Electronic Colloquium on Computational Complexity, Revision 1 of Report No. 114 (2013)

Locally Testable Codes and Cayley Graphs

Parikshit Gopalan* Salil Vadhan† Yuan Zhou‡

October 11, 2013

Abstract

We give two new characterizations of ($\mathbb{F}_2$-linear, smooth) locally testable error-correcting codes in terms of Cayley graphs over $\mathbb{F}_2^h$:

1. A locally testable code is equivalent to a Cayley graph over $\mathbb{F}_2^h$ whose set of generators is significantly larger than h and has no short linear dependencies, but yields a shortest-path metric that embeds into ℓ_1 with constant distortion. This extends and gives a converse to a result of Khot and Naor (2006), which showed that codes with large dual distance imply Cayley graphs that have no low-distortion embeddings into ℓ_1.
2. A locally testable code is equivalent to a Cayley graph over $\mathbb{F}_2^h$ that has significantly more than h eigenvalues near 1, which have no short linear dependencies among them and which "explain" all of the large eigenvalues. This extends and gives a converse to a recent construction of Barak et al. (2012), which showed that locally testable codes imply Cayley graphs that are small-set expanders but have many large eigenvalues.

Keywords: locally testable codes, Cayley graphs, metric embeddings, spectral graph theory, Fourier analysis.

*Microsoft Research Silicon Valley. Email: parik@microsoft.com.

†School of Engineering and Applied Sciences, Harvard University. E-mail: salil@seas.harvard.edu. Work done in part when on leave as a Visiting Researcher at Microsoft Research Silicon Valley and a Visiting Scholar at Stanford University. Supported in part by NSF grant CCF-1116616 and US-Israel BSF grant 2010196.

‡Computer Science Department, Carnegie Mellon University. E-mail: yuanzhou@cs.cmu.edu. Supported in part by a grant from the Simons Foundation (Award Number 252545). Work done in part when visiting Microsoft Research Silicon Valley.

ISSN 1433-8092

1 Introduction

In this work, we show that *locally testable codes* (with "smooth" testers) are equivalent to *Cayley graphs* with certain properties, thereby providing a new perspective from which to approach long-standing open problems about the achievable parameters of locally testable codes.

Before describing these results, we review the basics of both locally testable codes and Cayley graphs.

1.1 Locally Testable Codes

Informally, a *locally testable code (LTC)* is an error-correcting code in which one can distinguish received words that are in the code from those that are far from the code by a randomized test that probes only a few coordinates of the received word. Local testing algorithms for algebraic error-correcting codes (like the Hadamard code and the Reed–Muller code) were developed in the literature on program testing [BLR93, RS96], inspired the development of the field of property testing [GGR98], and played a key role in the constructions of multi-prover interactive proofs and the proof of the PCP Theorem [BFL91, FGL$^+$96, BFLS91, AS98, ALM$^+$98]. Indeed, they are considered to be the "combinatorial core" of PCPs (cf., [GS06, BGH$^+$06]), and thus understanding what is possible and impossible with locally testable codes can point the way to a similarly improved understanding of PCPs. See the surveys [Tre04, Gol11, Ben10].

We focus on the commonly studied case of linear codes over $\mathbb{F}_2$. Thus a *code* is specified by a linear subspace $\mathcal{C} \subset \mathbb{F}_2^n$. n is called the *blocklength* of the code, and $k = \dim(\mathcal{C})$ is its *rate*. The *minimum distance* is $d = \min_{x \neq y \in \mathcal{C}} d(x, y) = \min_{x \in \mathcal{C} - \{0,\}} |x|$, where $d(\cdot, \cdot)$ denotes Hamming distance and $|\cdot|$ denotes Hamming weight. These parameters are typically summarized by referring to $\mathcal{C}$ as an $[n, k, d]_2$ *linear code*.

A *local tester* for $\mathcal{C}$ is a randomized algorithm T that, when given oracle access to a *received word* $r \in \mathbb{F}_2^n$, makes at most a small number q of queries to symbols of r and accepts or rejects. If $r \in \mathcal{C}$, then T^r should accept with high probability (completeness), and if r is "far" from $\mathcal{C}$ in Hamming distance, then T^r should reject with high probability (soundness). It was shown in [BHR05] that any tester for a linear code can be converted into one with the following structure: the tester T randomly samples a string $\alpha \leftarrow \mathcal{D}$ and accepts if $\alpha \cdot r = 0$, where $\mathcal{D} = \mathcal{D}_T$ is some distribution on the *dual code* $\mathcal{C}^\perp = \{\alpha \in \mathbb{F}_2^n : \alpha \cdot c = 0 \forall\, c \in \mathcal{C}\}$. In particular, such a tester has perfect completeness (accepts with probability 1 if $r \in \mathcal{C}$). We say the tester (which is now specified solely by $\mathcal{D}$) has *(strong) soundness* δ if for every $r \in \mathbb{F}_2^n$,

$$\Pr_{\alpha \leftarrow \mathcal{D}}[\alpha \cdot r = 1] \geq \delta \cdot d(r, \mathcal{C}),$$

where $d(r, \mathcal{C}) = \min_{c \in \mathcal{C}} d(r, c)$. This formulation of soundness is often referred to as *strong soundness* in the literature. Typically, we want $\delta = \Omega(1/d)$, where d is the minimum distance of the code, so that received words at distance $\Omega(d)$ from $\mathcal{C}$ are rejected with constant probability. If there are received words at distance $\omega(d)$ from $\mathcal{C}$, then it is common to cap the rejection probability at a constant (e.g. require $\Pr_{\alpha \leftarrow \mathcal{D}}[\alpha \cdot r = 1] \geq \min\{\delta \cdot d(r, \mathcal{C}), 1/3\}$), but we ignore this issue in the introduction for simplicity. An alternate formulation known as *weak soundness* only requires that the test reject with probability $\Omega(1)$ when r is at distance βd from the code, for some fixed constant $\beta < 1/2$.

We are interested in two parameter regimes for LTCs:

Asymptotically Good LTCs Here we seek rate $k = \Omega(n)$, minimum distance $d = \Omega(n)$, soundness $\delta = \Omega(1/d) = \Omega(1/n)$, and query complexity $q = O(1)$. Unfortunately, we do not know whether such codes exist — this is the major open problem about LTCs first posed by [GS06] , and it is closely related to the long-standing open question about whether SAT has constant-query PCPs of linear

length (which would enable proving that various approximation problems require time $2^{\Omega(n)}$ under the exponential-time hypothesis). The closest we have is Dinur's construction [Din07], which has inverse-polylogarithmic (rather than constant) relative rate (i.e. $n = k \cdot \mathrm{polylog}(k)$). A recent result by Viderman additionally achieves strong soundness [Vid13].

Constant-distance LTCs Here we are interested in codes where the minimum distance d is a fixed constant, and the traditional coding question is how large the rate k can be as $n \to \infty$. BCH codes have (optimal) rate $k = n - (d/2) \cdot \log n$, but do not have any local testability properties. Reed–Muller codes yield the best known locally testable codes in this regime, with rate $k = n - O((\log n)^{\log d})$ and query complexity $q = O(n/d)$ to achieve soundness $\delta = \Omega(1/d)$ [BKS+10].[1] (This query complexity is optimal, as $\Omega(n/d)$ queries is needed to detect $d/2$ random corruptions to a codeword with constant probability.) An open problem is whether rate $k = n - c_d \cdot \log n$ is possible, for some constant c_d depending only on d (but not on n).

In terms of limitations of LTCs, there are a number of results shedding light on the structure of an LTC with good parameters (see the survey [Ben10]), but there are essentially no nontrivial upper-bounds on rate known for arbitrary $\mathbb{F}_2$-linear LTCs.

Instead of bounding the query complexity of our LTCs, it is convenient for us to work with *smooth LTCs*, where we simply require that the tester does not query any one coordinate too often. Formally, a tester, specified by a distribution $\mathcal{D}$ on $\mathcal{C}^\perp$, is *ε-smooth* if for every $i \in [n]$, $\Pr_{\alpha \leftarrow \mathcal{D}}[\alpha_i = 1] \leq \varepsilon$. This is analogous to the notion of smooth locally *decodable* codes (LDCs) defined by Katz and Trevisan [KT00]. Like in the case of LDCs, bounding smoothness is closely related to bounding query complexity, where query complexity q corresponds to smoothness $\Theta(q/n)$ (as would be the case for testers that make q uniformly distributed queries). In particular, we want the smoothness to be $\varepsilon = O(1/n)$ in the asymptotically good regime, and $\varepsilon = O(1/d)$ in the constant-distance regime.

