.

## 1 Introduction

This paper presents essentially optimal lower bounds on the size of linear codes

$$\mathbf{C} : \{0, 1\}^n \rightarrow \{0, 1\}^m$$

which have the property that, for constants  $\delta, \epsilon > 0$ , any bit of the message can be recovered with probability  $\frac{1}{2} + \epsilon$  by an algorithm reading only 2 bits of a codeword corrupted in up to  $\delta m$  positions. Such codes are known to be applicable to, among other things, the construction and analysis of information-theoretically secure private information retrieval schemes.

In this work, we show that  $m$  must be at least  $2^{\Omega(\frac{\delta}{1-2\epsilon}n)}$ . Our results extend work by Goldreich, Karloff, Schulman, and Trevisan [GKST02], who show that  $m$  must be at least  $2^{\Omega(\epsilon\delta n)}$ . Note that the prior bound does not grow arbitrarily large as the error probability of the decoder goes to zero ( $\epsilon \rightarrow \frac{1}{2}$ ), as intuitively it should; our new results have the correct qualitative behavior.

---

<sup>\*</sup> [kenjioba@eecs.berkeley.edu](mailto:kenjioba@eecs.berkeley.edu). Work supported by an NSF graduate fellowship.

The key to our improved bounds is an analysis which bypasses an intermediate reduction used in both prior works. The resulting improvement in the efficiency of the overall analysis is sufficient to achieve a lower bound optimal within a constant factor in the exponent. A construction of a locally decodable linear code matching this bound is presented.

Our work is structured as follows: In the remainder of this section, we briefly review the definitions and techniques employed in [KT00] and [GKST02]. In Section 2, we prove the main technical result of this paper, which establishes a relationship between the probability that an edge of a graph sampled from any distribution intersects any vertex-subset of a given size, and the size of a maximum matching in the graph. The analysis in this result seems independently interesting, and may be applicable in other contexts. In Section 3, we show how the combination of this result with the techniques of [GKST02] establishes lower bounds for this class of locally decodable codes, and present a construction of a family of these codes with size matching our bounds within a constant factor in the exponent.

### 1.1 Locally Decodable and Smooth Codes

Let  $\Sigma_1, \Sigma_2$  be arbitrary finite alphabets,  $\Sigma_i^n$  the set of strings of elements from  $\Sigma_i$  of length  $n$ , and for  $x, y \in \Sigma_i^n$ ,  $d(x, y)$  the number of positions  $i$  such that  $x_i \neq y_i$ .

**Definition 1 (Locally Decodable Code).** *For fixed constants  $\delta, \epsilon, q$ , a mapping*

$$\mathbf{C} : \Sigma_1^n \rightarrow \Sigma_2^m$$

*is a  $(q, \delta, \epsilon)$ -locally decodable code if there exists a probabilistic oracle machine  $A$  such that:*

- *$A$  makes at most  $q$  queries (without loss of generality,  $A$  makes exactly  $q$  queries).*
- *For every  $x \in \Sigma_1^n, y \in \Sigma_2^m$  with  $d(y, \mathbf{C}(x)) \leq \delta m$ , and  $i \in \{1, \dots, n\}$ ,*

$$\Pr [A^y(i) = x_i] \geq \frac{1}{|\Sigma_1|} + \epsilon$$

*where the probability is over the randomness of  $A$ .*

In this paper, we consider codes  $\mathbf{C}$  satisfying the above properties where, in addition,  $\Sigma_1, \Sigma_2$  are fields and  $\mathbf{C}$  is a *linear* mapping from  $\Sigma_1^n \rightarrow \Sigma_2^m$ . While all of our results are applicable to finite fields in general, and some to non-linear codes, we will for simplicity narrow our current discussion to linear codes on  $\mathbf{Z}_2$ . Also, while we have observed that our results are equally applicable to reconstruction algorithms making queries adaptively, we limit our comments in this abstract to algorithms making non-adaptive queries, and summarize some details of our proofs.

We begin by reviewing the definitions and techniques of [KT00] and [GKST02], which our results build upon.

