Electronic Colloquium on Computational Complexity, Report No. 6 (2007)

New Lower Bounds for General Locally Decodable Codes

David P. Woodruff
MIT
dpwood@mit.edu

Abstract

For any odd integer $q > 1$, we improve the lower bound for general q-query locally decodable codes $C : \{0,1\}^n \to \{0,1\}^m$ from $m = \Omega\left(n/\log n\right)^{\frac{q+1}{q-1}}$ to $m = \Omega\left(n^{\frac{q+1}{q-1}}\right)/\log n$. For example, for $q = 3$ we improve the previous bound from $\Omega(n^2/\log^2 n)$ to $\Omega(n^2/\log n)$. For *linear* 3-query locally decodable codes $C : \mathbb{F}^n \to \mathbb{F}^m$, we improve the lower bound further to $\Omega(n^2/\log\log n)$, and our bound holds for any (possibly infinite) field $\mathbb{F}$. Previously, the best lower bound for this case was $\Omega(n^2/\log^2 n)$, and held only for constant-sized $\mathbb{F}$. We are not aware of any previous non-trivial lower bounds for large $\mathbb{F}$ and $q > 2$ queries.

Our proofs use a random restriction of the message, hypergraph arguments, a new reduction from a q-query code to a generalization of a 2-query code, and quantum arguments. For linear codes our proofs are completely elementary. We work with random linear projections and use additional structure in the hypercube. The idea of using a random restriction (or projection for linear codes) is new in this context, and may be a powerful technique for future work.

ISSN 1433-8092

1 Introduction

Classical error-correcting codes allow one to encode an n-bit message x into a codeword $C(x)$ such that even if a constant fraction of the bits in $C(x)$ are corrupted, x can still be recovered. It is well-known how to construct codes C of length $O(n)$ that can tolerate a constant fraction of errors, even in such a way that allows decoding in linear time [14]. However, if one is only interested in recovering a few bits of the message, then these codes have the disadvantage that they require reading all (or most) of the codeword. This motivates the following definition.

A *locally decodable code* $C : \{0,1\}^n \to \{0,1\}^m$ is an encoding from n-bit strings to m-bit strings such that each bit x_i can be recovered with probability at least $\frac{1}{2} + \epsilon$ from $C(x)$ by a randomized algorithm that reads only q positions of $C(x)$, even if up to δm positions in $C(x)$ are corrupted. In typical applications, ϵ, δ, and q are constant, and the goal is to understand the tradeoff between q and m.

There is a large body of work on locally decodable codes [1, 2, 5, 7, 8, 11, 15, 16, 17]. For a survey, see [15]. Katz and Trevisan [7] were the first to formally define locally decodable codes. For 2 queries, Kerenidis and de Wolf [8] use tools from quantum information theory to show that $m = 2^{\Omega(n)}$, and the Hadamard code easily shows this is tight (see also [5, 11]). For $q > 2$ queries the best lower bound [8] is $m = \Omega\left(n/\log n\right)^{1+1/(\lceil \frac{q}{2} - 1 \rceil)}$, also due to Kerenidis and de Wolf. This is much smaller than that for 2 queries; however, there is also a much better upper bound[1] of $m \leq 2^{O(n^{c(q)})}$, for a small positive constant $c(q)$. This is obtained by combining a generic recursion of Beimel *et al* [1] with a recent result of Yekhanin [18] for 3 queries. With this state of affairs, it is hard to guess the optimal length of locally decodable codes.

It is quite difficult to prove lower bounds for general codes, and this has motivated researchers to study lower bounds in restricted models [5, 6, 12]. One natural subclass of these codes is the class of *linear locally decodable codes*, defined as follows. Let $\mathbb{F}$ be a field. A linear locally decodable code C is a linear transformation from $\mathbb{F}^n$ to $\mathbb{F}^m$ such that each coordinate of each $x \in \mathbb{F}^n$ can be recovered with probability at least $\frac{1}{|\mathbb{F}|} + \epsilon$ from $C(x)$ by a randomized algorithm that reads only q positions of $C(x)$, even if up to δm positions of $C(x)$ are corrupted. Here $\frac{1}{|\mathbb{F}|} = 0$ if $\mathbb{F}$ is infinite. All known constructions of locally decodable codes are linear, and all known lower bounds for linear codes before this work match, up to the dependence on ϵ and δ, those for general (not necessarily linear) codes.

Our Results: For any odd number of queries q, we improve the lower bound for *general* codes from $m = \Omega\left(n/\log n\right)^{1+1/(\lceil \frac{q}{2} - 1 \rceil)}$ to $m = \Omega\left(n\right)^{1+1/(\lceil \frac{q}{2} - 1 \rceil)} / \log n$. Next, for 3-query *linear* codes we improve our bound further from the previous $m = \Omega(n^2/\log^2 n)$ [8] to $m = \Omega(n^2/\log\log n)$, and our bound holds for any field $\mathbb{F}$, whereas the previous bound only holds for constant-sized $\mathbb{F}$.

Techniques: Given a locally decodable code, we use the reduction of [7, 8] to create a *smooth code*, that is, a code where for each $i \in [n]$, the decoder more or less uniformly distributes its queries over $C(x)$. Our smooth code is only good on average, that is, for each $i \in [n]$ and for most $x \in \{0,1\}^n$ the decoder correctly outputs x_i. Next, we find a small set T of $\Theta(n)$ heavily probed positions in the codeword that contain a lot of information about x. We restrict the set of possible x by fixing the assignment of the codeword to the positions in T. We show how to do this so as to still preserve a lot of entropy in x, while at the same time preserving the correctness of the decoder (on average). This sometimes reduces the number of queries of the decoder, since we can hardwire the values of positions in T into the decoder. We present a novel reduction from a q-query code to a generalization of a 2-query code, exploiting the fact that the decoder sometimes makes less than q queries. Finally, we generalize existing lower bounds for 2-query codes.

[1] Under a number-theoretic conjecture, this can be improved to $2^{O(n^{1/\log^{1-\alpha}\log n})}$ for any $\alpha > 0$ [18].

In [15], Trevisan asked whether one could reprove the results of [8] without using quantum information theory. Recently, Samorodnitsky [13] has shown how to do this for 2 queries. We note, though, that his proof is heavily inspired from the earlier work of [8]. The key idea in [13] is a new notion of entropy. This avoids the usage of a few very deep theorems in quantum information theory. With a significant effort, we can adapt his technique to our setting, and thus prove all of our results without quantum arguments. The only place we use his technique is to lower bound our generalized 2-query codes. However, since [13] is not published yet, in this version of the paper we give a proof of this step using quantum arguments. We believe this to be simpler, if one is willing to accept a few deep theorems in quantum information theory, and has the added benefit of showing how to extend the techniques of [8] to more general settings.

For 3-query linear codes, we use random projections and isoperimetric inequalities in the hypercube. The proof is a rather complicated (though elementary) packing argument, but the main idea is similar to that for general codes - we try to reduce a 3-query code to a 2 query-code. The idea is to repeatedly project coordinates of the codeword to 0, while at the same time preserving correctness. Another difficulty is that we need to handle adaptive decoders, which is non-trivial when $|\mathbb{F}|$ is super-constant.

Outline: In Section 2, we provide background. In Section 3, we provide our lower bound for general codes. In Section 4 we prove our lower bound for linear 3-query codes. For readability, we defer some proofs to Appendices 5, 6, 7, 8, and 9.

Notation: For positive integers z, $[z] \stackrel{\text{def}}{=} \{1, 2, \ldots, z\}$. We omit ceilings and floors if not essential.

2 Background

Definition 1 *([7]) Let $\delta, \epsilon \in (0,1)$, q an integer. We say $C : \{0,1\}^n \to \{0,1\}^m$ is a (q, δ, ϵ)-* locally decodable code *(LDC for short) if there is a probabilistic oracle machine A such that:*

- *In every invocation, A makes at most q queries.*
- *For every $x \in \{0,1\}^n$, for every $y \in \{0,1\}^m$ with $\Delta(y, C(x)) \le \delta m$, and for every $i \in [n]$,*

$$\Pr[A^y(i) = x_i] \;\geq\; \frac{1}{2} + \epsilon,$$

  where the probability is taken over the internal coin tosses of A. An algorithm A satisfying the above is called a (q, δ, ϵ)-local decoding algorithm for C (a decoder for short).

Since the code is binary[2], by the results of Katz and Trevisan [7] we may assume that for constant q, A queries non-adaptively. This only decreases ϵ by a constant factor. In all of the reductions between various codes that we discuss, non-adaptivity of the decoder is preserved.

Intuitively, a local-decoding algorithm A cannot query any particular location of y too often, as otherwise an adversary could ruin the success probability of A by corrupting only a few positions. This motivates the definition of a *smooth code*.

Definition 2 *([7]) For fixed c, ϵ, and integer q we say that $C : \{0,1\}^n \to \{0,1\}^m$ is a (q, c, ϵ)-* smooth code *if there exists a probabilistic oracle machine A such that for every $x \in \{0,1\}^n$,*

- *In every invocation A makes at most q queries.*
- *For every $i \in [n]$ and $j \in [m]$, $\Pr[A^{C(x)}(i)$ reads index $j] \le \frac{c}{m}$.*
- *For every $i \in [n]$, $\Pr[A^{C(x)}(i) = x_i] \;\geq\; \frac{1}{2} + \epsilon$.*

[2]For general codes, the binary setting is most often considered, and so we only consider non-binary codes when discussing linear codes, which we discuss later.

The probabilities are taken over the coin tosses of A. An algorithm A satisfying the above is called a (q, c, ϵ)-smooth decoding algorithm for C (a decoder for short).

Note that unlike a local-decoding algorithm, a smooth decoding algorithm is required to work only when given access to a valid codeword, rather than a possibly corrupt one. The following reduction from LDCs to smooth codes was observed by Katz and Trevisan.