As we show in Appendix A, given a code with a q-query tester that gives weak soundness, we can convert it into a code with similar parameters and a tester with smoothness $O(q/d)$ that gives weak soundness. In the reverse direction, we can convert an ε-smooth tester for a code into a $O(\varepsilon n)$ query tester while preserving weak soundness. However, we are unaware of such reductions that preserve strong soundness, and there are examples of codes in the literature (such as those in [Vid13]) which have q-query testers that give strong soundness, but which are not known to have smooth testers with strong soundness. [2] Below we focus mainly on smooth testers with strong soundness, because our result statements are cleanest for such testers, but there are variants of our results for smooth testers with weak soundness.

1.2 Cayley Graphs

Cayley graphs are combinatorial structures associated with finite groups and are useful for applications ranging from pure group theory to reasoning about the mixing rates of Markov chains to explicit constructions of expander graphs [Big93, HLW06]. In this paper, we focus on the case that the group is a finite-dimensional vector space $\mathcal{V}$ over $\mathbb{F}_2$. (So $\mathcal{V} \cong \mathbb{F}_2^h$ for some $h \in \mathbb{N}$.) Given a multiset $S \subseteq \mathcal{V}$, the Cayley (multi)graph $\mathrm{Cay}(\mathcal{V}, S)$ has vertex set $\mathcal{V}$ and edges $(x, x + s)$ for every $s \in S$ (with appropriate multiplicities if S is a multiset). Note that this is an $|S|$-regular undirected graph, since every element of $\mathcal{V}$ is its own additive inverse. If we take S to be a basis of $\mathcal{V}$, then $\mathrm{Cay}(\mathcal{V}, S)$ is simply the h-dimensional hypercube, where $h = \dim \mathcal{V}$. We will be interested in the properties of such graphs when $|S|$ is larger than h.

[1] This is a case where the rejection probabilities need to be capped at a constant, since there may be received words at distance $\omega(d)$ from $\mathcal{C}$.

[2] We are thankful to Or Meir [Mei13] for clarifying this and pointing out an error in a previous version of this paper.

1.3 LTCs and Metric Embeddings of Cayley Graphs

Our first result shows that locally testable codes are equivalent to Cayley graphs with low-distortion metric embeddings into ℓ_1. We refer the reader to [Mat02, Chapter 15] for background on metric embeddings. An *embedding* of a metric space (X_1, d_1) into metric space (X_2, d_2) with *distortion* $c \geq 1$ is a function $f : X_1 \to X_2$ such that for some $\alpha \in \mathbb{R}^+$ and every $x, y \in X_1$, we have

$$\alpha \cdot d_1(x, y) \leq d_2(f(x), f(y)) \leq c \cdot \alpha \cdot d_1(x, y).$$

A commonly studied case is when (X_1, d_1) is the shortest-path metric $d_{\mathcal{G}}$ on a graph $\mathcal{G}$, and (X_2, d_2) is an ℓ_p metric. Indeed, we will take (X_1, d_1) to be the shortest-path metric on a Cayley graph, and (X_2, d_2) to be an ℓ_1 metric. We will use the well-known characterization of ℓ_1 metrics as the cone of the "cut metrics": a finite metric space (X_2, d_2) is an ℓ_1 metric if and only if there is a constant $\alpha \in \mathbb{R}^+$ and a distribution F on boolean functions on X_2 such that for all $x, y \in X_2$, $d_2(x, y) = \alpha \cdot \Pr_{f \leftarrow F}[f(x) \neq f(y)]$.

Note that the shortest-path metric on the hypercube $\mathrm{Cay}(\mathbb{F}_2^h, \{e_1, \ldots, e_h\})$ is an ℓ_1 metric. Indeed, this metric is simply the Hamming distance d, and $d(x, y) = n \cdot \Pr_{i \leftarrow [n]}[x_i \neq y_i]$. We show that the existence of locally testable codes is equivalent to being able to approximate this property (i.e. have low-distortion embeddings into ℓ_1) even when the number of generators is noticeably larger than h. To avoid trivial ways of increasing the number of generators (like duplicating generators, or taking small linear combinations), we will also require that the generators are d-wise linearly independent (i.e. have no linear dependency of length smaller than d).

Theorem 1. *1. If there is an $\mathbb{F}_2$-linear code of blocklength n, rate k, and distance d with an ε-smooth local tester of strong soundness δ, then there is a Cayley graph $\mathcal{G} = \mathrm{Cay}(\mathbb{F}_2^h, S)$ such that $|S| = n$, $h = n - k$, S is d-wise linearly independent, and the shortest path metric on $\mathcal{G}$ embeds into ℓ_1 with distortion at most ε/δ.*

2. If there is a Cayley graph $\mathcal{G} = \mathrm{Cay}(\mathbb{F}_2^h, S)$ such that $|S| = n$, S is d-wise linearly independent, and the shortest path metric on $\mathcal{G}$ embeds into ℓ_1 with distortion at most c, then there is an $\mathbb{F}_2$-linear code of blocklength n, rate $k = n - h$, and distance d with an ε-smooth local tester of strong soundness δ, for some δ and ε such that $\varepsilon/\delta \leq c$.

Note that the theorem provides an exact equivalence between locally testable codes and ℓ_1 embeddings of Cayley graphs, except that the equivalence only preserves the ratio ε/δ rather than the two quantities separately. It turns out that this ratio is the appropriate parameter to measure when considering *strong* soundness. (See Section 3.1.) For weaker notions of soundness, we obtain equivalences with weaker notions of low-distortion embeddings, such as "single-scale embeddings" (where we replace the requirement that $d_2(f(x), f(y)) \geq \alpha d_1(x, y)$ with $d_1(x, y) \geq D \Rightarrow d_2(f(x), f(y)) \geq \alpha D$, see [Lee05] and references therein).

The theorem specializes as follows for the two main parameter regimes of interest:

Corollary 2. *There is an asymptotically good smooth LTC with strong soundness ($k = \Omega(n)$, $d = \Omega(n)$, $\varepsilon/\delta = O(1)$) iff there is a Cayley graph $\mathcal{G} = \mathrm{Cay}(\mathbb{F}_2^h, S)$ with $|S| = (1 + \Omega(1))h$ such that S is $\Omega(h)$-wise linearly independent and the shortest path metric on $\mathcal{G}$ embeds into ℓ_1 with distortion $O(1)$.*

Corollary 3. *For a constant d, there is a distance d smooth LTC of blocklength n with rate $k = n - c_d \log n$ and $\varepsilon/\delta = O(1)$ iff there is a Cayley graph $\mathcal{G} = \mathrm{Cay}(\mathbb{F}_2^h, S)$ with $|S| = 2^{h/c_d}$ such that S is d-wise linearly independent and the shortest path metric on $\mathcal{G}$ embeds into ℓ_1 with distortion $O(1)$.*

To interpret the theorem, let's consider what the conditions on the Cayley graph $\mathcal{G}$ mean. The condition that $|S| = n$ and the elements of S are d-wise independent means that locally, in balls of radius d, the graph $\mathcal{G}$ looks like the n-dimensional hypercube (which embeds into ℓ_1 with no distortion). However, it is squeezed into a hypercube of significantly lower dimension h (which may make even constant distortion impossible).

The canonical example of graphs that do not embed well into ℓ_1 are expanders. Specifically, an n-regular expander on $H = 2^h$ vertices with all nontrivial eigenvalues bounded away from 1 requires distortion $\Omega(h/\log n)$ to embed into ℓ_1. Roughly speaking, the reason is that by Cheeger's Inequality (or the Expander Mixing Lemma), cuts cannot distinguish random neighbors in the graph from random and independent pairs of vertices in the graph, and random pairs of vertices are typically at distance $\Omega(h/\log n)$.[3]

Thus, saying that a graph $\mathcal{G}$ embeds into ℓ_1 with constant distortion intuitively means that $\mathcal{G}$ is very far from being an expander. More precisely, to prove the nonexistence of an ℓ_1 embedding of distortion c amounts to exhibiting a distribution $\mathcal{D}_{\text{close}}$ on edges of $\mathcal{G}$ and a distribution $\mathcal{D}_{\text{far}}$ on pairs of vertices in $\mathcal{G}$ such that for every cut $f : \mathbb{F}_2^h \to \{0,1\}$,

$$\frac{\Pr_{(x,y)\leftarrow\mathcal{D}_{\text{close}}}[f(x) \neq f(y)]}{c} > \frac{\Pr_{(x,y)\leftarrow\mathcal{D}_{\text{far}}}[f(x) \neq f(y)]}{\mathbf{E}_{(x,y)\leftarrow\mathcal{D}_{\text{far}}}[d_{\mathcal{G}}(x,y)]}.$$

As discussed above, if $\mathcal{G}$ were an expander, we could take $\mathcal{D}_{\text{close}}$ to be the uniform distribution on edges and $\mathcal{D}_{\text{far}}$ to be the uniform distribution on pairs of vertices, and deduce a superconstant lower bound on c. Showing an impossibility result for LTCs amounts to finding such expander-like distributions $\mathcal{D}_{\text{close}}$ and $\mathcal{D}_{\text{far}}$ in an arbitrary Cayley graph with a large (size n) set of d-wise linearly independent generators S.