It was observed in [KT00] that a locally decodable code should have the property that a decoding algorithm  $A$  reads from each location in the code word with roughly uniform probability. This motivated the following definition:

**Definition 2 (Smooth Code).** *For fixed constants  $c, \epsilon, q$ , a mapping*

$$\mathbf{C} : \Sigma_1^n \rightarrow \Sigma_2^m$$

*is a  $(q, c, \epsilon)$ -smooth code if there exists a probabilistic oracle machine  $A$  such that:*

- *$A$  makes at most  $q$  queries (without loss of generality,  $A$  makes exactly  $q$  queries).*
- *For every  $x \in \Sigma_1^n$  and  $i \in \{1, \dots, n\}$ ,*

$$\Pr \left[ A^{\mathbf{C}(x)}(i) = x_i \right] \geq \frac{1}{|\Sigma_1|} + \epsilon.$$

- *For every  $i \in \{1, \dots, n\}, j \in \{1, \dots, m\}$ , the probability that on input  $i$  machine  $A$  queries index  $j$  is at most  $\frac{c}{m}$ .*

Intuitively, if a code is insufficiently smooth, so that a particular small subset of indices is queried with too high a probability, then corrupting that subset causes the decoding algorithm to fail with too high a probability. Thus, a locally decodable code must have a certain smoothness. Specifically, [KT00] proved:

**Theorem 1.** *If  $\mathbf{C} : \Sigma_1^n \rightarrow \Sigma_2^m$  is a  $(q, \delta, \epsilon)$ -locally decodable code, then  $\mathbf{C}$  is also a  $(q, \frac{q}{\delta}, \epsilon)$ -smooth code.*

The lower bounds for linear locally decodable codes in [GKST02] are proved by establishing lower bounds for smooth codes. The result for locally decodable codes follows by application of Theorem 1.

Smooth codes are closely related to the concept of information-theoretically secure private information retrieval schemes introduced in [CGKS98]. Briefly, the idea in these constructions is to allow a user to retrieve a value stored in a database in such a way that the database server does not learn significant information about what value was queried. It is easy to see that, in the information-theoretic setting, achieving privacy in this sense with a single database server requires essentially that the entire database be transferred to the user on any query. [CGKS98] showed, however, that by using 2 (non-colluding) servers, one can achieve privacy in this sense with a single round of queries and communication complexity  $O(n^{1/3})$ . [KT00] observed that if one interprets the query bits sent to the databases as indexes into a 2-query decodable code, then the smoothness parameter of a code can be interpreted as a statistical indistinguishability condition in the corresponding retrieval scheme. In this way, one can construct

and analyze smooth codes, and therefore locally decodable codes, from private information retrieval schemes and vice versa. We refer the reader to [GKST02] for a detailed discussion.

The basic technique for proving lower bounds for smooth codes introduced in [KT00] and extended in [GKST02] is to study, for each  $i \in \{1, \dots, n\}$ , the *recovery graph*  $G_i$  defined on vertex set  $\{1, \dots, m\}$  where  $(q_1, q_2)$  is an edge of  $G_i$  iff for all  $x \in \{0, 1\}^n$ ,

$$\Pr \left[ A^{\mathbf{C}^{(x)}}(i) = x_i \mid A \text{ queries } (q_1, q_2) \right] > \frac{1}{2}.$$

Such edges are called *good edges*. Then, one shows a lower bound on the size of a maximum matching in the recovery graphs  $G_i$  which is a function of the smoothness parameter of  $\mathbf{C}$ :

**Lemma 1** ([KT00], [GKST02]). *If  $\mathbf{C}$  is a  $(2, c, \epsilon)$ -smooth code with recovery graphs  $\{G_i\}_i$  then, for every  $i$ ,  $G_i$  has a matching of size at least  $\frac{\epsilon m}{c}$ .*

For *linear* smooth codes, it is easy to see that an edge  $(q_1, q_2)$  can be good for  $x_i$  iff  $x_i$  is a linear combination of  $q_1, q_2$ . To simplify matters, one narrows the analysis to codes in which these linear combinations are non-trivial:

**Definition 3 (Non-Degenerate Code)**. *A linear code  $\mathbf{C}$  is non-degenerate if none of the entries in the range of  $\mathbf{C}$  is a scalar multiple of an input entry.*

We can assume non-degeneracy in smooth codes with only a constant factor modification in length and recovery parameters:

**Theorem 2** ([GKST02]). *For  $n > \frac{4c}{\epsilon}$ , let  $\mathbf{C} : \{0, 1\}^n \rightarrow \{0, 1\}^m$  be a  $(q, c, \epsilon)$ -smooth code. Then there exists a  $(q, c, \frac{\epsilon}{2})$ -smooth code  $\mathbf{C}' : \{0, 1\}^{n'} \rightarrow \{0, 1\}^{m'}$  with  $n' \geq \frac{n}{2}, m' \leq m$  in which for all  $i \in \{1, \dots, n'\}, j \in \{1, \dots, m'\}$ , the  $j$ -th bit of  $\mathbf{C}'(x)$  is not a scalar multiple of  $x_i$ .*

Putting the pieces together, we have that a non-degenerate  $(2, c, \epsilon)$ -smooth code has for every  $i \in \{1, \dots, n\}$  a recovery graph  $G_i$  containing a matching of size at least  $\frac{\epsilon m}{c}$ , and for each of the edges  $(q_1, q_2)$  in this matching,  $x_i$  is in the span of  $q_1, q_2$ , but is not a scalar multiple of  $q_1$  or  $q_2$ . Thus, the preconditions for the following key result of [GKST02] are satisfied:

**Lemma 2**. *Let  $q_1, \dots, q_m$  be linear functions on  $x_1, \dots, x_n \in \{0, 1\}^n$  such that for every  $i \in \{1, \dots, n\}$  there is a set  $M_i$  of at least  $\gamma m$  disjoint pairs of indices  $j_1, j_2$  such that  $x_i = q_{j_1} + q_{j_2}$ . Then  $m \geq 2^{\gamma n}$ .*

Composing this with the degenerate to non-degenerate reduction of Theorem 2, we have:

**Theorem 3** ([GKST02]). *Let  $\mathbf{C} : \{0, 1\}^n \rightarrow \{0, 1\}^m$  be a  $(2, c, \epsilon)$ -smooth linear code. Then  $m \geq 2^{\frac{\epsilon n}{4c}}$ .*

Finally, composing this with the locally decodable to smooth reduction, this says:

**Theorem 4 ([GKST02]).** *Let  $C : \{0, 1\}^n \rightarrow \{0, 1\}^m$  be a  $(2, \delta, \epsilon)$ -locally decodable linear code. Then  $m \geq 2^{\frac{\epsilon \delta n}{8}}$ .*

Note that Lemma 2 yields a lower bound which is exponential in the fraction of vertices in  $\{0, 1\}^m$  covered by a matching in every recovery graph of the code. Thus, if we can prove a tighter lower bound on the size of these matchings, then we get a corresponding improvement in the exponent in the final lower bound. This is exactly the method used in this paper. In particular, we achieve an optimized bound on the size of the matchings in the recovery graphs by bypassing the reduction to smooth codes and instead arguing directly about locally decodable codes. The resulting direct reduction is strong enough to yield a tight final lower bound.

## 2 Blocking Distributions and Matchings

In this section, we prove a combinatorial theorem regarding the relationship between the probability that an edge of a graph sampled from any distribution intersects any vertex-subset of a given size, and the size of a maximum matching in the graph. This is the primary technical tool which allows us to optimize the lower bounds of [GKST02]. Further, the analysis seems independently interesting and may be applicable in other contexts.

Let  $G(V, E)$  be an undirected graph on  $n$  vertices,  $w : E \rightarrow \mathbf{R}^+$  a probability distribution on the edges of  $G$ ,  $\mathcal{W}$  the set of all such distributions, and  $S$  a subset of  $V$ . Our concern in this section is to establish a bound on the following parameter of  $G$  based on the size of a maximum matching in  $G$ :

**Definition 4 (Blocking Probability).** *Let  $X^w$  denote a random edge of  $G$  sampled according to distribution  $w$ . Define the blocking probability  $\beta_\delta(G)$  as*

$$\beta_\delta(G) = \min_{w \in \mathcal{W}} \left( \max_{S \subseteq V, |S| \leq \delta n} \Pr[X^w \cap S \neq \emptyset] \right).$$

One can think of  $\beta_\delta(G)$  as the value of a game in which the goal of the first player (the decoding algorithm) is to sample an edge from  $G$  which avoids a vertex in a  $\delta n$ -set selected by the second player (the channel adversary), whose goal is to maximize the probability of blocking the edge selected by the first player.