Theorem 3 *([7]) Let $C : \{0,1\}^n \to \{0,1\}^m$ be a (q, δ, ϵ)-LDC. Then C is also a $(q, q/\delta, \epsilon)$-smooth code.*

We use the following weaker notion of a smooth code that is only good on average.

Definition 4 *A (q, c, ϵ)-*smooth code that is good on average *satisfies the first two conditions of a (q, c, ϵ)-smooth code, but the third condition is relaxed to the following: for every $i \in [n]$,*

$$\frac{1}{2^n} \sum_{x \in \{0,1\}^n} \Pr[A^{C(x)}(i) = x_i] \geq \frac{1}{2} + \epsilon.$$

We use a graph-theoretic interpretation of smooth codes given in [5, 7]. Although not stated explicitly there, their results also hold for smooth codes that are good on average. Let $C : \{0,1\}^n \to \{0,1\}^m$ be a (q, c, ϵ)-smooth code that is good on average, and let algorithm A be a (q, c, ϵ)-smooth decoding algorithm for C. We say that a given invocation of A *reads* a set $s \subseteq [m]$ if the set of indices that A reads in that invocation equals s. Since A is restricted to read at most q entries, $|s| \leq q$.

We say that s is *good* for i if $\Pr[A^{C(x)}(i) = x_i \mid A \text{ reads } s] \geq \frac{1}{2} + \frac{\epsilon}{2}$, where the probability is over x uniformly drawn from $\{0,1\}^n$ and the internal coin tosses of A.

Definition 5 *([7]) Fixing a code $C : \{0,1\}^n \to \{0,1\}^m$ and a q-query recovery algorithm A, the recovery hypergraphs for $i \in [n]$, denoted G_i, consist of the vertex set $[m]$ and the hyperedge set $C_i = \{s \subseteq [m] \mid s \text{ is good for } i\}$.*

Lemma 6 *([7]) Let C be a (q, c, ϵ)-smooth code that is good on average, and let $\{G_i\}_{i=1}^n$ be the set of recovery hypergraphs. Then, for every i, the hypergraph $G_i = ([m], C_i)$ has a matching M_i of sets of size q with $|M_i| \geq \frac{\epsilon m}{cq}$.*

For positive constants $c_1, c_2, \ldots$, $\Theta_{c_1, c_2, \ldots}(f(n))$ denotes the class $g(c_1, c_2, \ldots)\Theta(f(n))$ of functions, where g is an arbitrary positive function. Similarly define $O_{c_1,c_2,\ldots}(f(n))$ and $\Omega_{c_1,c_2,\ldots}(f(n))$.

Lemma 7 *([8], implicit) For constant q, if $C : \{0,1\}^n \to \{0,1\}^m$ is a (q, c, ϵ)-smooth code, then there is a $(q, O_{q,c,\epsilon}(1), \Omega_{q,c,\epsilon}(1))$-smooth code $C' : \{0,1\}^n \to \{0,1\}^{\Theta(m)}$ that is good on average, and further, for each $i \in [n]$, the decoder A' of C' just picks a random q-set $\{j_1, \ldots, j_q\} \in M_i$ and either outputs $C(x)_{j_1} \oplus C(x)_{j_2} \oplus \cdots \oplus C(x)_{j_q}$ or $C(x)_{j_1} \oplus C(x)_{j_2} \oplus \cdots \oplus C(x)_{j_q} \oplus 1$. The decision of which to output is based solely on $\{j_1, \ldots, j_q\}$ (this follows from non-adaptivity).*

The intuitive justification for Lemma 7 is as follows. Using Fourier analysis, one can show that if from q Boolean functions one can recover x_i with probability greater than $1/2 + \epsilon$, then from some sum of the functions one can recover x_i with probability greater than $1/2 + \epsilon/2^q$. Now, if the decoder often takes the sum of less than q functions, we can increase the length of the code by a constant fraction by adding many zero functions to the code, and now the decoder, by adding zero functions, can be assumed to always take the sum of q positions.

The following lemma will simplify notation. We defer its simple proof to Appendix 5.

Lemma 8 *If C is a (q, c, ϵ)-smooth code that is good on average for which for each $i \in [n]$, the decoder A picks a random q-set $\{j_1, \ldots, j_q\} \in M_i$ and either outputs $C(x)_{j_1} \oplus C(x)_{j_2} \oplus \cdots \oplus C(x)_{j_q}$ or $C(x)_{j_1} \oplus C(x)_{j_2} \oplus \cdots \oplus C(x)_{j_q} \oplus 1$, then there is a $(q, 2c, \epsilon)$-smooth code C' that is good on average for which for each $i \in [n]$ there is a bit $b_i \in \{0,1\}$ for which the decoder A' picks a random q-set $\{j_1, \ldots, j_q\} \in M_i$ and outputs $C(x)_{j_1} \oplus C(x)_{j_2} \oplus \cdots \oplus C(x)_{j_q} \oplus b_i$.*

In the remainder of the paper, we assume we have a code $C : \{0,1\}^n \to \{0,1\}^m$ that is a (q, c, ϵ)-smooth code that is good on average. Thus, for each i, there is a matching M_i of q-sets with $|M_i| \geq \beta(\epsilon, c, q)m$, for some function β of q, c, and ϵ, in the corresponding recovery hypergraphs G_i. We can also assume on input i that the decoder A just picks a random q-set $\{j_1, \ldots, j_q\} \in M_i$ and outputs $C(x)_{j_1} \oplus C(x)_{j_2} \oplus \cdots \oplus C(x)_{j_q} \oplus b_i$.

Via these reductions, the lower bounds on m we obtain for C give lower bounds for (q, δ, ϵ)-locally decodable codes with the same asymptotic dependence on n (for q, δ, and ϵ constant).

3 The lower bound for general codes

Suppose we are given a (q, c, ϵ)-smooth code $C : \{0,1\}^n \to \{0,1\}^m$ that is good on average (as defined in Section 2), and suppose that q is odd. We may identify the coordinates of the encoding with m functions $f_1, \ldots, f_m : \{0,1\}^n \to \{0,1\}$. By the results in the previous section, there is a positive constant $\beta \stackrel{\text{def}}{=} \beta(q, c, \epsilon)$ such that for all $i \in [n]$, there is a collection M_i of at least βm disjoint sets in $[m]$ of size q, and a bit b_i, such that for all $e \in M_i$,

$$\Pr_{x \in \{0,1\}^n}\left[b_i \oplus \bigoplus_{j \in e} f_j(x) = x_i\right] \geq \frac{1}{2} + \epsilon.$$

Our goal is to construct a related 2-query code which is easier to analyze.

3.1 A small set incident to many edges in the recovery hypergraphs

Consider the multi-hypergraph G with vertex set $[m]$ and hyperedge set $\uplus_{i=1}^n M_i$, that is, a hyperedge e occurs in G once for each M_i that it occurs in. For readability, we use the term hypergraph to refer to a multi-hypergraph, that is, a hypergraph which may have repeated hyperedges (which we sometimes just refer to as edges). We first claim that we can find a non-empty induced sub-hypergraph G' of G with minimum degree βn. Our proof is a straightforward generalization of Proposition 1.2.2 in [3] to hypergraphs, and thus, we defer it to Appendix 6.

Lemma 9 *There exists a non-empty hypergraph $G' \subseteq G$ with minimum degree at least βn.*

Let $v \in G'$ be an arbitrary vertex, and let $N(v)$ denote v's neighbors in G'. Consider the set $T = \{v\} \cup N(v)$. We would like to argue that T contains many vertices. To do this, we use the following generalization of Theorem 2 in [7].

Theorem 10 *Let $F : \{0,1\}^n \to R$ be a function. Assume there is an algorithm B such that for some set $J \subseteq [n]$ of indices, for any $j \in J$,*

$$\Pr[B(F(x), j) = x_j] \geq \frac{1}{2} + \epsilon,$$

where the probability is over both x uniformly drawn from $\{0,1\}^n$ and the coin tosses of B. Then $\log|R| \geq (1 - H(\frac{1}{2} + \epsilon))|J|$, where H is the binary entropy function.

Proof: Let $I(x; F(x)) = H(F(x)) - H(F(x) \mid x) = H(x) - H(x \mid F(x))$ denote the mutual information between x and $F(x)$. Then, $I(x; F(x)) \leq H(F(x)) \leq \log|R|$. On the other hand, using the chain rule and subadditivity of entropy, as well as Fano's inequality (p. 536 of [10]),

$$\begin{aligned} I(x; F(x)) &= H(x) - H(x \mid F(x)) \geq H(x) - \sum_{i=1}^n H(x_i \mid F(x)) \\ &\geq H(x) - (n - |J|) - \sum_{j \in J} H(x_j \mid F(x)) \geq |J| - |J| H(\frac{1}{2} + \epsilon), \end{aligned}$$

and combining the two inequalities establishes the lemma. ■

Claim 11 $|T| \geq \beta \cdot (1 - H(1/2+\epsilon))n$.

Proof: Observe that T contains an edge e in M_i for at least βn different i. This follows from the fact that v has degree at least βn, and for each $i \in [n]$, there is at most one edge $e \in M_i$ containing v since the M_i are matchings.

Let J denote the set of these i. It follows by the definition of an edge that the encoding of x by the functions in T has a decoding algorithm that recovers x_j, $j \in J$, with probability at least $\frac{1}{2}+\epsilon$. By the previous theorem, $|T| \geq |J|(1 - H(1/2+\epsilon)) \geq \beta \cdot (1 - H(1/2+\epsilon))n$. ■

Let $0 < \alpha \ll \beta \cdot (1 - H(1/2+\epsilon))$ be a constant to be determined, and remove all but αn vertices from T. For each $i \in [n]$, let $M_i' \subseteq M_i$ be the set of all edges in M_i incident to at least one vertex in T. Since each of the αn vertices in T has degree at least βn, and since each edge e in any M_i can be incident to at most q vertices of T, we have $\sum_{i=1}^n |M_i'| \geq \alpha \beta n^2/q = \alpha \Theta_{q,c,\epsilon}(n^2)$. Here the constant in the $\Theta_{q,c,\epsilon}(\cdot)$ may depend on q, c, and ϵ, but does not depend on α.