Our construction of a Cayley graph from an LTC in Item 1 of Theorem 1 is a "quotient of hypercube" construction previously analyzed by Khot and Naor [KN06]. Specifically, they showed that if we start from a code $\mathcal{C}$ whose dual code $\mathcal{C}^\perp$ has large minimum distance, then the resulting Cayley graph $\mathcal{G}$ requires large distortion to embed into ℓ_1. Our contributions are to show that we can replace the hypothesis with the weaker condition that $\mathcal{C}$ is not locally testable, and to establish a tight converse by constructing LTCs from Cayley graphs with low-distortion embeddings.

1.4 LTCs and Spectral Properties of Cayley Graphs.

In our second result, we show that locally testable codes are equivalent to Cayley graphs with spectral properties similar to the "ε-noisy hypercube". We call such graphs derandomized hypercubes.

For Cayley graphs over $\mathbb{F}_2$ vector spaces (and more generally abelian groups), the spectrum can be described quite precisely using Fourier analysis. Let M be the transition matrix for the random walk on $\mathrm{Cay}(\mathcal{V}, S)$, i.e. the adjacency matrix divided by $|S|$. Then, regardless of the choice of S, the eigenvectors of M are exactly of the form $\chi_b(x) = (-1)^{b(x)}$ where $b : \mathcal{V} \to \mathbb{F}_2$ ranges over all $\mathbb{F}_2$-linear functions. (If we pick a basis so that $\mathcal{V} = \mathbb{F}_2^h$, then each such linear function is of the form $b(x) = \sum_i b_i x_i$.) The eigenvalue of M associated with χ_b is $(1/|S|) \cdot \sum_{s\in S} \chi_b(s)$. In particular, if S is a λ-biased space [NN93] for λ bounded away from 1, then all the nontrivial eigenvalues have magnitude at most λ, and hence the graph $\mathrm{Cay}(\mathcal{V}, S)$ is an expander. In contrast, for the case of the hypercube ($S = \{e_1, \ldots, e_h\}$ for a basis $e_1, \ldots, e_h$ of $\mathcal{V}$), the eigenvalue associated with $b = (b_1, \ldots, b_h)$ is $1 - 2|b|/h$ where $|b|$ is the Hamming weight of b, so there are $\binom{h}{i}$ eigenvalues of value $1 - 2i/h$.

[3] Actually, for Cayley graphs over $\mathbb{F}_2$ vector spaces, the bound can be improved to $\Omega(h/\log(n/h))$, using the fact that there are at most $\binom{n}{t}$ (rather than n^t) vertices at distance t from any given vertex.

In this section, it will be useful to generalize the notion of Cayley graph from multisets to distributions over $\mathcal{V}$. If S is a distribution over $\mathcal{V}$, then $\mathrm{Cay}(\mathcal{V}, S)$ is a weighted graph where we put weight $\Pr[S = s]$ on the edge $(x, x + s)$ for every $x, s \in \mathbb{F}_2^h$. $\Pr[S = s]$ is also the $(x, x + s)$ entry of the transition matrix of the random walk on $\mathrm{Cay}(\mathcal{V}, S)$. Now, the eigenvalues are $\lambda(b) = \mathbf{E}_{s \leftarrow S}[\chi_b(s)]$. Here a useful example is the ε-noisy hypercube, where $\mathcal{V} = \mathbb{F}_2^h$ and $S = (S_1, \ldots, S_h)$ has each coordinate independently set to 1 with probability ε, and hence the eigenvalues are $\lambda(b) = (1 - 2\varepsilon)^{|b|}$.

Neither the hypercube nor the ε-noisy hypercube are very good expanders, as they have eigenvalues of $1 - 2/h$ and $1 - 2\varepsilon$, respectively, corresponding to eigenvectors χ_b with $|b| = 1$ (which in turn correspond to the "coordinate cuts," partitioning $\mathbb{F}_2^h$ into the sets $\{x : x_i = 1\}$ and $\{x : x_i = 0\}$). However, their spectral properties do imply that small sets expand well. Indeed, Kahn, Kalai, and Linial [KKL88] showed that the indicator vectors of "small" sets in $\mathbb{F}_2^h$ are concentrated on the eigenvectors χ_b where $|b|$ is large, and hence small sets expand well in both the hypercube and ε-noisy hypercube (where "small" is $|\mathcal{V}|^{1-\Omega(1)}$ in the case of the hypercube, and $o(|\mathcal{V}|)$ in the case of the ε-noisy hypercube).

Our spectral characterization of locally testable codes is as follows.

Theorem 4. *There is an $\mathbb{F}_2$-linear code of blocklength n, rate k, and distance d with an ε-smooth local tester of strong soundness δ if and only if there is a Cayley graph $\mathcal{G} = \mathrm{Cay}(\mathbb{F}_2^h, \mathcal{D})$ (for some distribution $\mathcal{D}$ on $\mathbb{F}_2^h$) and a set $S = \{b_1, \ldots, b_n\}$ of linear maps $b_i : \mathbb{F}_2^h \to \mathbb{F}_2$ satisfying:*

1. *$h = n - k$*
2. *S is d-wise linearly independent*
3. *$\lambda(b_i) \geq 1 - 2\varepsilon$ for $i = 1, \ldots, n$.*
4. *For every linear map $b : \mathbb{F}_2^h \to \mathbb{F}_2$, $\lambda(b) \leq 1 - 2\delta \cdot \mathrm{rank}_S(b)$, where $\mathrm{rank}_S(b) = \min\{|T| : T \subseteq S, b = \sum_{i \in T} b_i\}$.*

Let's compare these properties with those of the ε-noisy hypercube. Recall that, in the ε-noisy hypercube, the coordinate cuts $S = \{e_1, \ldots, e_h\}$ are linearly independent and all give eigenvalues of $1 - 2\varepsilon$. And for every b, $\lambda(b) = (1 - 2\varepsilon)^{|b|} = (1 - 2\varepsilon)^{\mathrm{rank}_S(b)} = 1 - \Omega(\varepsilon \cdot \mathrm{rank}_S(b))$ (provided $\mathrm{rank}_S(b) \leq O(1/\varepsilon)$).

Like in our metric embedding result, the main difference here is that we are asking for the set S to be of size larger than h (while retaining d-wise independence among the generators), so we need to squeeze many large eigenvalues into a low-dimensional space. One reason that these spectral properties are interesting is that they imply that the graph $\mathcal{G}$ is a small-set expander for sets of size $|\mathcal{V}|/\exp(d)$(see Lemma 18).

One direction of the above theorem (from LTCs to Cayley graphs) is extracted from the work of Barak et al. [BGH$^+$12], who used locally testable codes (in the constant distance regime) to construct small-set expanders that have a large number of large eigenvalues (as a function of the number $H = 2^h$ of vertices). Such graphs provide barriers to improving the analysis of the Arora–Barak–Steurer algorithm for approximating small-set expansion and unique games [ABS10], and were also used by Barak et al. [BGH$^+$12] to construct improved integrality gap instances for semidefinite programming relaxations of the unique games problem. Our contribution is showing that the connection can be reversed, when formulated appropriately (in terms of spectral properties rather than small-set expansion).

We can specialize Theorem 4 to the two parameter regimes of interest to us (see Corollaries 16 and 17 in Section 4.4 for precise statements). The existence of asymptotically good smooth LTCs with strong soundness is equivalent to the existence of Cayley graphs whose eigenvalue spectrum resembles the n-dimensional Boolean hypercube (for eigenevalues in the range $[0.5, 1]$) but where the number of vertices is $2^{(1-\Omega(1))n}$. In the constant d regime, the existence of $[n, n - c_d \log n, d]_2$ LTCs blocklength n with

smoothness $\varepsilon = O(1/d)$, and soundness $\delta = \Omega(1/d)$ is equivalent to the existence of Cayley graphs whose eigenvalue spectrum resembles the n-dimensional Boolean hypercube (for eigenvalues in the range $[0.5, 1]$) but where the number of vertices is n^{c_d}.