For a graph  $G$ , the *independence number*  $\alpha(G)$  of  $G$  is the size of a maximum independent set of vertices in  $G$ , and the *defect*  $d(G)$  of  $G$  is the number of vertices left uncovered in a maximum matching in  $G$ . We begin our analysis by observing that  $d(G)$  is a lower bound on  $\alpha(G)$ . We then define a relaxation

of the optimization problem for the blocking probability on graphs with a given independence number. For this relaxed problem, we define a special family of distributions and show that some distribution in this family optimizes the blocking probability. Finally, we exhibit a lower bound on the blocking probability of a particular set of  $\delta n$  vertices with respect to any distribution in this family of distributions.

**Lemma 3.** *For a graph  $G$ ,  $\alpha(G) \geq d(G)$ .*

*Proof.* Choose any maximum matching in  $G$ . The vertices left uncovered in this matching must be an independent set, for an edge between any of these vertices would allow us to increase the size of the matching by at least one.

Define the graph  $K(n, \alpha)$  with vertex set

$$K_1(n, \alpha) \cup K_2(n, \alpha), |K_1(n, \alpha)| = \alpha n, |K_2(n, \alpha)| = (1 - \alpha)n,$$

such that the edge set of  $K(n, \alpha)$  is the union of the edge set of the complete bipartite graph with bipartition  $(K_1(n, \alpha), K_2(n, \alpha))$  and the  $(1 - \alpha)n$ -clique on  $K_2(n, \alpha)$ .

**Lemma 4.** *Let  $G$  be a graph with defect  $\alpha n$ . Then*

$$\beta_\delta(G) \geq \beta_\delta(K(n, \alpha)).$$

*Proof.* By Lemma 3,  $G$  has an independent set  $S$  of size at least  $\alpha n$ . With a labeling of vertices of  $K(n, \alpha)$  which sets  $K_1(n, \alpha)$  to an arbitrary  $\alpha n$ -subset of  $S$ , it is easy to see that the edge set of  $K(n, \alpha)$  contains the edge set of  $G$ . Therefore, the optimization of  $w$  on  $K(n, \alpha)$  is a relaxation of the optimization of  $w$  on  $G$  (a distribution  $w$  on  $G$  can be expressed as a distribution  $w'$  on  $K(n, \alpha)$  in which any edge of  $K(n, \alpha)$  not in  $G$  has probability 0). The claim follows.

We will focus on the following special class of distributions on  $K(n, \alpha)$  and show that the blocking probability of  $K(n, \alpha)$  is always optimized by some distribution in this class:

**Definition 5 (( $\lambda_1, \lambda_2$ )-Symmetric Distribution).** *An edge distribution  $w$  on the graph  $K(n, \alpha)$  is ( $\lambda_1, \lambda_2$ )-symmetric if for every edge  $e \in (K_1(n, \alpha), K_2(n, \alpha))$ ,  $w(e) = \lambda_1$ , and for every edge  $e \in (K_2(n, \alpha), K_2(n, \alpha))$ ,  $w(e) = \lambda_2$ .*

**Lemma 5.** *Let  $w_1, \dots, w_k$  be edge distributions on  $G$  such that*

$$\max_{S \subseteq V, |S| \leq \delta n} \Pr[X^{w_i} \cap S \neq \emptyset] = \beta_\delta(G).$$

*Then for any convex combination of the distributions  $w = \sum_i \gamma_i w_i$ ,*

$$\max_{S \subseteq V, |S| \leq \delta n} \Pr[X^w \cap S \neq \emptyset] = \beta_\delta(G).$$

*Proof.* For every  $S \subseteq V$ ,

$$\Pr[X^w \cap S \neq \emptyset] = \sum_i \gamma_i \Pr[X^{w_i} \cap S \neq \emptyset]$$

since this is simply the sum over edge weights of edges of  $G$  incident to  $S$ . By the condition on the  $w_i$ , for any subset  $S$  with  $|S| \leq \delta n$ ,

$$\begin{aligned} \Pr[X^w \cap S \neq \emptyset] &\leq \sum_i \gamma_i \beta_\delta(G) \\ &= \beta_\delta(G) \sum_i \gamma_i \\ &= \beta_\delta(G). \end{aligned}$$