3.2 Randomly restricting $\Theta(n)$ coordinates

Let $T \subseteq [m]$ be the set of size exactly αn chosen in the previous section. Consider the multiset F_T of αn functions f_j, where $j \in T$. For each $x \in \{0,1\}^n$, the tuple $(f_j(x) \mid j \in T)$ is a string in $\{0,1\}^{\alpha n}$. Thus, we may partition $\{0,1\}^n$ into $2^{\alpha n}$ equivalence classes (some of which may be empty) L_b, where $b \in \{0,1\}^{\alpha n}$. Here L_b denotes all $x \in \{0,1\}^n$ for which $(f_j(x) \mid j \in T) = b$.

Say an equivalence class L_b is *bad* if $|L_b| \leq 2^{n-2\alpha n}$. If L_b is not bad, then it is *good.* Say an $x \in \{0,1\}^n$ is *bad* if $x \in L_b$ for a bad L_b. If x is not bad, then it is *good.* As there are $2^{\alpha n}$ different L_b, the total number of bad x is at most $2^{\alpha n} 2^{n-2\alpha n} = 2^{n-\alpha n}$. Let $X \subseteq \{0,1\}^n$ be the set of all good $x \in \{0,1\}^n$. Then $|X| \geq 2^n - 2^{n-\alpha n}$.

Consider any $i \in [n]$, and let $e \in M_i'$. By a union bound, we have,

$$\Pr_{x \in X}\left[b_i \oplus \bigoplus_{j \in e} f_j(x) = x_i\right] \geq \frac{1}{2} + \epsilon - \frac{2^{n-\alpha n}}{2^n} \geq \frac{1}{2} + \frac{\epsilon}{2},$$

for any $\alpha > 0$ and n sufficiently large. This holds for every i and every $e \in M_i'$. As our goal will be to fix the values of functions in F_T, we now try to find a good class with special properties.

Lemma 12 *There exists a good equivalence class L and an index set $I \subset [n]$ with $|I| = \Theta_{q,c,\epsilon}(n)$, for which for all $i \in I$, there are at least $\alpha \Theta_{q,c,\epsilon}(n)$ different $e \in M_i'$ for which*

$$\Pr_{x \in L}\left[b_i \oplus \bigoplus_{j \in e} f_j(x) = x_i\right] \geq \frac{1}{2} + \frac{\epsilon}{4}.$$

Proof: Consider the probability distribution P on good equivalence classes L_b defined by: $\Pr[P = L_b] = \frac{|L_b|}{|X|}$. For each $i \in [n]$ and each $e \in M_i'$, define the random variable

$$Y_{i,e} = \Pr_{x \in P}\left[b_i \oplus \bigoplus_{j \in e} f_j(x) \neq x_i\right].$$

Then

$$\mathbf{E}[Y_{i,e}] = \sum_{L_b} \frac{L_b}{|X|} \Pr_{x \in L_b}\left[b_i \oplus \bigoplus_{j \in e} f_j(x) \neq x_i\right] \leq \frac{1}{2} - \frac{\epsilon}{2}.$$

It follows by the Markov bound that

$$\Pr\left[Y_{i,e} \geq \frac{1}{2} - \frac{\epsilon}{4}\right] \leq \left(\frac{1}{2} - \frac{\epsilon}{2}\right) / \left(\frac{1}{2} - \frac{\epsilon}{4}\right) = \gamma < 1.$$

It follows that with probability at least $1-\gamma$, $Y_{i,e}$ is at most $\frac{1}{2} - \frac{\epsilon}{4}$. Define the indicator random variable $J_{i,e}$ which is 1 iff $Y_{i,e} \leq \frac{1}{2} - \frac{\epsilon}{4}$. Then $\mathbf{E}[J_{i,e}] \geq 1-\gamma$. Since $\sum_{i=1}^{n} |M_i'| \geq \alpha\Theta_{q,c,\epsilon}(n^2)$, by linearity of expectations,

$$\mathbf{E}\left[\sum_{i,e} J_{i,e}\right] \geq (1-\gamma)\alpha\Theta_{q,c,\epsilon}(n^2) = \alpha\Theta_{q,c,\epsilon}(n^2).$$

So there exists a good equivalence class L for which $\alpha\Theta_{q,c,\epsilon}(n^2)$ edges e in $\uplus_{i=1}^{n} M_i'$ satisfy $\Pr_{x\in L}\left[b_i \oplus \bigoplus_{j\in e} f_j(x) = x_i\right] \geq \frac{1}{2} + \frac{\epsilon}{4}$. Say such an e is *good*. For each i, $|M_i'| \leq \alpha n$. Moreover, the average number a of good e in M_i' is $\alpha\Theta_{q,c,\epsilon}(n)$. Let r denote the number of different $i \in [n]$ for which the number of good e in M_i' is at most $a/2$. Then r is subject to the following constraint: $\frac{a}{2}\cdot r + \alpha n(n-r) \geq an$. Solving,

$$n - r \geq \frac{an - \frac{ar}{2}}{\alpha n} = \frac{a}{\alpha} - \frac{ar}{2\alpha n} \geq \frac{a}{\alpha} - \frac{a}{2\alpha} = \frac{a}{2\alpha} = \Theta_{q,c,\epsilon}(n).$$

Thus, for a set $I \subseteq [n]$ of size $\Theta_{q,c,\epsilon}(n)$, for each $i \in I$ the number of good e in M_i' is at least $\alpha\Theta_{q,c,\epsilon}(n)$. ■

3.3 Reducing the number of queries to 2

Fix a set of indices $I \subseteq [n]$ guaranteed by Lemma 12. By relabeling indices if necessary, we may assume $I = [\Theta_{q,c,\epsilon}(n)]$. We construct a new code C'. Let η be a positive constant to be determined. Also, define the function $h(q) = 2/(q-1)$.

Consider all $m' = \binom{m}{\eta\frac{m}{n^{h(q)}}}$ functions $g_B : \{0,1\}^n \to \{0,1\}$ formed as follows: choose any set $B \subseteq [m]$ of size exactly $\eta\frac{m}{n^{h(q)}}$, and let $g_B = \bigoplus_{j\in B} f_j$. Let C' be the code which takes an $x \in L$, and applies each of these m' functions to x. The code has length m'.

Lemma 13 *There exists a constant $\eta > 0$ such that for each $i \in I$, $[m']$ contains a matching W_i of disjoint pairs (indexed by sets) $\{B, B'\}$ and a bit $b_{i,B,B'} \in \{0,1\}$, such that*

$$\Pr_{x\in L}[g_B(x) \oplus g_{B'}(x) \oplus b_{i,B,B'} = x_i] \geq \frac{1}{2} + \frac{\epsilon}{4}.$$

Moreover, $|W_i| \geq \frac{m'}{4}$ for all $i \in I$.

Proof: Fix an $i \in I$. M_i' has at least $\alpha\Theta_{q,c,\epsilon}(n)$ different e for which $\Pr_{x\in L}\left[b_i \oplus \bigoplus_{j\in e} f_j(x) = x_i\right] \geq \frac{1}{2} + \frac{\epsilon}{4}$. As before, call such an e a *good* edge. For each good edge $e \in M_i'$, e is incident to at least one vertex in T. Arbitrarily choose one such vertex, and denote it by v_e. Next, of the remaining $q-1$ vertices in e, arbitrarily partition them into two sets A_e and B_e. So $e = v_e \cup A_e \cup B_e$. We need the following lemma due to Katz and Trevisan [7]:

Lemma 14 *([7]) Let H be a hypergraph on m vertices whose hyperedges all contain c or fewer vertices. Let H have a matching M of size γm for any $0 < \gamma < 1/c$. Then there exists a $t = \Theta(\gamma^{-\frac{1}{c}} m^{\frac{c-1}{c}})$ such that for a collection B of t randomly selected vertices of H,*

$$\Pr_B[B \textit{ contains an edge of } M] > 3/4.$$

Consider the hypergraph H on vertex set $[m]$ whose hyperedges have size $(q-1)/2$ and are the sets A_e and B_e for each good edge $e \in M_i'$. Then the hyperedges form a matching M of H of size $2|M_i'| \geq \alpha\Theta_{q,c,\epsilon}(n)$. Choose a subset $B \subseteq [m]$ of size $\eta\frac{m}{n^{h(q)}}$ uniformly at random. By setting $\gamma = \alpha\Theta_{q,c,\epsilon}(n)/m$ and $c = (q-1)/2 = 1/h(q)$ in Lemma 14, for a sufficiently large constant η (that may depend on α), $\Pr[B$ contains an A_e or $B_e] > 3/4$.

For any good edge e' in M_i', the probability that e' satisfies $|e' \cap B| > (q-1)/2$ is at most

$$\sum_{j>(q-1)/2} \binom{q}{j} \frac{\binom{m-j}{\eta\frac{m}{n^{h(q)}}-j}}{\binom{m}{\eta\frac{m}{n^{h(q)}}}} \leq 2^q \left(\frac{\eta}{n^{h(q)}}\right)^{(q+1)/2} = \Theta_{q,c,\epsilon}\left(n^{-\frac{q+1}{q-1}}\right).$$

By a union bound, the probability there exists a good edge in M_i' contained in B is at most $|M_i'| \cdot \Theta_{q,c,\epsilon}\left(n^{-\frac{q+1}{q-1}}\right) = o(1)$. By another union bound, for at least half of the functions g_B,

1. B contains either A_e or B_e for at least one good edge $e \in M_i'$, and
2. For any good edge $e' \in M_i'$, $|e' \cap B| \leq (q-1)/2$.