Like our metric embedding result, Theorem 4 and its corollaries have analogues for weaker notions of soundness for the locally testable codes. Specifically, Item 4 changes in a way that is analogous to the soundness condition, for example only requiring that $\lambda(b)$ is small when $\mathrm{rank}_S(b)$ is large.

1.5 Perspective

For many of the problems about constructing codes or Cayley graphs studied in theoretical computer science, the main challenge is finding an *explicit* construction. Indeed, we know that a randomly chosen code has good rate and distance and that a randomly chosen set of generators yields a Cayley graph with high expansion, and much of the research on these topics is aimed at trying to match these parameters with efficient deterministic algorithms.

Locally testable codes (and the equivalent types of Cayley graphs that we formulate) are intriguing in that they combine properties of random objects (such as large distance) with very non-random properties (the existence of a local tester). Thus the major open questions (such as whether there are asymptotically good LTCs) are *existential* — do there even exist objects with the given parameters, regardless of the complexity of constructing them?

Our hope is that the alternative characterizations developed in this paper will be useful in approaching some of these existential questions, either positively (e.g. by using graph operations to construct Cayley graphs with the properties discussed above, analogously to Meir's construction of LTCs [Mei09]) or negatively (e.g. by reasoning about expander-like subgraphs of Cayley graphs, as discussed above in Section 1.3).

The connection between metric embeddings and local testability gives a new perspective on existing results in this area, for instance we use it to give a simple LTC-based proof of the non-embeddability result of Khot and Naor [KN06] (see Section 3.2). Similarly, the connection to derandomized hypercubes has been used by [BGH$^+$12, KM13] to construct improved integrality gap instances for semidefinite programming relaxations of combinatorial optimization problems.

2 Locally Testable Codes revisited

In this section we reformulate the properties of Locally Testable Codes in terms of cosets, which makes our equivalences easier to show.

Recall that a local tester for an $[n, k, d]_2$ binary linear code $\mathcal{C}$ is specified by a a distribution $\mathcal{D}$ on $\mathcal{C}^\perp$. The tester $\mathcal{D}$ is ε-smooth if for every $i \in [n]$, $\Pr_{\alpha \leftarrow \mathcal{D}}[\alpha_i = 1] \leq \varepsilon$. For $v \in \mathbb{F}_2^n$ let $d(v, \mathcal{C}) = \min_{c \in \mathcal{C}} d(v, c)$ and $\mathrm{Rej}(v, \mathcal{D}) = \Pr_{\alpha \leftarrow \mathcal{D}}[\alpha \cdot v = 1]$. We say that $\mathcal{D}$ has soundness δ if $\mathrm{Rej}(v, \mathcal{D}) \geq \delta d(v, \mathcal{C})$ for all $v \in \mathbb{F}_2^n$. We say that an $[n, k, d]_2$ linear code is (ε, δ)-locally testable if it has a tester $\mathcal{D}$ which has smoothness ε and soundness δ. By considering received words at distance 1 from the code, we get $\delta \leq \mathrm{Rej}(e_i, \mathcal{D}) \leq \varepsilon$. The upper bound is an easy consequence of the smoothness. Ideally, we want $\delta = \Omega(\varepsilon)$.

Given $v \in \mathcal{V}$, let $\bar{v} \in \mathcal{V}/\mathcal{C}$ denote the coset of $\mathcal{C}$ containing it. Let $\bar{\mathcal{E}} = \{\bar{e}_1, \ldots, \bar{e}_n\}$ denote the coset representatives of the basis vectors $\{e_1, \ldots, e_n\}$. The $\bar{e_i}$s are not independent over $\mathbb{F}_2$, indeed we have

$$\sum_{i \in S} \bar{e}_i = 0 \iff \sum_{i \in S} e_i \in \mathcal{C}$$

Hence the shortest non-trivial linear dependence is of weight exactly d.

For $\bar{v} \in \mathcal{V}/\mathcal{C}$, there could be several ways to write it as a linear combination over $\bar{\mathcal{E}}$. We have

$$\bar{v} = \sum_{i \in S} \bar{e_i} \iff v + \sum_{i \in S} e_i \in \mathcal{C}$$

Hence if we define $d(\bar{v}, \mathcal{C}) = d(v, \mathcal{C})$ for any $v \in \bar{v}$ (the exact choice does not matter), it follows that $d(\bar{v}, \mathcal{C}) = \mathrm{rank}_{\bar{\mathcal{E}}}(v)$. Similarly, for every $c \in \mathcal{C}$, $v \in \mathcal{V}$ and $\alpha \in \mathcal{C}^{\perp}$, $\alpha(v) = \alpha(v + c)$. Hence

$$\mathrm{Rej}(v, \mathcal{D}) = \mathrm{Rej}(v + c, \mathcal{D}) \tag{1}$$

This lets us define $\mathrm{Rej}(\bar{v}, \mathcal{D}) = \mathrm{Rej}(v, \mathcal{D})$ for any $v \in \bar{v}$.

We can now rephrase smoothness and soundness in terms of coset representatives.

$$\Pr_{\alpha \leftarrow \mathcal{D}}[\alpha_i = 1] = \Pr_{\alpha \leftarrow \mathcal{D}}[\alpha(e_i) = 1] = \mathrm{Rej}(\bar{e_i}, \mathcal{D}) \tag{2}$$

Thus $\mathcal{D}$ is ε-smooth if every $\bar{e_i}$ is rejected with probability at most ε.

We say a set S of vectors in an $\mathbb{F}_2$-linear space is d-wise independent if every $T \subseteq S$ where $|T| < d$ is linearly independent over $\mathbb{F}_2$. For a set of vectors $S = \{s_1, \ldots, s_n\}$ which span a space $\mathcal{T}$, we use $\mathrm{rank}_S(t)$ for $t \in \mathcal{T}$ to denote the smallest k such that t can be expressed as the sum of k vectors from S. With this notation, $\mathcal{D}$ has soundness δ if for every $\bar{v} \in \mathcal{V}/\mathcal{C}$ such that $\mathrm{rank}_{\bar{\mathcal{E}}} \geq d'$, $\mathrm{Rej}(\bar{v}) \geq \delta d'$.

We summarize these observations in the following lemma:

Lemma 5. *Let $\mathcal{C}$ be an $[n, k]_2$ code and let $\mathcal{D}$ be a tester for $\mathcal{C}$.*

- *$\mathcal{C}$ has distance d iff the set $\bar{\mathcal{E}}$ is d-wise independent.*
- *The tester $\mathcal{D}$ is ε-smooth iff $\mathrm{Rej}(\bar{e_i}, \mathcal{D}) \leq \varepsilon$ for all $i \in [n]$.*
- *For $\bar{v} \in \mathcal{V}/\mathcal{C}$, $d(\bar{v}, \mathcal{C}) = \mathrm{rank}_{\bar{\mathcal{E}}}(\bar{v})$. Hence $\mathcal{D}$ has soundness δ iff for every $\bar{v} \in \mathcal{V}/\mathcal{C}$,*

$$\mathrm{Rej}(\bar{v}, \mathcal{D}) \geq \delta \cdot \mathrm{rank}_{\bar{\mathcal{E}}}(\bar{v})$$

3 Locally Testable Codes and Metric Embeddings

Let $S = \{s_1, \ldots, s_n\} \subset \mathbb{F}_2^h$ be set of $n \geq h$ generators of $\mathbb{F}_2^h$ that are d-wise independent. Let $\mathcal{G} = \mathrm{Cay}(\mathbb{F}_2^h, S)$ be the Cayley graph whose edges correspond to the set S. The graph $\mathcal{G}$ can be naturally associated with a code $\mathcal{C}_{\mathcal{G}}$ which consists of all vectors $c = (c_1, \ldots, c_n)$ such that $\sum_i c_i s_i = 0$. It is easy to see that $\mathcal{C}_{\mathcal{G}}$ is an $[n, n - h, d]_2$ linear code.