Therefore,

$$\max_{S \subseteq V, |S| \leq \delta n} \Pr[X^w \cap S \neq \emptyset] \leq \beta_\delta(G).$$

However, by definition of  $\beta_\delta(G)$ , this must be at least  $\beta_\delta(G)$ . Therefore,

$$\max_{S \subseteq V, |S| \leq \delta n} \Pr[X^w \cap S \neq \emptyset] = \beta_\delta(G).$$

The *automorphism group* of a graph  $G$  is the set of permutations  $\pi$  on the vertices of  $G$  such that  $(\pi(i), \pi(j)) \in E \iff (i, j) \in E$ . Let  $\Gamma$  be the automorphism group of  $K(n, \alpha)$ .

**Lemma 6.** *There exists a  $(\lambda_1, \lambda_2)$ -symmetric distribution  $w$  such that*

$$\max_{S \subseteq V, |S| \leq \delta n} \Pr[X^w \cap S \neq \emptyset] = \beta_\delta(K(n, \alpha)).$$

*Proof.* Let  $w'$  be any distribution which optimizes the blocking probability of  $K(n, \alpha)$ . It is obvious that if  $w'$  is such a distribution, then so is  $\pi(w')$  for  $\pi \in \Gamma$  (where we extend the action of  $\Gamma$  to the edges of  $G$  in the natural way). By Lemma 5, the distribution

$$w = \frac{1}{|\Gamma|} \sum_{\pi \in \Gamma} \pi(w')$$

optimizes the blocking probability of  $K(n, \alpha)$ . We claim that  $w$  is a  $(\lambda_1, \lambda_2)$ -symmetric distribution: For any edge  $e \in E$  and  $\sigma \in \Gamma$ ,

$$\begin{aligned} w(e) &= \frac{1}{|\Gamma|} \sum_{\pi \in \Gamma} w'(\pi(e)) \\ &= \frac{1}{|\Gamma|} \sum_{\pi \in \Gamma} w'(\pi\sigma(e)) \\ &= w(\sigma(e)) \end{aligned}$$

where the second step is the usual group-theoretic trick of permuting terms in summations over  $\Gamma$ . Therefore, if  $e, e' \in E$  are in the same orbit under the action of  $\Gamma$ ,  $w(e) = w(e')$ . It is easy to verify that  $\Gamma$  is the direct product of the group of permutations of the vertices of  $K_1(n, \alpha)$  and  $K_2(n, \alpha)$ , and so there are exactly two edge-orbits of  $K(n, \alpha)$  under  $\Gamma$ , one consisting of the edges  $(K_1(n, \alpha), K_2(n, \alpha))$  and the other  $(K_2(n, \alpha), K_2(n, \alpha))$ . This is exactly the condition for a  $(\lambda_1, \lambda_2)$ -symmetric distribution.

Finally, we need to compute a lower bound on the blocking probability for a  $(\lambda_1, \lambda_2)$ -symmetric distribution:

**Lemma 7.** *Let  $w$  be a  $(\lambda_1, \lambda_2)$ -symmetric distribution on  $K(n, \alpha)$ . Then there exists a subset  $S \subseteq V$  with  $|S| \leq \delta n$  such that*

$$\Pr[X^w \cap S \neq \emptyset] \geq \min\left(\frac{\delta}{1-\alpha}, 1\right).$$

*Proof.* We will study a blocking set which selects any  $\delta n$  vertices of  $K_2(n, \alpha)$ . Note that, by  $(\lambda_1, \lambda_2)$ -symmetry, it does not matter which  $\delta n$  vertices we select. Further, we can assume that  $\delta < 1 - \alpha$ , for if  $\delta \geq 1 - \alpha$  we can cover all of  $K_2(n, \alpha)$  and thereby achieve blocking probability 1.