Arbitrarily impose a total order on the good edges $e \in M_i'$. Fix any B satisfying the two properties above. Consider the multiset B' defined as follows: let e be the smallest good edge for which either $A_e \subseteq B$ or $B_e \subseteq B$. Note that for any given e, at most one of A_e, B_e can occur in B by the second property above. If $A_e \subseteq B$, define $B' = B_e \cup B \setminus A_e$, else define $B' = A_e \cup B \setminus B_e$. Note that in either case B' is a set (rather than a multiset), since $|e \cap B| = (q-1)/2$. Then

$$g_B(x) \oplus g_{B'}(x) = \bigoplus_{j \in A_e} f_j(x) \oplus \bigoplus_{k \in B_e} f_k(x) = f_{v_e}(x) \oplus \bigoplus_{j \in e} f_j(x).$$

Now for $x \in L$, $f_{v_e}(x)$ is constant. Define $b_{i,B,B'} = b_i \oplus f_{v_e}$. Then, since e is good,

$$\Pr_{x \in L}[g_B(x) \oplus g_{B'}(x) \oplus b_{i,B,B'} = x_i] \geq \frac{1}{2} + \frac{\epsilon}{4}.$$

Let ψ denote our map on sets satisfying the two properties above, so $\psi(B) = B'$. First, we claim B' satisfies the two properties above. Both properties follow from the fact that the good edges are disjoint, and thus $B' \cap e' = B \cap e'$ for all $e' \neq e$, while $|B' \cap e| = |B \cap e| = (q-1)/2$.

We claim that ψ is invertible on sets B which satisfy the two properties above. To see this, let e' be the smallest good edge for which either $A_{e'} \subseteq B'$ or $B_{e'} \subseteq B'$. It follows from the way we constructed B' that either $A_{e'} \subseteq B$ or $B_{e'} \subseteq B$. Then e' cannot be smaller than e in the total order since e is the smallest edge in B for which either $A_e \subseteq B$ or $B_e \subseteq B$. Since $A_e \subseteq B'$ or $B_e \subseteq B'$, we must then have $e' = e$. So, $\psi(B') = B$.

Thus, B uniquely determines B' and vice versa, and so we may group at least $1/2$ of the elements of $[m']$ into disjoint pairs, giving a matching W_i of size at least $m'/4$. ■

3.4 The quantum tools

At this point we have a set $I \subseteq [n]$ with $|I| = \Theta_{q,c,\epsilon}(n)$ and $m' = \binom{m}{\eta\frac{m}{n^{h(q)}}}$ functions g_B with the following property: for each $i \in I$, $[m']$ contains a matching W_i of disjoint pairs (indexed by sets) $\{B, B'\}$ and a bit $b_{i,B,B'} \in \{0,1\}$, such that

$$\Pr_{x \in L}[g_B(x) \oplus g_{B'}(x) \oplus b_{i,B,B'} = x_i] \geq \frac{1}{2} + \frac{\epsilon}{4}.$$

Moreover, $|W_i| \geq \frac{m'}{4}$ for all $i \in I$. Unfortunately, we cannot apply the 2-query lower bound of Kerenidis and de Wolf [8] directly since I may not equal $[n]$ and L may not equal $\{0,1\}^n$. We need to generalize Nayak's [9] lower bound for quantum random access codes to apply it to our setting. For readability, we defer this to Appendix 7. There we show that these constraints imply $m = \Omega_{q,c,\epsilon}(n^{1+h(q)}/\log n)$. Thus,

Theorem 15 *For odd q, any (q, c, ϵ)-smooth code $C : \{0,1\}^n \to \{0,1\}^m$ that is good on average satisfies*

$$m = \Omega_{q,c,\epsilon}\left(\frac{n^{1+2/(q-1)}}{\log n}\right).$$

Using the reductions in Section 2, we obtain,

Theorem 16 *For $\delta, \epsilon \in (0,1)$ and for any odd integer $q > 1$, if $C : \{0,1\}^n \to \{0,1\}^m$ is a (q, δ, ϵ)-locally decodable code, then*

$$m = \Omega_{q,\delta,\epsilon}\left(\frac{n^{1+2/(q-1)}}{\log n}\right).$$

So, for instance, if $q = 3$, the bound is $\Omega_{q,c,\epsilon}(n^2/\log n)$, improving the previous bound [8] of $\Omega_{q,c,\epsilon}(n^2/\log^2 n)$. If $q = 5$, the bound is $\Omega_{q,c,\epsilon}(n^{3/2}/\log n)$, improving the previous best bound [8] of $\Omega_{q,c,\epsilon}(n^{3/2}/\log^{3/2} n)$.

4 Linear 3-query lower bounds

4.1 The random projection

Assume we have a linear $(3, \delta, \epsilon)$-LDC $C : \mathbb{F}^n \to \mathbb{F}^m$ for an arbitrary (possibly infinite) field $\mathbb{F}$. Recall the model is that for every $x \in \mathbb{F}^n$, for every $y \in \mathbb{F}^m$ with[3] $\Delta(C(x), y) \le \delta m$, and for every $i \in [n]$, the decoder A satisfies $\Pr[A^y(i) = x_i] \ge \frac{1}{|\mathbb{F}|} + \epsilon$, where the probability is over the coin tosses of A. A queries at most 3 coordinates of y. In Appendix 8, we prove the following.

Theorem 17 *Let $C : \mathbb{F}^n \to \mathbb{F}^m$ be a linear $(3, \delta, \epsilon)$-LDC. Then C is also a linear $(3, \delta/9, 2/3 - 1/|\mathbb{F}|)$-LDC with a non-adaptive decoder.*

This greatly improves known reductions to non-adaptive codes (since it holds for any $\mathbb{F}$), but it only holds for linear codes. Thus, we may assume that we have a non-adaptive decoder by changing δ and ϵ by constant factors. Then, by similar reductions to those given in Section 2 for non-adaptive decoders (extended straightforwardly to arbitrary fields), C is also a $(3, 3/\delta, \epsilon)$-smooth code.

Since C is linear, we may identify its coordinates j with vectors f_j in $\mathbb{F}^n$ computing the function $\langle f_j, x\rangle$. By the reductions in Section 2, for every $i \in [n]$, the recovery hypergraph G_i has a matching M_i of $[m]$ of size $\Theta_{q,c,\epsilon}(m)$ such that, if $e \in M_i$, then $u_i \in \text{span}(f_j \mid j \in e)$, where u_i denotes the unit vector in direction i. This follows from the observation that if $u_i \notin \text{span}(f_j \mid j \in e)$, then e contains no information about x_i, and so any algorithm, when reading e, can output x_i with probability at most $1/2$.

We may assume, by increasing m by at most a factor of 3, that every hyperedge in M_i has size exactly 3, and moreover, for every such edge $e = \{j_1, j_2, j_3\} \in M_i$, we have $u_i = \gamma_1 f_{j_1} + \gamma_2 f_{j_2} + \gamma_3 f_{j_3}$, where $\gamma_1, \gamma_2, \gamma_3$ are non-zero elements of $\mathbb{F}$. Indeed, we may append $2m$ constant functions which always output 0 to the end of C. Then, if $e = \{j_1, j_2, j_3\} \in M_i$ either has size less than 3 or satisfies $u_i = \gamma_1 f_{j_1} + \gamma_2 f_{j_2} + \gamma_3 f_{j_3}$ for some $\gamma_k = 0$, we can replace the γ_k with 1 and replace j_k with an index corresponding to one of the zero functions.

Consider the non-empty hypergraph $G' \subseteq G$ with minimum degree βn given in Section 3.1. In Section 3.1, we found a set T which contained an edge e in M_i for at least βn different i. It follows that the rank of the vectors in T is at least βn, so we can remove vectors from T so that we are left with a set T of exactly αn linearly independent vectors.

[3]Here $\Delta(C(x), y)$ refers to the number of positions in $C(x)$ and y that differ.

Let $v_1, \ldots, v_T$ denote the vectors in T. Extend $\{v_1, \ldots, v_T\}$ to a basis of $\mathbb{F}^n$ by adding a set of $n - \alpha n$ *unit vectors* U. Define a linear projection L as follows:

$$\begin{aligned} L(v) &= 0 \text{ for all } v \in T \\ L(v) &= v \text{ for all } v \in U \end{aligned}$$

Since L is specified on a basis, it is specified on all of $\mathbb{F}^n$ by linearity.

Recall that M_i' denotes the collection of edges in M_i that are incident to some vertex in T. Let $e = \{j_1, j_2, j_3\}$ be an edge in some M_i'. Then there are non-zero $\gamma_1, \gamma_2, \gamma_3 \in \mathbb{F}$ for which $\gamma_1 f_{j_1} + \gamma_2 f_{j_2} + \gamma_3 f_{j_3} = u_i$. By linearity,

$$L(u_i) = L(\gamma_1 f_{j_1} + \gamma_2 f_{j_2} + \gamma_3 f_{j_3}) = \gamma_1 L(f_{j_1}) + \gamma_2 L(f_{j_2}) + \gamma_3 L(f_{j_3}).$$

By definition of M_i', $|\{j_1, j_2, j_3\} \cap T| > 0$, so one of the following must be true:

$$L(u_i) \in \text{span}(L(f_{j_1}), L(f_{j_2})), \ L(u_i) \in \text{span}(L(f_{j_1}), L(f_{j_3})), \text{ or } L(u_i) \in \text{span}(L(f_{j_2}), L(f_{j_3})).$$

Thus, for each such edge $e \in \{j_1, j_2, j_3\}$, by removing exactly one element $j_\ell \in \{j_1, j_2, j_3\}$ for which $L(f_{j_\ell}) = 0$, we may define matchings W_i of disjoint pairs $\{j, k\}$ of $[m]$ such that if $\{j, k\} \in W_i$, then $L(u_i) \in \text{span}(L(f_j), L(f_k))$. Moreover, $\sum_{i=1}^n |W_i| = \sum_{i=1}^n |M_i'| \geq \alpha \Theta_{c,\epsilon}(n^2)$.