Similarly, one can start from an $[n, k, d]_2$ linear code $\mathcal{C}$ and construct a Cayley graph $\mathcal{G}_{\mathcal{C}}$ on $\mathbb{F}_2^{n-k}$. We take the vertex set to be $\mathbb{F}_2^n/\mathcal{C}$. We add an edge $(\bar{x}, \bar{y})$ if there exist $x \in \bar{x}$ and $y \in \bar{y}$ such that $d(x, y) = 1$. It is easy to see that is equivalent to taking $S = \{\bar{e_1}, \ldots, \bar{e_n}\}$, and this set is d-wise independent by the distance property of $\mathcal{C}$.

It is easy to see that this construction inverts the previous construction. Henceforth we will fix a code $\mathcal{C}$ and a graph $\mathcal{G}$ that can be derived from one another. The vertex set of $\mathcal{G}$ is given by $V(\mathcal{G}) = \mathbb{F}_2^n/\mathcal{C}$ and the edge set $E(\mathcal{G})$ by $\{\bar{x}, \bar{x} + \bar{e_i}\}$ for $\bar{x} \in \mathbb{F}_2^n/\mathcal{C}$ and $i \in [n]$.

Lemma 6. *Let $d_{\mathcal{G}}$ denote the shortest path metric on $\mathcal{G}$. We have*

$$d_{\mathcal{G}}(\bar{x}, \bar{y}) = d_{\mathcal{G}}(\bar{x} + \bar{y}, 0) = d(x + y, \mathcal{C})$$

Proof: If $d_{\mathcal{G}}(\bar{x}, \bar{y}) \leq d$ then there exists $T \subset [n]$ of size d such that

$$\bar{y} = \bar{x} + \sum_{j \in T} \bar{e}_j \quad \Rightarrow \quad \bar{x} + \bar{y} = \sum_{j \in T} \bar{e}_j$$

hence $d(\bar{x} + \bar{y}, \mathcal{C}) \leq d$. Similarly, if $d(x + y, \mathcal{C}) \leq d$, that gives an $\bar{x}$–$\bar{y}$ path of length d in $\mathcal{G}$. ∎

An ℓ_1-embedding of the shortest path metric $d_{\mathcal{G}}$ on the graph $\mathcal{G}$ is a distribution $\mathcal{D}$ over Boolean functions $f : V(\mathcal{G}) = \mathbb{F}_2^n/\mathcal{C} \to \{\pm 1\}$. The distance $\delta(\bar{x}, \bar{y})$ between a pair of vertices $\bar{x}$ and $\bar{y}$ under this embedding is

$$\delta(\bar{x}, \bar{y}) = \Pr_{f \in \mathcal{D}}[f(\bar{x}) \neq f(\bar{y})]$$

We define the stretch of an edge $(\bar{x}, \bar{y})$ to be the ratio $\delta(\bar{x}, \bar{y})/d(\bar{x}, \bar{y})$. The distortion $c_{\mathcal{D}}$ of the embedding $\mathcal{D}$ is the ratio of the maximum to the minimum stretch of any pair of vertices. It is given by

$$c_{\mathcal{D}} = \frac{\max_{\bar{x}, \bar{y} \in V(\mathcal{G})} \delta(\bar{x}, \bar{y})/d_{\mathcal{G}}(\bar{x}, \bar{y})}{\min_{\bar{x}, \bar{y}} \delta(\bar{x}, \bar{y})/d_{\mathcal{G}}(\bar{x}, \bar{y})}$$

It follows by the triangle inequality that the stretch is maximized by some edge. Hence we get

$$c_{\mathcal{D}} = \frac{\max_{\bar{x} \in V(\mathcal{G}), i \in [n]} \delta(\bar{x}, \bar{x} + \bar{e}_i)}{\min_{x,y} \delta(\bar{x}, \bar{y})/d_{\mathcal{G}}(\bar{x}, \bar{y})} \tag{3}$$

The minimum c achieved over all ℓ_1-embeddings of $\mathcal{G}$ is denoted $c_1(\mathcal{G})$.

Definition 7. *An embedding $\mathcal{D}$ of $\mathcal{G} = \mathrm{Cay}(\mathbb{F}_2^h, S)$ into ℓ_1 is linear if $\mathcal{D}$ is supported on functions $\chi_\alpha(x) = (-1)^{\alpha(x)}$ where α is an $\mathbb{F}_2$-linear function on $V(\mathcal{G})$.*

The space of linear functions on $\mathbb{F}_2^n/\mathcal{C}$ is isomorphic to $\mathcal{C}^\perp$. In a linear embedding, we have

$$\begin{aligned}
\delta(\bar{x}, \bar{y}) &= \Pr_{\alpha \in \mathcal{D}}[\chi_\alpha(\bar{x}) \neq \chi_\alpha(\bar{y})] \\
&= \Pr_{\alpha \in \mathcal{D}}[\chi_\alpha(\bar{x} + \bar{y}) \neq 1] \\
&= \delta(\bar{x} + \bar{y}, 0)
\end{aligned}$$

Thus, the distance δ is invariant under shifting in linear embeddings, just like the shortest path distance $d_{\mathcal{G}}$. Indeed, the next lemma shows that we can replace any embedding by a linear embedding without increasing the distortion.

Lemma 8. *There is a linear embedding $\mathcal{D}$ of $\mathcal{G}$ into ℓ_1 achieving distortion $c_1(\mathcal{G})$.*

To prove this lemma, we set up some machinery. Let $f : \mathbb{F}_2^n/\mathcal{C} \to \{\pm 1\}$ be a Boolean function on $\mathbb{F}_2^n/\mathcal{C}$. We can extend f to a function on all of $\mathbb{F}_2^n$ by setting $f(x) = f(\bar{x})$. (We will henceforth switch freely between both notions). The resulting function is invariant under cosets of $\mathcal{C}$, which implies that its Fourier spectrum is supported on $\mathcal{C}^\perp$. Hence we have

$$f(x) = \sum_{\alpha \in \mathcal{C}^\perp} \hat{f}(\alpha) \chi_\alpha(x)$$

Further, we have $\sum_{\alpha \in \mathcal{C}^\perp} \hat{f}(\alpha)^2 = 1$.

We now proceed to the proof of Lemma 8.

Proof: Given an arbitrary distribution $\mathcal{D}$ on Boolean functions, we define a new distribution $\mathcal{D}'$ on functions where we sample $a \in \mathbb{F}_2^n$ uniformly at random, $f \in \mathcal{D}$, and return the function $f' : \mathbb{F}_2^n \to \{\pm 1\}$ defined by $f'(x) = f(x+a)$. We will show that $c'_{\mathcal{D}} \leq c_{\mathcal{D}}$.

Let $\delta'(x,y) = \Pr_{f' \in \mathcal{D}'}[f'(x) \neq f'(y)]$. We have

$$\begin{aligned}\delta'(x,y) &= \Pr_{a \in \mathbb{F}_2^n, f \in \mathcal{D}}[f(x+a) \neq f(y+a)] \\ &= \mathbf{E}_{a \in \mathbb{F}_2^n}\left[\Pr_{f \in \mathcal{D}'}[f(x+a) \neq f(y+a)]\right] \\ &= \mathbf{E}_{a \in \mathbb{F}_2^n}[\delta(x+a, y+a)].\end{aligned}$$

Let a_1 and a_2 be the values of a that minimize and maximize $\delta(x+a, y+a)$ respectively. Then

$$\delta(x+a_1, y+a_1) \leq \delta'(x,y) \leq \delta(x+a_2, y+a_2)$$

Hence we have

$$\frac{\delta(x+a_1, y+a_1)}{d(x+a_1, y+a_1)} \leq \frac{\delta'(x,y)}{d(x,y)} \leq \frac{\delta(x+a_2, y+a_2)}{d(x+a_2, y+a_2)}$$

since all the denominators are equal. But this implies that

$$\begin{aligned}\min_{x,y} \frac{\delta'(x,y)}{d(x,y)} &\geq \min_{x,y} \frac{\delta(x,y)}{d(x,y)}, \\ \max_{x,y} \frac{\delta'(x,y)}{d(x,y)} &\leq \max_{x,y} \frac{\delta(x,y)}{d(x,y)}\end{aligned}$$

and hence

$$c_{\mathcal{D}'} = \frac{\max_{x,y} \frac{\delta'(x,y)}{d(x,y)}}{\min_{x,y} \frac{\delta'(x,y)}{d(x,y)}} \leq \frac{\max_{x,y} \frac{\delta(x,y)}{d(x,y)}}{\min_{x,y} \frac{\delta(x,y)}{d(x,y)}} = c_{\mathcal{D}}.$$