Placing a blocking set in this manner and summing up over edges and weights, we achieve blocking probability

$$(\delta n)(\alpha n)\lambda_1 + \frac{1}{2}(\delta n)(\delta n - 1)\lambda_2 + (\delta n)(1 - \alpha - \delta)n\lambda_2.$$

Since  $w$  is a probability distribution, we must have

$$(\alpha n)(1 - \alpha)n\lambda_1 + \frac{1}{2}(1 - \alpha)n((1 - \alpha)n - 1)\lambda_2 = 1.$$

Using this to eliminate  $\lambda_1$  from the first expression, we obtain blocking probability

$$\delta \left( \frac{1}{1-\alpha} + \frac{1}{2}n^2(1-\alpha-\delta)\lambda_2 \right).$$

Since  $\delta < 1 - \alpha$ , the second term in the sum is positive (and, obviously, optimized when  $\lambda_2 = 0$ ), so the blocking probability must be at least

$$\frac{\delta}{1-\alpha}.$$

It is now easy to prove our main result:

**Theorem 5.** *Let  $G$  be a graph with defect  $\alpha n$ . Then*

$$\beta_\delta(G) \geq \min\left(\frac{\delta}{1-\alpha}, 1\right).$$

*Proof.* By Lemma 4,  $\beta_\delta(G) \geq \beta_\delta(K(n, \alpha))$ . By Lemma 6, the blocking probability of  $K(n, \alpha)$  is optimized by some  $(\lambda_1, \lambda_2)$ -symmetric distribution. By Lemma 7, there exists a subset of  $\delta n$  vertices which blocks any such distribution with probability at least  $\min\left(\frac{\delta}{1-\alpha}, 1\right)$ . Therefore,  $\beta_\delta(G) \geq \beta_\delta(K(n, \alpha)) \geq \min\left(\frac{\delta}{1-\alpha}, 1\right)$ .

## 2.1 Probabilistic Proof of Theorem 5

An anonymous referee observed the following probabilistic proof of Theorem 5. This argument does not characterize the optimal strategies for the blocking game, as in our original analysis, but is sufficient to prove our ultimate result.

Fix an arbitrary edge-distribution  $w$  on  $K(n, \alpha)$  and, for  $\delta < 1 - \alpha$  as before, select a subset  $S$  of  $\delta n$  vertices of  $K_2(n, \alpha)$  uniformly at random. The resulting blocking probability  $\beta$  can be written as a sum  $\beta = \sum_e \beta_e$  over edges  $e$ , where  $\beta_e$  is a random variable with value  $w_e$  if  $S$  intersects  $e$ , or 0 otherwise. By linearity of expectation,

$$E(\beta) = \sum_e E(\beta_e) = \sum_e w_e \Pr[S \cap e \neq \emptyset]$$

where the randomness is over the selection of the subset  $S$ . Clearly,  $S$  intersects each edge  $e$  with probability at least  $\frac{\delta n}{(1-\alpha)n} = \frac{\delta}{1-\alpha}$ , so this expectation is at least

$$\sum_e w_e \frac{\delta}{1-\alpha} = \frac{\delta}{1-\alpha} \sum_e w_e = \frac{\delta}{1-\alpha}.$$

In particular, there must exist some subset  $S$  achieving this expectation, proving the theorem. In fact, our original analysis shows that, for the natural family of symmetric optimal strategies  $w$ , every  $\delta n$ -subset  $S$  of  $K_2(n, \alpha)$  achieves this blocking probability.

## 3 Lower Bounds

In this section, we apply Theorem 5 to our original problem of finding lower bounds for locally decodable linear codes. We also present a construction of a family of 2-query decodable linear codes with size matching our bounds within a constant factor in the exponent.

### 3.1 Degenerate to Non-Degenerate Reduction

We require a degenerate to non-degenerate reduction analogous to Theorem 2. Note that we cannot use Theorem 2 directly as this argues about smooth codes, whereas the point in our analysis is to bypass the use of smooth codes.

**Theorem 6.** Let  $\mathbf{C} : \{0, 1\}^n \rightarrow \{0, 1\}^m$  be a  $(2, \delta, \epsilon)$ -locally decodable linear code where  $n$  is large enough so that  $\frac{2(n+1)}{2^n} \leq \delta/201$ . Then there exists a non-degenerate  $(2, \frac{\delta}{2.01}, \epsilon)$ -locally decodable code  $\mathbf{C}' : \{0, 1\}^n \rightarrow \{0, 1\}^{2m}$ .