Say an index $i \in [n]$ *survives* if $L(u_i) = u_i$, and say an edge e *survives* if $e \in M_i'$ for an i that survives. If i survives, then $u_i \in U$, as otherwise we would have $u_i = \sum_{v \in T} \gamma_v v + \sum_{u \in U} \gamma_u u$ for some coefficients $\gamma_v, \gamma_u \in \mathbb{F}$. Applying L to both sides we would obtain $u_i = \sum_{u \in U} \gamma_u u$, which is impossible unless $u_i \in U$.

Recall that each of the αn vertices v in T has degree at least βn in G'. For any such $v \in T$, there are at least $\beta n - \alpha n$ edges $e \in \uplus_i M_i'$ containing v that survive. Thus, since each edge that survives can be incident to at most q elements of T, and since $\alpha \ll \beta$,

$$\sum_{i \text{ that survive}} |W_i| \geq \alpha n(\beta - \alpha)n/q = \alpha \Omega_{c,\epsilon}(n^2).$$

For i that do not survive, we set $W_i = \emptyset$. We need a theorem due to Dvir and Shpilka [4].

Theorem 18 *([4]) Let $\mathbb{F}$ be any field, and let $a_1, \ldots, a_m \in \mathbb{F}^n$. For every $i \in [n]$, let M_i be a set of disjoint pairs of indices $\{j_1, j_2\}$ such that $u_i \in span(a_{j_1}, a_{j_2})$. Then,*

$$\sum_{i=1}^n |M_i| \leq m \log m + m.$$

Applying Theorem 18 to our setting, we have m vectors $L(f_j) \in \mathbb{F}^n$ and matchings W_i with $\sum_i |W_i| = \alpha \Omega_{c,\epsilon}(n^2)$. We conclude that,

Theorem 19 *For $\delta, \epsilon \in (0, 1)$, if $C : \mathbb{F}^n \to \mathbb{F}^m$ is a linear $(3, \delta, \epsilon)$-locally decodable code, then*

$$m = \Omega_{\delta,\epsilon}(n^2 / \log n),$$

which is independent of the field $\mathbb{F}$.

4.2 The additional optimizations

We improve the bound to $m = \Omega_{\delta,\epsilon}(n^2 / \log \log n)$. Let $N(T)$ denote all vertices $u \in G'$ that are incident to a vertex $v \in T$, that is, $N(T)$ denotes the neighbors of the set T in the hypergraph G'. Let U' be a random subset of U size exactly αn. We define a linear projection L' as follows:

$$\begin{aligned} L'(v) &= 0 \text{ for all } v \in T \cup U' \\ L'(v) &= v \text{ for all } v \in U \setminus U' \end{aligned}$$

Let e be a 3-edge that survives. In Section 4.1, we showed that if we apply L to each vertex (identified with a vector) in G', there are at least $\alpha\Omega_{c,\epsilon}(n^2)$ 3-edges e that survive. We say that e is *zeroed out* if $e \in M'_i$ and $u_i \in U'$.

Claim 20 *There exists an L' for which $\alpha^2\Omega_{c,\epsilon}(n^2)$ 3-edges survive and are zeroed out.*

Proof: Fix a 3-edge that survives. Since U' is a random subset of U of αn unit vectors, e is zeroed out with probability at least $\alpha/(1-\alpha) > \alpha$. By linearity of expectations, there exists an L' for which at least $\alpha^2\Omega_{c,\epsilon}(n^2)$ edges that survive are zeroed out. ■

Fix such an L'. Define a multigraph H on vertex set $N(T)$ as follows. For distinct u, v, there is an edge $\{u, v\}$ for each 3-edge e containing u, v that survives and is zeroed out. Then, by the previous claim, for large enough n the number of edges in H is at least $\alpha^2\lambda n^2$ for a positive constant λ, which depends on c and ϵ. Let $P_1, \ldots, P_r$ be the connected components in H. Let p_j be the number of vertices in P_j.

Lemma 21 *The number of edges in P_j, for any j, is at most $p_j \log p_j + p_j$.*

Proof: Let $\{u, v\}$ be an edge in P_j. Then there is a 3-edge $e = \{w, u, v\} \in M'_i$, $w \in T$, for some $i \in [n]$ for which $u, v \in e$, e survives, and e is zeroed out. Then $\gamma_1 w + \gamma_2 u + \gamma_3 v = u_i$ for non-zero $\gamma_1, \gamma_2, \gamma_3$ in $\mathbb{F}$. Since e survives, $L(u_i) = u_i$. Since $w \in T$, $L(w) = 0$. By linearity, $\gamma_2 L(u) + \gamma_3 L(v) = u_i$. Moreover, for each $i \in [n]$, each vertex $u \in P_j$ can occur in at most one edge $e \in M'_i$, so we obtain matchings W'_i, where an edge $\{u, v\}$ in P_j is in W'_i iff there is a 3-edge $e \in M'_i$ for which $u, v \in e$ and e survives and is zeroed out. By Theorem 18,

$$\sum_i |W'_i| \leq p_j \log p_j + p_j.$$

Since the number of edges in P_j is just $\sum_i |W'_i|$, this completes the proof. ■

We assume that $|N(T)| \leq \alpha^2\lambda n^2/(3 \log\log n)$, as otherwise since $m \geq |N(T)|$, we immediately have the desired bound. Thus, we have the following conditions on the p_j:

1. $\alpha^2\lambda n^2 \leq \sum_j p_j \log p_j + p_j$
2. $\sum_j p_j \leq \alpha^2\lambda n^2/(3 \log\log n)$

We can use the second condition to simplify the first condition to $\alpha^2\lambda n^2/2 \leq \sum_j p_j \log p_j$, which holds for sufficiently large n. In Appendix 9 we show these conditions imply:

Lemma 22 *There exists a set S of $\alpha^2 n$ indices j for which $\sum_{j \in S} p_j \geq \alpha^2 n \log n$.*

Fix such a set S guaranteed by this lemma. Form the set $V(S)$ from S by including exactly one element of each P_j for $j \in S$. Let I be a maximum-sized subset of linearly independent vectors of $V(S) \cup T \cup U'$. Then $|I| \leq |V(S)| + |T| + |U'| \leq \alpha^2 n + \alpha n + \alpha n \leq 3\alpha n$. Extend I to a basis by adding a set of unit vectors J. Define the linear projection:

$$\begin{aligned} L''(v) &= 0 \text{ for all } v \in I \\ L''(v) &= v \text{ for all } v \in J \end{aligned}$$

Claim 23 *Let P be a connected component of H. If a vertex $a \in P$ (identified with a vector) is such that $L''(a) = 0^n$, then all vertices b in P satisfy $L''(b) = 0^n$.*

Proof: Consider any vertex b in P, and let $a = a_0, a_1, a_2, \ldots, a_k = b$ be a path from a to b in P. Since $\{a_0, a_1\}$ is an edge in H, there is a 3-edge $e = \{w, a_0, a_1\}$ in some M'_i for which $w \in T$ and e is zeroed out. This means that $L''(w) = 0$ and $L''(u_i) = 0$. But, for some non-zero $\gamma_1, \gamma_2, \gamma_3 \in \mathbb{F}$, $\gamma_1 w + \gamma_2 a_0 + \gamma_3 a_1 = u_i$. By linearity, these conditions imply that $\gamma_2 L''(a_0) + \gamma_3 L''(a_1) = 0^n$. Thus, if $a = a_0$ satisfies $L''(a) = 0^n$, then $L''(a_1) = 0^n$. By induction, this means that $L''(b) = 0^n$. ■

Corollary 24 *For any $v \in \cup_{j \in S} P_j$, $L''(v) = 0$.*

Proof: This follows from Claim 23 and the fact that L'' vanishes on $V(S)$. ■

Define $P = \cup_{j \in S} P_j$, so that $|P| \geq \alpha^2 n \log n$. Let $N(P)$ be the vertices neighboring P in G'. Each vertex in P has degree at least βn, so it is incident to at least $\beta n - 3\alpha n = \Omega_{c,\epsilon}(n)$ 3-edges in $\cup_{i \in J} M_i$, provided α is a small enough constant. Thus, since any 3-edge is incident to at most 3 elements of P, P is collectively incident to at least $\alpha^2 n \log n(\beta - 3\alpha)n/3 = \alpha^2 \Omega_{c,\epsilon}(n^2 \log n)$ 3-edges in $\cup_{i \in J} M_i$. Since L'' vanishes on P but preserves unit vectors in J, this gives rise to matchings W_i on the multiset of vectors $L''(N(P))$. Here, $N(P)$ is identified with a multiset of vectors, and $L''(N(P))$ is the multiset formed by applying L'' to each element of $N(P)$. Moreover, $\sum_i |W_i| \geq \alpha^2 \Omega_{c,\epsilon}(n^2 \log n)$. By Theorem 18, $|N(P)| \geq \alpha^2 \Omega_{c,\epsilon}(n^2 \log n / \log n)$. Thus, $m \geq |N(P)| \geq \alpha^2 \Omega_{c,\epsilon}(n^2)$. Recall that this is under the assumption that $|N(T)| \leq \alpha^2 \lambda n^2/(3 \log \log n)$. But, $m \geq |N(T)|$. We conclude,

Theorem 25 *For $\delta, \epsilon \in (0,1)$, if $C : \mathbb{F}^n \to \mathbb{F}^m$ is a linear $(3, \delta, \epsilon)$-locally decodable code, then*

$$m = \Omega_{\delta,\epsilon}(n^2/\log \log n),$$

which is independent of the field $\mathbb{F}$.