Next we show that there is a linear embedding $\mathcal{D}''$ with distortion $c_{\mathcal{D}''} = c_{\mathcal{D}'}$. The embedding is simple to describe: we first sample $f \in \mathcal{D}$, we then sample $\chi_\alpha \in \mathcal{C}^\perp$ with probability $\hat{f}(\alpha)^2$. We denote this

distribution on $\mathcal{C}^\perp$ by $\hat{f}^2$. Note that

$$\begin{aligned}
\delta'(x,y) &= \Pr_{a\in\mathbb{F}_2^n, f\in\mathcal{D}}[f(x+a)\neq f(y+a)] \\
&= \mathbf{E}_{f\in\mathcal{D}}\left[\frac{1}{2}\mathbf{E}_{a\in\mathbb{F}_2^n}[1-f(x+a)f(y+a)]\right] \\
&= \mathbf{E}_{f\in\mathcal{D}}\left[\frac{1}{2}\mathbf{E}_{a\in\mathbb{F}_2^n}\left[1-\left(\sum_{\alpha\in\mathcal{C}^\perp}\hat{f}(\alpha)\chi_\alpha(x+a)\right)\left(\sum_{\beta\in\mathcal{C}^\perp}\hat{f}(\beta)\chi_\beta(y+a)\right)\right]\right] \\
&= \mathbf{E}_{f\in\mathcal{D}}\left[\frac{1}{2}\left(1-\sum_{\alpha,\beta\in\mathcal{C}^\perp}\hat{f}(\alpha)\hat{f}(\beta)\chi_\alpha(x)\chi_\beta(y)\mathbf{E}_{a\in\mathbb{F}_2^n}[\chi_\alpha(a)\chi_\beta(a)]\right)\right] \\
&= \mathbf{E}_{f\in\mathcal{D}}\left[\frac{1}{2}\left(1-\sum_{\alpha\in\mathcal{C}^\perp}\hat{f}(\alpha)^2\chi_\alpha(x)\chi_\alpha(y)\right)\right] \\
&= \mathbf{E}_{f\in\mathcal{D}}\left[\sum_{\alpha\in\mathcal{C}^\perp}\hat{f}(\alpha)^2\frac{(1-\chi_\alpha(x)\chi_\alpha(y)}{2}\right] \\
&= \Pr_{f\in\mathcal{D}}\Pr_{\alpha\in\hat{f}^2}[\chi_\alpha(x)\neq\chi_\alpha(y)] \\
&= \delta''(x,y).
\end{aligned}$$

From this it follows that $c''_{\mathcal{D}} = c'_{\mathcal{D}} \leq c_{\mathcal{D}}$.

The lemma follows by taking $\mathcal{D}$ to be the ℓ_1 embedding of $\mathcal{G}$ that minimizes distortion. ∎

We now prove the main result of this section.

Theorem 9. *We have $c_1(\mathcal{G}) \leq c$ iff there exists an (ε,δ)-tester for $\mathcal{C}$ where $\delta \geq \varepsilon/c$.*

Proof: For linear embeddings, we can use shift invariance to simplify the expression for distortion. Since $\mathcal{D}$ is a distribution on $C^\perp$, we can view it as a tester for $\mathcal{C}$. Note that $\mathrm{Rej}(\bar{x},\mathcal{D}) = \delta(\bar{x},0)$. Recall by Equation (3)

$$c_{\mathcal{D}} = \frac{\max_{\bar{x}\in V(\mathcal{G}),i\in[n]}\delta(\bar{x},\bar{x}+\bar{e}_i)}{\min_{x,y}\delta(\bar{x},\bar{y})/d_{\mathcal{G}}(\bar{x},\bar{y})}$$

We can use $\delta(\bar{x},\bar{y}) = \delta(\bar{x}+\bar{y},0) = \mathrm{Rej}(\bar{x}+\bar{y},\mathcal{D})$ to rewrite this as

$$c_{\mathcal{D}} = \frac{\max_{i\in[n]}\mathrm{Rej}(\bar{e}_i,\mathcal{D})}{\min_{\bar{x}\in V(\mathcal{G})}\mathrm{Rej}(\bar{x},\mathcal{D})/d(\bar{x},0)} \tag{4}$$

Given a linear embedding specified by a distribution $\mathcal{D}$ that gives distortion $c_{\mathcal{D}}$, we view $\mathcal{D}$ as a tester. By definition, it has smoothness ε for

$$\varepsilon \geq \max_{i\in[n]}\mathrm{Rej}(\bar{e}_i,\mathcal{D}) \tag{5}$$

and has soundness δ for

$$\delta \leq \min_{\bar{x}\in V(\mathcal{G})}\mathrm{Rej}(\bar{x},\mathcal{D})/d(\bar{x},0) \tag{6}$$

since any such δ satisfies the condition

$$\mathrm{Rej}(\bar{x}, \mathcal{D}) \geq \delta d(\bar{x}, 0).$$

By taking ε, δ to satisfy Equations 5 and 6 with equality, we get $\delta = \varepsilon / c_{\mathcal{D}}$.

In the other direction, assume we have a (ε, δ)-tester for $\mathcal{C}$ where $\delta \geq \varepsilon / c$. Note that ε, δ must satisfy Equations 5 and 6. Plugging these into Equation 4, we get

$$c_{\mathcal{D}} = \frac{\max_{i \in [n]} \mathrm{Rej}(\bar{e}_i, \mathcal{D})}{\min_{\bar{x} \in V(\mathcal{G})} \mathrm{Rej}(\bar{x}, \mathcal{D}) / d(\bar{x}, 0)} \leq \frac{\varepsilon}{\delta} \leq c.$$

■

3.1 Boosting the soundness

Theorem 9 implies the existance of an (ε, δ)-tester for $\mathcal{C}$, where

$$\frac{\varepsilon}{c_1(\mathcal{G})} \leq \delta \leq \varepsilon.$$

While this is the best ratio possible between ε and δ, Theorem 9 does not seem to guarantee the right absolute values for them. In this section, we show that one can achieve this by repeating the tests. First, we identify the right absolute values.

Let t denote the covering radius of $\mathcal{C}$, and let $\bar{x}$ be a codeword at distance t from $\mathcal{C}$. Since

$$1 \geq \mathrm{Rej}(\bar{x}, \mathcal{D}) \geq \delta t$$

we get $\delta \leq 1/t$. Given this upper bound, we would like ε to be $\Theta(1/t)$ and δ to be $\Theta(1/(c_1(\mathcal{G})t))$. We show that this is possible, with a small loss in constants (which we do not attempt to optimize).

Theorem 10. *There is a $(1/(4t), 1/(16c_1(G)t))$-tester for $\mathcal{C}$.*

We defer the proof of this result to Appendix B.

Note that if $d = \Omega(n)$, then d and t differ by a constant factor. However, when $d = o(n)$, it could be that $t = \omega(d)$. In this case, we could relax the soundness requirement for words at distance $\omega(d)$ as follows:

$$\mathrm{Rej}(\bar{x}, \mathcal{D}) \geq \begin{cases} \delta d(\bar{x}, \mathcal{C}) & \text{if } d(\bar{x}, \mathcal{C}) \leq d \\ \delta d & \text{if } d(\bar{x}, \mathcal{C}) \geq d \end{cases}$$

It is possible to get such a tester where $\varepsilon = O(1/d)$, $\delta = \Omega(1/(c_1(\mathcal{G})d))$ using the same argument as above, but replacing t with d. We omit the details.

3.2 Relation to previous work

This equivalence allows us to reformulate results about LTCs in the language of metric embeddings and vice versa. We present two examples where we feel such reformulations are particularly interesting.

Embedding lower bounds from dual distance: Khot and Naor show the following lower bound for $c_1(\mathcal{G})$ in terms of its dual distance.

Theorem 11. *[KN06, Theorem 3.4] Let $\mathcal{C}$ be an $[n, n-h]_2$ code and let $\mathcal{G}$ be the associated Cayley graph. Let $d^\perp$ denote its dual distance. Then*

$$c_1(\mathcal{G}) \geq \Omega\left(d^\perp \frac{h}{n \log(n/h)}\right)$$

In the setting where $h/n = \Omega(1)$ (which is necessary to have constant relative distance), this gives a lower bound of $\Omega(d^\perp)$. Thus their result can be seen as the embedding analogue of the result of BenSasson et al. [BHR05], who showed that the existence of low-weight dual codewords is a necessary condition for local testability. Our results allow for a simple alternative proof of Theorem 11.