*Proof.* Write the  $i$ th entry  $y_i$  of the codeword as  $y_i = a_i \cdot x$  where  $a_i \in \{0, 1\}^n$ . A straightforward probabilistic argument shows that there exists a vector  $r \in \{0, 1\}^n$  such that the Hamming weights of  $r$  and  $a_i + r$  are at least 2 for a fraction at least  $\left(1 - \frac{2(n+1)}{2^n}\right)$  of the  $y_i$ . Let  $S$  be the set of  $y_i$  satisfying this property. Note that for  $y_i \in S$ ,  $(a_i + r) \cdot x$  is not a scalar multiple of an input entry. We form a non-degenerate code  $\mathbf{C}' : \{0, 1\}^n \rightarrow \{0, 1\}^{2m}$  from  $\mathbf{C}$  by setting  $y'_i = (a_i + r) \cdot x$  for  $y_i \in S$ ,  $y'_i = (1, \dots, 1) \cdot x$  for all other indices, and adding a set of  $m$  codeword bits  $y''_i = r \cdot x$  for all  $i \in \{1, \dots, m\}$ .

We claim that  $\mathbf{C}'$  is a  $(2, \frac{\delta}{2.01}, \epsilon)$ -locally decodable code. Let  $A$  be a recovery algorithm for  $\mathbf{C}$ , and recall that an edge for  $A$  can be good only if the answer of  $A$  is a linear combination of the entries it queries. Without loss of generality, we can assume that  $A$  only queries good edges (otherwise, we can ignore the answers to the queries and output a random coin flip). We implement a recovery algorithm  $A'$  for  $\mathbf{C}'$  as follows: If  $A$  takes a non-trivial linear combination of queries  $y_i, y_j$ , then  $A'$  simulates  $A$  but executes queries  $y'_i, y'_j$ ; if  $A$  is the identity on a query  $y_i$ , then  $A'$  makes queries  $y'_i, y''_i$  and takes the (non-trivial) linear combination  $y'_i + y''_i$ , which for  $i \in S$  equals  $(a_i + r) \cdot x + r \cdot x = a_i \cdot x = y_i$ . Finally, we note that if at most  $\frac{\delta}{2.01}$  entries of a codeword of  $\mathbf{C}'$  are corrupted, then  $A'$  exactly simulates the behavior of  $A$  when interacting with some code word with at most

$$\frac{\delta}{2.01}2m + |\bar{S}| \leq \frac{200}{201}\delta m + \frac{2(n+1)}{2^n}m \leq \frac{200}{201}\delta m + \frac{1}{201}\delta m = \delta m$$

corrupt entries. By the decoding condition on  $A$ ,  $A'$  succeeds with probability at least  $\frac{1}{2} + \epsilon$ .

So, we can essentially assume that we have a non-degenerate locally decodable code.

### 3.2 Lower Bound for $(2, \delta, \epsilon)$ -Locally Decodable Linear Codes on $\mathbf{Z}_2$

**Theorem 7.** Let  $\mathbf{C} : \{0, 1\}^n \rightarrow \{0, 1\}^m$  be a  $(q, \delta, \epsilon)$ -locally decodable linear code for  $0 < \delta, \epsilon < \frac{1}{2}$ . Then for sufficiently large  $n$ ,  $m \geq 2^{\frac{1}{4.03} \frac{\delta}{1-2\epsilon} n}$ .

*Proof.* By Theorem 6, for sufficiently large  $n$ , the existence of  $\mathbf{C}$  implies the existence of a non-degenerate  $(2, \frac{\delta}{2.01}, \epsilon)$ -locally decodable code  $\mathbf{C}' : \{0, 1\}^n \rightarrow \{0, 1\}^{2m}$ . As before, we can assume that the recovery algorithm  $A'$  for  $\mathbf{C}'$  only queries good edges. On one hand, for all  $i \in \{1, \dots, n\}$  and  $y \in \{0, 1\}^{2m}$  such that  $d(y, \mathbf{C}'(x)) \leq \frac{\delta}{2.01}(2m)$ ,

$$\Pr [A'^y(i) \neq x_i] \leq \frac{1}{2} - \epsilon.$$

On the other, if  $\alpha(2m)$  is the minimum over all  $i \in \{1, \dots, n\}$  of the defect in the recovery graph  $G_i$  of  $A'$ , then