Acknowledgment: The author thanks Piotr Indyk, Swastik Kopparty, and Sergey Yekhanin for helpful comments.

References

[1] Amos Beimel, Yuval Ishai, Eyal Kushilevitz, and Jean-Franois Raymond. Breaking the $O(n^{\frac{1}{2k-1}})$ barrier for information-theoretic private information retrieval. In *FOCS*, 2002.

[2] Amit Deshpande, Rahul Jain, T. Kavitha, Jaikumar Radhakrishnan, and Satyanarayana V. Lokam. Better lower bounds for locally decodable codes. In *IEEE Conference on Computational Complexity*, pages 184–193, 2002.

[3] Reinhard Diestel. Graph theory. *Springer-Verlag Graduate Texts in Mathematics*, 2005.

[4] Zeev Dvir and Amir Shpilka. Locally decodable codes with 2 queries and polynomial identity testing for depth 3 circuits. In *STOC*, 2005.

[5] Oded Goldreich, Howard J. Karloff, Leonard J. Schulman, and Luca Trevisan. Lower bounds for linear locally decodable codes and private information retrieval. In *CCC*, 2002.

[6] T. Itoh. On lower bounds for the communication complexity of private information retrieval. *IEICE Trans. Fund. of Electronics, Commun. and Comp. Sci*, E84-A(1):157–164, 2001.

[7] Jonathan Katz and Luca Trevisan. On the efficiency of local decoding procedures for error-correcting codes. In *STOC*, 2000.

[8] I. Kerenidis and R. de Wolf. Exponential lower bound for 2-query locally decodable codes. In *STOC*, 2003.

[9] Ashwin Nayak. Optimal lower bounds for for quantum automata and random access codes. In *FOCS*, 1999.

[10] M. A. Nielsen and I. Chuang. Quantum computation and quantum information. *Cambridge University Press*, 2000.

[11] K. Obata. Optimal lower bounds for 2-query locally decodable linear codes. In *RANDOM, 2483: 39-50*, 2002.

[12] A. Razborov and S. Yekhanin. An $\omega(n^{1/3})$ lower bound for bilinear group based private information retrieval. In *FOCS*, 2006.

[13] A. Samorodnitsky. Manuscript, presented at the ipam workshop on locally decodable codes. 2006.

[14] M. Sipser and D. A. Spielman. Expander codes. *IEEE Trans. Inform. Theory, 42:1710-1722*, 1996.

[15] L. Trevisan. Some applications of coding theory in computational complexity. *Quaderni di Matematica 13:347-424*, 2004.

[16] S. Wehner and R. de Wolf. Improved lower bounds for locally decodable codes and private information retrieval. In *ICALP, 3580: 1424-1436*, 2005.

[17] D. Woodruff and S. Yekhanin. A geometric approach to information theoretic private information retrieval. In *CCC, pp. 275-284*, 2005.

[18] S. Yekhanin. New locally decodable codes and private information retrieval schemes. *ECCC TR06-127*, 2006.

5 Appendix: A simplification

Lemma 26 *If C is a (q, c, ϵ)-smooth code that is good on average for which for each $i \in [n]$, the decoder A picks a random q-set $\{j_1, \ldots, j_q\} \in M_i$ and either outputs $C(x)_{j_1} \oplus C(x)_{j_2} \oplus \cdots \oplus C(x)_{j_q}$ or $C(x)_{j_1} \oplus C(x)_{j_2} \oplus \cdots \oplus C(x)_{j_q} \oplus 1$, then there is a $(q, 2c, \epsilon)$-smooth code C' that is good on average for which for each $i \in [n]$ there is a bit $b_i \in \{0, 1\}$ for which the decoder A' picks a random q-set $\{j_1, \ldots, j_q\} \in M_i$ and outputs $C(x)_{j_1} \oplus C(x)_{j_2} \oplus \cdots \oplus C(x)_{j_q} \oplus b_i$.*

Proof: For each $i \in [n]$, for at least half of the q-sets in M_i, either A outputs $C(x)_{j_1} \oplus C(x)_{j_2} \oplus \cdots \oplus C(x)_{j_q}$ or A outputs $C(x)_{j_1} \oplus C(x)_{j_2} \oplus \cdots \oplus C(x)_{j_q} \oplus 1$. In the first case, we set $b_i = 0$ and in the second $b_i = 1$. We remove all q-sets from M_i for which A does not output $C(x)_{j_1} \oplus C(x)_{j_2} \oplus \cdots \oplus C(x)_{j_q} \oplus b_i$. On input i, the new decoder A' of C' picks a random q-set from the remaining ones in M_i, and outputs $C(x)_{j_1} \oplus C(x)_{j_2} \oplus \cdots \oplus C(x)_{j_q} \oplus b_i$. Then C' is a $(q, 2c, \epsilon)$-smooth code that is good on average satisfying the condition of the lemma. ■

6 Appendix: Finding a small set incident to many edges in the recovery graphs

For the hypergraph G of Section 3.1 on m vertices with at least βmn hyperedges, we let $e(G)$ denote the number of its hyperedges and $v(G)$ the number of its vertices. Consider the following algorithm:

Min-Degree(G):

1. $G(0) \leftarrow G$.
2. $x(0) \leftarrow \frac{e(G(0))}{v(G(0))}$.
3. $i \leftarrow 0$.
4. While there is a vertex $v_i \in G(i)$ with $\deg(v_i) < x(i)$,
   - $i \leftarrow i + 1$.
   - $G(i) \leftarrow G(i-1) \setminus \{v_{i-1}\}$.
   - $x(i) \leftarrow \frac{e(G(i))}{v(G(i))}$.
5. Output $G' = G(i)$.

Lemma 27 Min-Degree *outputs a non-empty $G' \subseteq G$ with minimum degree at least βn.*

Proof: It is clear that Min-Degree terminates since step 4 can be iterated at most m times. Suppose then that $G' = G(i')$. Evidently,

$$G = G(0) \supseteq G(1) \supseteq G(2) \supseteq \cdots \supseteq G(i') = G'.$$

When we delete v_{i-1} from $G(i-1)$ to form $G(i)$, we remove $\deg(v_{i-1}) < x(i-1)$ edges and one vertex. It follows that

$$x(i) = \frac{e(G(i))}{v(G(i))} \geq \frac{e(G(i-1))}{v(G(i-1))} = x(i-1),$$

and thus $x(i') \geq \frac{e(G)}{v(G)}$. Note that G' is non-empty since $x(0) > 0$ and $x(i') \geq x(0)$. Thus, since G' has no vertex that can be deleted, it follows that its minimum degree is at least $x(0) \geq \frac{e(G)}{v(G)} \geq \frac{\beta mn}{m} = \beta n$. ■

7 Appendix: The quantum arguments

7.1 Quantum background

We borrow notation from [8]. For more background on quantum information theory, see [10].

A *density matrix* is a positive semi-definite (PSD) complex-valued matrix with trace 1. A *quantum measurement* on a density matrix ρ is a collection of PSD matrices $\{M_j\}$ satisfying $\sum_j M_j^\dagger M_j = I$, where I is the identity matrix ($A^\dagger$ denotes the conjugate-transpose of A). The set $\{M_j\}$ defines a probability distribution X on indices j given by $\Pr[X = j] = \mathrm{tr}(M_j^\dagger M_j \rho)$.

We use the notation AB to denote a bipartite quantum system, given by some density matrix ρ^{AB}, and A and B to denote its subsystems. More formally, the density matrix of ρ^A is $\mathrm{tr}_B(\rho^{AB})$, where tr_B is a map known as the *partial trace* over system B. For given vectors $|a_1\rangle$ and $|a_2\rangle$ in the vector space of A, and $|b_1\rangle$ and $|b_2\rangle$ in the vector space of B,

$$\mathrm{tr}_B(|a_1\rangle\langle a_2| \otimes |b_1\rangle\langle b_2|) \stackrel{\text{def}}{=} |a_1\rangle\langle a_2| \mathrm{tr}(|b_1\rangle\langle b_2|),$$

and $\mathrm{tr}_B(\rho^{AB})$ is then well-defined by requiring tr_B to be a linear map.

$S(A)$ is the *Von Neumann entropy* of A, which is defined to be $\sum_{i=1}^d \lambda_i \log_2 \frac{1}{\lambda_i}$, where the λ_i are the eigenvalues of A. $S(A \mid B) = S(AB) - S(B)$ is the *conditional entropy* of A given B, and $S(A; B) = S(A) + S(B) - S(AB) = S(A) - S(A \mid B)$ is the *mutual information* between A and B.