Proof of Theorem 11. By Theorem 9, there exists a (ε, δ)-tester $\mathcal{D}$ so that $c_1(\mathcal{G}) = \epsilon/\delta$. We may assume without loss of generality that $\mathcal{D}$ is supported on non-zero code-words in $\mathcal{C}^\perp$, each of which is of weight at least $d^\perp$, so we have $\varepsilon \geq d^\perp/n$.

As in Section 3.1, we have $\delta \leq 1/t$ where t is the covering radius of $\mathcal{C}$. We lower bound t by a standard volume argument:

$$2^{n-h} \cdot \sum_{t'=0}^{t} \binom{n}{t'} \geq 2^n \quad \Rightarrow \quad t = \Omega\left(\frac{h}{\log(n/h)}\right).$$

So we have

$$c_1(\mathcal{G}) \geq \frac{\varepsilon}{\delta} \geq \frac{d^\perp}{n} t = \Omega\left(d^\perp \frac{h}{n \log(n/h)}\right).$$

□

Lower bounds for basis testers: A basis tester for a code $\mathcal{C}$ is a tester $\mathcal{D}$ which is supported on a basis for $\mathcal{C}$. Ben-Sasson et al. showed a strong lower bound for such testers [BSGK+10]. Their main result when restated in our notation says:

Theorem 12. *[BSGK+10, Theorem 5] Let $\mathcal{C}$ be an $[n, k, d]_2$ code with an (ε, δ)-basis tester. Then*

$$\frac{\varepsilon}{\delta} \geq \frac{kd}{3n}.$$

If $\mathcal{C}$ is an $[n, k, d]_2$ code, a basis tester for $\mathcal{C}$ yields an embedding into $(n-k)$-dimensional space. Hence their result implies that any linear embedding of $\mathbb{F}_2^n/\mathcal{C}$ into $(n-k)$-dimensional space requires distortion $\Omega(kd/n)$ (even though $\mathbb{F}_2^n/\mathcal{C}$ has dimension $n-k$ as a vector space over $\mathbb{F}_2$), and hence low-distortion embeddings must have larger support. Note that since our reduction from arbitrary embeddings to linear embeddings could blow up the support, this does not imply a similar bound for arbitrary embeddings.

4 Locally Testable Codes and Derandomized Hypercubes

4.1 Derandomized Hypercubes

We consider Cayley graphs over groups of characteristic 2. Let $\mathcal{A} = \mathbb{F}_2^h$ for some $h > 0$. A distribution $\mathcal{D}$ over $\mathcal{A}$ gives rise to a weighted graph $\mathrm{Cay}(\mathcal{A}, \mathcal{D})$ where the weight of edge (α, β) equals $\mathcal{D}(\alpha + \beta)$.

The symmetry of Cayley graphs makes it easy to compute their eigenvectors and eigenvalues explicitly. Let $\mathcal{A}^*$ denote the space of all linear functions $b : \mathcal{A} \to \mathbb{F}_2$. The characters of the group $\mathcal{A}$ are in 1-1 correspondence with linear functions: $b \in \mathcal{A}^*$ corresponds to a character $\chi_b : \mathcal{A} \to \{\pm 1\}$ given by $\chi_b(\alpha) = (-1)^{b(\alpha)}$. The eigenvectors of $\mathrm{Cay}(\mathcal{A}, \mathcal{D})$ are precisely the characters $\{\chi_b\}_{b \in \mathcal{A}^*}$. The corresponding eigenvalues are given by $\lambda(b) = \mathbf{E}_{\alpha \in \mathcal{D}}[\chi_b(\alpha)]$.

As mentioned earlier, the Cayley graphs we are interested in can be viewed as derandomizations of the ε-noisy hypercube, which retain many of the nice spectral properties of the Boolean hypercube. Before defining them formally, we list these properties that we would like preserved (at least approximately).

1. **Large Eigenvalues.** There are h "top" eigenvectors $\{\chi_{e_i}\}_{i=1}^h$ whose eigenvalues satisfy $\lambda(e_i) \geq 1-\varepsilon$.
2. **Linear Independence.** The linear functions $\{e_1, \ldots, e_h\}$ corresponding to the top eigenvectors are linearly independent over $\mathbb{F}_2$.
3. **Spectral Decay.** For $a \in \mathcal{A}^*$, if $a = \sum_{i \in S} e_i$, then $\lambda(a) \leq (1-\varepsilon)^{|S|}$.

We are interested in Cayley graphs whose threshold rank n is possibly (much) larger than h. But this means that the corresponding dual vectors which lie in the space $\mathcal{A}^*$ of dimension $h < n$ can no longer be linearly independent. So we relax the Linear Independence condition, and only ask that there should be no *short* linear dependencies between these vectors. The Spectral Decay condition will stay the same, except that we need to modify the notion of rank to account for linear dependencies.

Definition 13. *Let $\mathrm{Cay}(\mathcal{A}, \mathcal{D})$ be a Cayley graph on the group $\mathcal{A} = \mathbb{F}_2^h$. Let $\mu, \nu \in [0,1]$ and $d \in \{1, \ldots, n\}$. Let $\mathcal{B}^* = \{b_1, \ldots, b_n\}$ be a d-wise independent set of generators for $\mathcal{A}^*$ of cardinality n. We say that $\mathcal{B}^*$ is a (μ, ν)-spectrum generator for $\mathrm{Cay}(\mathcal{A}, \mathcal{D})$ if it satisfies the following properties:*

- **Large Eigenvalues.** *$\lambda(b) \geq 1 - \mu$ for every $b \in \mathcal{B}^*$.*
- **Spectral Decay.** *For $a \in \mathcal{A}^*$, $\lambda(a) \leq 1 - \nu \cdot \mathrm{rank}_{\mathcal{B}^*}(a)$.*

Note that any set of generators $\mathcal{B}^*$ for $\mathcal{A}^*$ gives us some values of n, d, μ and ν. We would like n, d to be large. Also, applying the Spectral decay condition to $b \in \mathcal{B}^*$, we see that

$$1 - \mu \leq \lambda(b) \leq 1 - \nu$$

hence $\mu \geq \nu$. Ideally, we would like them to be within a constant factor of each other.

We refer to such graphs as "derandomized hypercubes". The reason is that if there is a generating set of size n which is significantly larger than the dimension h, then the resulting graph has spectral properties that resemble the n dimensional hypercube, although it has only $2^h \ll 2^n$ vertices. Every Cayley graph $\mathrm{Cay}(\mathcal{A}, \mathcal{D})$ together with a generating set $\mathcal{B}^*$ gives us a derandomized hypercube, the parameters n, d, μ, ν tell us how good the derandomization is (just like any code $\mathcal{C}$ and dual distribution $\mathcal{D}$ gives us local tester, whose quality is governed by the parameters it achieves).

4.2 Derandomized hypercubes from Locally Testable Codes

Barak et al. proposed the following construction of Derandomized Hypercubes from any Locally Testable Code [BGH$^+$12]. Given $\mathcal{C}$ which is an $[n, k, d]_2$ linear code with a local tester $\mathcal{D}$, they consider the Cayley graph $C(\mathcal{C}^\perp, \mathcal{D})$ on $\mathcal{C}^\perp \cong \mathbb{F}_2^{n-k}$ whose edge weights are distributed according to $\mathcal{D}$.

Theorem 14. *Let $\mathcal{C}$ be an $[n,k,d]_2$ linear code for $d \geq 3$, and let $\mathcal{D}$ be an (ε,δ)-tester for $\mathcal{C}$. There exists a d-wise independent set $\bar{\mathcal{E}}$ of size n which is a $(2\varepsilon, 2\delta)$-spectrum generator for $\mathrm{Cay}(\mathcal{C}^\perp, \mathcal{D})$.*

Proof. Observe that $(\mathcal{C}^\perp)^* \cong \mathcal{V}/\mathcal{C}$. This is because each $v \in \mathcal{V}$ defines a linear function on $\mathcal{C}^\perp$ given by $v(\alpha) = \alpha(v)$, and v, v' define the same function iff the lie in the same coset of $\mathcal{C}$.

We take $\bar{\mathcal{E}} = \{\bar{e}_1, \ldots, \bar{e}_n\}$ to be the cosets corresponding to the received words $e_1, \ldots, e_n$. By Lemma 5, since $\mathcal{C}$ has distance d, the set $\bar{\mathcal{E}}$ is d-wise independence. We will show that it is a $(2\varepsilon, 2\delta)$-spectrum generator for $\mathrm{Cay}(\mathcal{C}^\perp, \mathcal{D})$.