$$\Pr[A'^y(i) \neq x_i] \geq \frac{1}{2} \frac{\delta/2.01}{1-\alpha}$$

for, by Theorem 5, there exists a fraction  $\frac{\delta}{2.01}$  of vertices  $S$  such that an adversary which sets the values of  $S$  to random coin flips causes  $A'$  to read a blocked edge, and therefore have probability  $\frac{1}{2}$  of outputting an incorrect response, with probability at least  $\frac{\delta/2.01}{1-\alpha}$ . Therefore,

$$\frac{1}{2} \frac{\delta/2.01}{1-\alpha} \leq \frac{1}{2} - \epsilon \implies \alpha \leq 1 - \frac{\delta/2.01}{1-2\epsilon}$$

which is equivalent to saying that there exists for all  $i \in \{1, \dots, n\}$  a set of at least  $\frac{1}{2} \frac{\delta/2.01}{1-2\epsilon} (2m)$  disjoint pairs of indices  $j_1, j_2$  such that  $x_i = q_{j_1} + q_{j_2}$ . Then by Lemma 2,  $2m \geq 2^{\frac{1}{2} \frac{\delta/2.01}{1-2\epsilon} n}$  or  $m \geq 2^{\frac{1}{4.02} \frac{\delta}{1-2\epsilon} n-1}$  which is at least, say,  $2^{\frac{1}{4.03} \frac{\delta}{1-2\epsilon} n}$  for sufficiently large  $n$ .

### 3.3 Matching Upper Bound

Finally, we show that the lower bound of Theorem 7 is optimal within a constant factor in the exponent. The following construction was observed earlier by Luca Trevisan.

The *Hadamard code* on  $x \in \{0, 1\}^n$  is given by

$$y_i = a_i \cdot x$$

where  $a_i$  runs through all  $2^n$  vectors in  $\{0, 1\}^n$ . Hadamard codes are locally decodable with 2 queries as, for any  $i \in \{1, \dots, n\}$  and  $r \in \{0, 1\}^n$ ,

$$x_i = r \cdot x + (r + e_i) \cdot x = e_i \cdot x$$

where  $e_i$  is the  $i$ th unit vector in  $\{0, 1\}^n$ . It is easy to see that the recovery graphs of this code are perfect matchings on the  $n$ -dimensional hypercube, and the code has recovery parameter  $\epsilon = \frac{1}{2} - 2\delta$ .

For given  $\delta, \epsilon$ , let  $c = \frac{1-2\epsilon}{4\delta}$ . It can be shown that for feasible values of  $\delta, \epsilon$ ,  $1 - 2\epsilon \geq 4\delta$  so that  $c \geq 1$ . We divide the input bits into  $c$  blocks of  $\frac{n}{c}$  bits, and encode each block with the Hadamard code on  $\{0, 1\}^{\frac{n}{c}}$ . The resulting code has length  $\frac{1-2\epsilon}{4\delta} 2^{\frac{4\delta}{1-2\epsilon} n}$  which is, say, less than  $2^{4.01 \frac{\delta}{1-2\epsilon} n}$  for sufficiently large  $n$ . Finally, since each code block has at most a fraction  $c\delta$  of corrupt entries, the code achieves recovery parameter

$$\frac{1}{2} - 2c\delta = \frac{1}{2} - 2 \left( \frac{1-2\epsilon}{4\delta} \right) \delta = \epsilon$$

as required.

## 4 Acknowledgements

The author thanks Luca Trevisan for many helpful discussions and the anonymous referees for several corrections and observations, including the probabilistic argument in Section 2.1.

## References

- [CGKS98] B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan. Private information retrieval. *Journal of the ACM*, 45(6):965–982, 1998.
- [GKST02] O. Goldreich, H. Karloff, L. Schulman, and Luca Trevisan. Lower bounds for linear locally decodable codes and private information retrieval. In *Proc. of the 17th IEEE CCC*, 2002.
- [KT00] J. Katz and Luca Trevisan. On the efficiency of local decoding procedures for error-correcting codes. In *Proc. of the 32nd ACM STOC*, 2000.