7.2 Our argument

For $x \in L$, let $C'(x)$ be its encoding. Replace the jth entry of $C'(x)$ with $(-1)^{C'(x)_j}$. We can represent $C'(x)$ as a vector in a state space of $\log m'$ qubits $|j\rangle$, where $j \in [m']$. That is, the vector space it lies in has dimension m', and its standard basis consists of all vectors $|b\rangle$, where $b \in \{0,1\}^{\log m'}$ (we can assume m' is a power of 2). Define

$$\rho_x = \frac{1}{m'} C'(x)^\dagger C'(x).$$

It is easy to verify that ρ_x is a density matrix. Next, we adapt the argument given on p.24 of [8]. Consider the following $n + \log m'$ qubit quantum system XM:

$$\frac{1}{|L|} \sum_{x \in L} |x\rangle\langle x| \otimes \rho_x.$$

We use X to denote the first system, X_i for its individual qubits, and M for the second subsystem. By Theorem 11.8.4 of [10],

$$S(XM) = S(X) + \frac{1}{|L|}\sum_{x\in L} S(\rho_x) \geq S(X) = \log_2 |L|.$$

Since M has $\log m'$ qubits, $S(M) \leq \log m'$, hence

$$S(X:M) = S(X) + S(M) - S(XM) \leq S(M) \leq \log m'.$$

Using a chain rule for relative entropy and a highly non-trivial inequality known as the strong subadditivity of the Von Neumann entropy (for proofs of these facts, see Chapter 11 of [10]), we get

$$S(X \mid M) = \sum_{i=1}^{n} S(X_i \mid X_1, \ldots, X_{i-1}M) \leq \sum_{i=1}^{n} S(X_i \mid M).$$

In Theorem 34 below, we show that if $i \notin I$, then $S(X_i \mid M) \leq 1$, and if $i \in I$, then $S(X_i \mid M) \leq H(\frac{1}{2} + \frac{\epsilon}{8})$. Putting everything together,

$$\begin{aligned}
\log_2 |L| - |I|H(\frac{1}{2} + \frac{\epsilon}{8}) - (n - |I|) &\leq S(X) - \sum_{i=1}^{n} S(X_i \mid M) \\
&\leq S(X) - S(X \mid M) = S(X:M) \\
&\leq \log m' = \log \binom{m}{\eta\frac{m}{n^{h(q)}}} \\
&= O_{q,c,\epsilon}\left(\frac{m}{n^{h(q)}}\log n\right).
\end{aligned}$$

To complete the argument, we would like to show that for a small enough constant α,

$$\log_2 |L| - |I|H(\frac{1}{2} + \frac{\epsilon}{8}) - (n - |I|) = \Omega_{q,c,\epsilon}(n). \tag{1}$$

Recall that $|I| = \Theta_{q,c,\epsilon}(n)$ and $|L| \geq 2^{n-2\alpha n}$. Thus, equation 1 holds if $n - 2\alpha n - |I|H(\frac{1}{2}+\frac{\epsilon}{8}) - n + |I| = |I|(1 - H(\frac{1}{2}+\frac{\epsilon}{8})) - 2\alpha n = \Omega_{q,c,\epsilon}(n)$. This in turn holds for α sufficiently small. Thus, $\Omega_{q,c,\epsilon}(n) = O_{q,c,\epsilon}\left(\frac{m}{n^{h(q)}}\log n\right)$, so $m = \Omega_{q,c,\epsilon}(n^{1+h(q)}/\log n)$. We conclude,

Theorem 28 *For odd q, any (q,c,ϵ)-smooth code $C : \{0,1\}^n \to \{0,1\}^m$ that is good on average satisfies*

$$m = \Omega_{q,c,\epsilon}\left(\frac{n^{1+2/(q-1)}}{\log n}\right).$$

7.3 The missing piece

To complete the proof, it remains to bound $S(X_i \mid M)$. We use Theorem 4.2 of [9], which is a theorem due to Holevo.

Theorem 29 *[9] Let $x \to \sigma_x$ be any quantum encoding of bit strings, let X be a random variable with a distribution given by $\Pr[X = x] = p_x$, and let $\sigma = \sum_x p_x\sigma_x$ be the state corresponding to the encoding of the random variable X. If Y is any random variable obtained by performing a measurement on σ_x, then*

$$I(X;Y) \leq S(\sigma) - \sum_x p_x S(\sigma_x),$$

where $I(X;Y) = H(X) - H(X \mid Y)$ is the classical mutual information between X and Y.

Let q_i be the fraction of different $x \in L$ for which $x_i = 0$.

Lemma 30 *Let Y be any random variable obtained by performing a measurement on the encoding M. Then,*

$$S(X_i \mid M) \leq H(q_i) - I(X_i; Y).$$

Proof: By definition, $S(X_i \mid M) = S(X_i M) - S(M)$. Consider the following two matrices:

$$A = \sum_{x \in L | x_i = 0} \rho_x, \quad B = \sum_{x \in L | x_i = 1} \rho_x.$$

By Theorem 11.8.4 of [10],

$$S(X_i M) = H(q_i) + q_i S(A) + (1 - q_i) S(B).$$

By Theorem 29,

$$I(X_i; Y) \leq S(M) - q_i S(A) - (1 - q_i) S(B).$$

Thus,

$$\begin{aligned} S(X_i M) - S(M) &\leq H(q_i) + q_i S(A) + (1 - q_i) S(B) - I(X_i; Y) - q_i S(A) - (1 - q_i) S(B) \\ &= H(q_i) - I(X_i; Y), \end{aligned}$$

which completes the proof. ■

Corollary 31 *Suppose $i \notin I$. Then $S(X_i \mid M) \leq 1$.*

Now suppose that $i \in I$. We choose a quantum measurement $\{E_j\}$ as follows. Initialize $\{E_j\} \leftarrow \emptyset$. For each of the $m'/4$ disjoint pairs $\{B, B'\} \in W_i$, add the two projections to $\{E_j\}$:

$$\frac{1}{\sqrt{2}}(|B\rangle - |B'\rangle)(\langle B| - \langle B'|), \quad \frac{1}{\sqrt{2}}(|B\rangle + |B'\rangle)(\langle B| + \langle B'|).$$

For the remaining $m'/2$ coordinates B, use the projections $|B\rangle\langle B|$. Observe that $\{E_j\}$ is in fact a quantum measurement, since in an appropriate basis $\sum_j E_j^\dagger E_j = I$. We may identify the first $m'/2$ coordinates of $[m']$ with those sets B occurring in W_i.

Lemma 32 *For $i \in I$, there is an algorithm A that, when measuring ρ_x with quantum measurement $\{E_j\}$, outputs x_i with probability at least $\frac{1}{2} + \frac{\epsilon}{8}$. Here, the probability is over $x \in L$, the randomness of A, and the distribution defined by $\{E_j\}$.*

Proof: When measuring with $\{E_j\}$, with probability

$$\sum_{j > m'/2} \mathrm{tr}(E_j^\dagger E_j \rho_x) = \frac{1}{2},$$

the outcome is in $\{m'/2 + 1, m'/2 + 2, \ldots, m'\}$. In this case, A just outputs a random coin toss.

Now we compute the probability the outcome is j for some $j \in [m'/2]$. We use the notation $j = (B, B', -)$ to refer to the projection $\frac{1}{\sqrt{2}}(|B\rangle - |B'\rangle)(\langle B| - \langle B'|)$ and $j = (B, B', +)$ to refer to the projection $\frac{1}{\sqrt{2}}(|B\rangle + |B'\rangle)(\langle B| + \langle B'|)$. Recall that $\rho_x = \frac{1}{m'} \sum_{B,B'} C'(x)_B \cdot C'(x)_{B'} |B\rangle\langle B'|$ (recall that for all j, we have replaced the jth entry of $C'(x)$ with $(-1)^{C'(x)_j}$). By definition, the probability the outcome is $j = (B, B', -)$ is

$$\begin{aligned} \mathrm{tr}\left(\frac{1}{\sqrt{2}}(|B\rangle - |B'\rangle)(\langle B| - \langle B'|)\frac{1}{\sqrt{2}}(|B\rangle - |B'\rangle)(\langle B| - \langle B'|)\rho_x\right) &= \frac{1}{2}\mathrm{tr}((|B\rangle - |B'\rangle)(\langle B| - \langle B'|)\rho_x), \\ &= \frac{1}{m'} - \frac{1}{m'} C'(x)_B \cdot C'(x)_{B'} \end{aligned}$$

This probability is 0 if $C'(x)_B = C'(x)_{B'}$, and is $\frac{2}{m'}$ otherwise. Similarly, the probability the outcome is $j = (B, B', +)$ is 0 if $C'(x)_B \neq C'(x)_{B'}$, and is $\frac{2}{m'}$ otherwise.

It follows that if $j \in [m'/2]$, A can output $g_B \oplus g_{B'} \oplus b_{i,B,B'}$ for some $\{B, B'\} \in W_i$. In this case, it is correct with probability at least $\frac{1}{2} + \frac{\epsilon}{4}$, by definition of W_i. It follows that A, when measuring ρ_x for random $x \in L$, outputs x_i with probability at least

$$\frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \left(\frac{1}{2} + \frac{\epsilon}{4}\right) = \frac{1}{2} + \frac{\epsilon}{8},$$

which completes the proof. ■

Corollary 33 *Suppose $i \in I$. Then $S(X_i \mid M) \leq H(\frac{1}{2} + \frac{\epsilon}{8})$.*

Proof: Let Y be the output of algorithm A in Lemma 32. By Lemma 30,

$$\begin{aligned} S(X_i \mid M) &\leq H(q_i) - I(X_i; Y) \\ &= H(q_i) - (H(X_i) - H(X_i \mid Y)) \\ &= H(q_i) - H(q_i) + H(X_i \mid Y) \\ &= H(X_i \mid Y). \end{aligned}$$

Now, by Fano's inequality (see p. 536 of [10]), since $\Pr[X_i = Y] \geq \frac{1}{2} + \frac{\epsilon}{8}$,

$$H(X_i \mid Y) \leq H(\frac{1}{2} + \frac{\epsilon}{8}).$$

Thus, $S(X_i \mid M) \leq H(\frac{1}{2} + \frac{\epsilon}{8})$. ■

By combining Corollary 33 and Corollary 31, we have shown the needed missing theorem.

Theorem 34 *If $i \notin I$, then $S(X_i \mid M) \leq 1$. If $i \in I$, then $S(X_i \mid M) \leq H(\frac{1}{2} + \frac{\epsilon}{8})$.*

8 Appendix: From adaptive decoders to non-adaptive decoders

Theorem 35 *Let $C : \mathbb{F}^n \to \mathbb{F}^m$ be a linear $(3, \delta, \epsilon)$-LDC. Then C is also a linear $(3, \delta/9, 2/3 - 1/|\mathbb{F}|)$-LDC with a non-adaptive decoder.*

Proof: Since C is a linear code, each of its coordinates can be identified with a vector $f_j \in \mathbb{F}^n$, with the function for that coordinate computing $\langle f_j, x\rangle$, where the inner product is over $\mathbb{F}$. Define the ordered list of vectors $B = f_1, \ldots, f_m$.

Fix some $i \in [n]$, and let $\mathcal{C}_i$ be the collection of all sets $S \subseteq [m]$, with $|S| \leq 3$, for which $u_i \in \text{span}(f_j \mid j \in S)$, where u_i denotes the unit vector in direction i. Let $D_i \subseteq [m]$ be a smallest dominating set of $\mathcal{C}_i$, that is, a set for which for all $S \in \mathcal{C}_i$, $|S \cap D_i| > 0$.

Claim 36 *$|D_i| > \delta m$.*

Proof: Suppose not. Consider the following adversarial strategy: given a codeword $C(x)$, replace all coordinates $C(x)_j$ for $j \in D_i$, with 0. Denote the new string $\tilde{C}(x)$. The coordinates of $\tilde{C}(x)$ compute the functions $\langle \tilde{f}_j, x\rangle$, where $\tilde{f}_j = f_j$ if $j \notin D_i$, and $\tilde{f}_j = 0$ otherwise. Let $\tilde{B}$ be the ordered list of vectors $\tilde{f}_1, \ldots, \tilde{f}_m$.

Define 3-span$(\tilde{B})$ to be the (possibly infinite) list of all vectors in the span of each subset of $\tilde{B}$ of size at most 3. We claim that $u_i \notin$ 3-span$(\tilde{B})$. Indeed, if not, then let $S \subseteq [m]$ be a smallest set for which $u_i \in \text{span}(\tilde{f}_j \mid j \in S)$. Then $|S| \leq 3$. If S is empty, this is impossible.

Otherwise, $u_i \in \text{span}(f_j \mid j \in S)$, and so $S \cap D_i \neq \emptyset$, so there is some $\ell \in S \cap D_i$. Since $\tilde{f}_\ell = 0$, it follows that $u_i \in \text{span}(\tilde{f}_j \mid j \in (S \setminus \{\ell\}))$. But $|S \setminus \{\ell\}| < |S|$, which contradicts that S was smallest.

Let A be the decoder of C, where A computes $A^y(i, r)$ on input index $i \in [n]$ and random string r. Here, for any $x \in \mathbb{F}^n$, we let the string $y = y(x)$ be defined by the adversarial strategy given above. For any $x \in \mathbb{F}^n$, $A^y(i, r)$ first probes coordinate j_1 of y, learning the value $\langle \tilde{f}_{j_1}, x \rangle$. Next, depending on the answer it receives, it probes coordinate j_2, learning the value $\langle \tilde{f}_{j_2} x \rangle$. Finally, depending on the answer it receives, it probes coordinate j_3, learning the value $\langle \tilde{f}_{j_3} x \rangle$. Consider the affine subspace V of dimension $d \geq n-3$ of all $x \in \mathbb{F}^n$ which cause $A^y(i, r)$ to read positions j_1, j_2, and j_3. Let V_0 be the affine subspace of V of all x for which $A^y(i, r)$ outputs x_i. Since the output of $A^y(i, r)$ is fixed given that it reads positions j_1, j_2, and j_3, and since $u_i \notin \text{span}(\tilde{f}_{j_1}, \tilde{f}_{j_2}, \tilde{f}_{j_3})$, it follows that the dimension of V_0 is at most $d - 1$.

Suppose first that $\mathbb{F}$ is a finite field. Then for any fixed r, the above implies $A^y(i, r)$ is correct on at most a $\frac{1}{|\mathbb{F}|}$ fraction of $x \in \mathbb{F}^n$ since $\frac{|V_0|}{|V|} \leq \frac{1}{|\mathbb{F}|}$ for any set of three indices j_1, j_2, and j_3 that A can read. Thus, by averaging, there exists an $x \in \mathbb{F}^n$ for which

$$\Pr[A^y(i) = x_i] \leq \frac{1}{|\mathbb{F}|},$$

where the probability is over the random coins r of A. This contradicts the correctness of A.

Now suppose that $\mathbb{F}$ is an infinite field. We will show that there exists an $x \in \mathbb{F}^n$ for which

$$\Pr[A^y(i) = x_i] = 0,$$

contradicting the correctness of the decoder.

For each random string r, there is a finite non-empty set G_r of linear constraints over $\mathbb{F}$ that any $x \in \mathbb{F}^n$ must satisfy in order for $A^y(i, r) = x_i$. Consider the union $\cup_r G_r$ of all such linear constraints. Since the number of different r is finite, this union contains a finite number of linear constraints.

Since $\mathbb{F}$ is infinite, we claim that we can find an $x \in \mathbb{F}^n$ which violates all constraints in $\cup_r G_r$. We prove this by induction on n. If $n = 1$, then the constraints have the form $x_1 = c_1, x_1 = c_2, \ldots, x_1 = c_s$ for some finite s. Thus, by choosing $x_1 \notin \{c_1, c_2, \ldots, c_s\}$, we are done. Suppose, inductively, that our claim is true for $n - 1$. Now consider $\mathbb{F}^n$. Consider all constraints in $\cup_r G_r$ that have the form $x_1 = c$ for some $c \in \mathbb{F}$. There are a finite number of such constraints, and we can just choose x_1 not to equal any of these values c, since $\mathbb{F}$ is infinite. Now, substituting this value of x_1 into the remaining constraints, we obtain constraints (each depending on at least one variable) on $n - 1$ variables $x_2, \ldots, x_n$. By induction, we can choose the values to these $n-1$ variables so that all constraints are violated. Since we haven't changed x_1, the constraints of the form $x_1 = c$ are still violated. This completes the proof. ■

It follows that since $|D_i| > \delta m$ and D_i is a smallest dominating set of $\mathcal{C}_i$, we can greedily construct a matching M_i of $\delta m/3$ disjoint triples $\{j_1, j_2, j_3\}$ of $[m]$ for which $u_i \in \text{span}(f_{j_1}, f_{j_2}, f_{j_3})$.

Consider the new behavior of the decoder: on input $i \in [n]$, choose a random triple $\{j_1, j_2, j_3\} \in M_i$, and compute u_i as $\gamma_1 \langle f_{j_1}, x \rangle + \gamma_2 \langle f_{j_2}, x \rangle + \gamma_3 \langle f_{j_3}, x \rangle$, where $u_i = \gamma_1 f_{j_1} + \gamma_2 f_{j_2} + \gamma_3 f_{j_3}$. Since the adversary can now corrupt at most $\delta m/9$ positions, it follows that with probability at least $2/3$, the positions queried by the decoder are not corrupt and it outputs x_i. Note that the new decoder also makes at most 3 queries. ■

9 Appendix: A structural lemma

We have the following conditions on the p_j:

1. $\alpha^2 \lambda n^2/2 \leq \sum_j p_j \log p_j$

2. $\sum_j p_j \le \alpha^2\lambda n^2/(3\log\log n)$.

Lemma 37 *There exists a set S of $\alpha^2 n$ indices j for which $\sum_{j\in S} p_j \ge \alpha^2 n\log n$.*

Proof: Put $s = \alpha^2 n$. We may assume, by relabeling if necessary, that $p_1 \ge p_2 \ge \cdots \ge p_r$. Consider the following program:

$$\begin{aligned} \min \quad & \sum_{j=1}^{s} p_j \\ \text{s.t.} \quad & \alpha^2\lambda n^2/2 \le \sum_j p_j \log p_j \\ & \sum_j p_j \le \alpha^2\lambda n^2/(3\log\log n) \\ & \forall j,\ p_j \ge 0 \end{aligned}$$

We have relaxed the integrality requirements on the p_j, and allowed $p_j = 0$, whereas previously $p_j \ge 1$. This cannot increase the optimum. If $p_j = 0$, we define $p_j \log p_j = 0$ (note that $\lim_{p_j\to 0} p_j\log p_j = 0$, and $p_j \log p_j$ is continuous for $p_j > 0$). Consider an optimal solution $\mathbf{p} = (p_1, \ldots, p_r)$ to this program with cost OPT. If $OPT \ge \alpha^2 n\log n$, the lemma follows, so assume $OPT < \alpha^2 n \log n$.

We claim there is another optimal solution of the form

$$\mathbf{p}' = (OPT - (s-1)p_s, p_s, p_s, \ldots, p_s, q, 0, 0, \ldots, 0),$$

where $0 \le q \le p_s$ and $\sum_j p'_j = \sum_j p_j \le \alpha^2\lambda n^2/(3\log\log n)$. This follows from the convexity of the $x\log x$ function for $x \ge 0$ (where $0\log 0$ is defined to be 0), so that we again have $\alpha^2\lambda n^2/2 \le \sum_j p'_j \log p'_j$. Note that the objective function evaluated at $\mathbf{p}'$ is again OPT.

Let t be the number of positions in $\mathbf{p}'$ whose value is at least p_s. Since $OPT < \alpha^2 n\log n$, we also have $OPT - (s-1)p_s < \alpha^2 n \log n$, and thus, for sufficiently large n we must have,

$$\alpha^2\lambda n^2/3 \le t p_s \log p_s,$$

and also

$$t p_s \le \alpha^2\lambda n^2/(3\log\log n).$$

Putting these inequalities together, this implies

$$\alpha^2\lambda n^2/3 \le \alpha^2\lambda n^2 \log p_s/(3\log\log n),$$

or

$$\log\log n \le \log p_s.$$

On the other hand, since $OPT < \alpha^2 n\log n$, we have $sp_s < \alpha^2 n \log n$, and since $s = \alpha^2 n$, this means $p_s < \log n$. But then $\log\log n \le \log p_s < \log\log n$, a contradiction.

Thus, $OPT \ge \alpha^2 n \log n$, and the proof is complete. ■

ECCC ISSN 1433-8092
http://eccc.hpi-web.de/