We bound the eigenvalues using the correspondence between the spectrum of $\mathrm{Cay}(\mathcal{C}^\perp, \mathcal{D})$ and the soundness of the tester $\mathcal{D}$ established by Barak et al.. For $\bar{v} \in \mathcal{V}/\mathcal{C}$, let $\chi_{\bar{v}}$ and $\lambda(\bar{v})$ denote the corresponding eigenvalue. [BGH+12, Lemma 4.5] says that

$$\lambda(\bar{v}) = 1 - 2\mathrm{Rej}(\bar{v}, \mathcal{D}). \tag{7}$$

- **Smoothness implies Large Eigenvalues.** By Lemma 5 the smoothness of $\mathcal{D}$ implies $\mathrm{Rej}(\bar{e}_i, \mathcal{D}) \leq \varepsilon$. By Equation 7,

$$\lambda(\bar{e}_i) = 1 - 2\mathrm{Rej}(e_i, \mathcal{D}) \geq 1 - 2\varepsilon.$$

- **Soundness implies Spectral decay.** Fix $\bar{v} \in \mathcal{V}/\mathcal{C}$ so that $\mathrm{rank}_{\mathcal{B}^*}(\bar{v}) \geq d'$. By Lemma 5, the soundness of $\mathcal{D}$ implies $\mathrm{Rej}(\bar{v}, \mathcal{D}) \geq \delta d'$. By Equation 7,

$$\lambda(\bar{v}) = 1 - 2\mathrm{Rej}(\bar{v}, \mathcal{D}) \leq 1 - 2\delta d'.$$

□

4.3 Locally Testable Codes from Derandomized Hypercubes

We show how to start from a Cayley graph on $\mathcal{A} = \mathbb{F}_2^h$ and a set of generators for $\mathcal{A}^*$ and get a locally testable code from it. Our construction takes a Cayley graph $\mathrm{Cay}(\mathcal{A}, \mathcal{D}')$ and a (d, μ, ν)-spectrum generator $\mathcal{B}^*$.

We define the locally testable code $\mathcal{C}$ by specifying the dual code $\mathcal{C}^\perp$ and the tester $\mathcal{D}$. We view elements $\alpha \in \mathcal{A}$ as messages, and embed them into $\mathbb{F}_2^n$ using the map

$$f(\alpha) = (b_1(\alpha), \ldots, b_n(\alpha)). \tag{8}$$

Since $\mathcal{B}^*$ generates $\mathcal{A}^*$, the mapping f is injective. Its image is a h-dimensional subspace of $\mathbb{F}_2^n$ which we denote by $\mathcal{C}^\perp$. The LTC will be $\mathcal{C}$, which is the dual of $\mathcal{C}^\perp$. The distribution $\mathcal{D}'$ on A induces a distribution $\mathcal{D} = f(\alpha)_{\alpha \in \mathcal{D}'}$ on $\mathcal{C}^\perp$, which is the tester for $\mathcal{C}$.

Theorem 15. *Let $\mathrm{Cay}(\mathcal{A} = F_2^h, \mathcal{D}')$ be a Cayley graph and let $\mathcal{B}^* = \{b_1, \ldots, b_n\}$ be a d-wise independent (μ, ν)-spectrum generator for it. Let $\mathcal{C}$ be the dual of the code specified by Equation 8. Then $\mathcal{C}$ is an $[n, n-h, d]_2$ linear code and $\mathcal{D}$ is a $(\mu/2, \nu/2)$-tester for $\mathcal{C}$.*

Proof. It is clear that $\mathcal{C}^\perp$ is an $[n,h]_2$ code, and hence $\mathcal{C}$ is an $[n, n-h]_2$ code. Recall that $f : \mathcal{A} \to \mathcal{C}^\perp$ is an isomorphism. Since $\mathbb{F}_2^n/\mathcal{C} \cong (\mathcal{C}^\perp)^*$, f induces an isomorphism $g : \mathcal{A}^* \to \mathbb{F}_2^n/\mathcal{C}$, with property that for $a \in \mathcal{A}^*$ and $\alpha \in \mathcal{A}$,

$$g(a)(f(\alpha)) = a(\alpha). \tag{9}$$

We observe that $g(b_i) = \bar{e}_i$, since

$$\bar{e}_i(f(\alpha)) = e_i \cdot f(\alpha) = b_i(\alpha).$$

Since $\mathcal{B}^*$ is d-wise independent, so is $\bar{\mathcal{E}}$, which by Lemma 5 implies that $\mathcal{C}$ has distance d. This also implies that for any $a \in \mathcal{A}^*$,

$$a = \sum_{i \in S} b_i \iff g(a) = \sum_{i \in S} \bar{e}_i$$

Hence by Lemma 5, we have $d(g(a), \mathcal{C}) = \mathrm{rank}_{\mathcal{B}^*}(a)$. We can now deduce the local testability of $\mathcal{C}$ from the spectral properties of $\mathcal{B}^*$.

- **Large Eigenvalues Imply Smoothness.** In order to bound the smoothness of $\mathcal{D}$ we need to bound

  $$\mathrm{Rej}(\bar{e}_i, \mathcal{D}) = \Pr_{\alpha \in \mathcal{D}'}[\bar{e}_i \cdot f(\alpha) = 1] = \Pr_{\alpha \in \mathcal{D}'}[b_i(\alpha) = 1]$$

  We have

  $$1 - \mu \leq \lambda(b_i) = \mathbf{E}_{\alpha \in \mathcal{D}}[(-1)^{b_i(\alpha)}] = 1 - 2 \Pr_{\alpha \in \mathcal{D}}[b_i(\alpha) = 1]$$

  which implies that

  $$\mathrm{Rej}(\bar{e}_i, \mathcal{D}) \geq \frac{\mu}{2}.$$

- **Spectral decay implies soundness.**

  Consider $\bar{v} \in \mathbb{F}_2^n / \mathcal{C}$ such that $d(\bar{v}, \mathcal{C}) \geq d'$. Let $\bar{v} = g(a)$ for $a \in \mathcal{A}^*$, so that $\mathrm{rank}_{\mathcal{B}^*}(g(a)) \geq d'$. From the spectral decay property of $\mathcal{B}^*$,

  $$1 - \nu d' \geq \lambda(a) = \mathbf{E}_{\alpha \in \mathcal{D}'}[(-1)^{a(\alpha)}] = 1 - 2 \Pr_{\alpha \in \mathcal{D}'}[a(\alpha) = 1]$$

  hence

  $$\Pr_{\alpha \in \mathcal{D}'}[a(\alpha) = 1] \geq \frac{\nu d'}{2}.$$

  The soundness of the tester follows by noting that

  $$\mathrm{Rej}(\bar{v}, \mathcal{D}) = \Pr_{\alpha \in \mathcal{D}'}[g(a) \cdot f(\alpha) = 1] = \Pr_{\alpha \in \mathcal{D}'}[a(\alpha) = 1]$$

□

4.4 Some consequences of this equivalence

This equivalence lets us reformulate questions regarding LTCs as questions regarding the existence of certain families of derandomized hypercubes.

Corollary 16. *There exists an asymptotically good family of codes $\{\mathcal{C}_n\}$ where $\mathcal{C}_n$ has blocklength n and is $(O(1/n), \Omega(1/n))$-locally testable iff for infintitely many h there exists a Cayley graph $\mathcal{G}_h = \mathrm{Cay}(\mathbb{F}_2^h, \mathcal{D})$ and a set $\mathcal{B}^*$ of generators for $(\mathbb{F}_2^h)^*$ such that*

- *the elements of $\mathcal{B}^*$ are $(\rho_0 h)$-wise independent for $\rho_0 > 0$,*
- *$|\mathcal{B}^*| \geq \rho_1 h$ for $\rho_1 > 1$,*
- *$\mathcal{B}^*$ is an $(O(1/h), \Omega(1/h))$-spectrum generator for $\mathcal{G}_h$.*

Corollary 17. *Let $d \geq 3$. There exists an asymptotic family of codes $\{\mathcal{C}_n\}$ where $\mathcal{C}_n$ has parameters $[n, n - c_d \log n, d]_2$ and is $(\varepsilon, \Omega(\varepsilon))$-locally testable iff for infinitely many h there exists a Cayley graph $\mathcal{G}_h = \mathrm{Cay}(\mathbb{F}_2^h, \mathcal{D})$ and a set $\mathcal{B}^*$ of generators for $(\mathbb{F}_2^h)^*$ such that*

- *the elements of $\mathcal{B}^*$ are d-wise independent,*
- *$|\mathcal{B}^*| \geq 2^{h/c_d}$,*
- *$\mathcal{B}^*$ is an $(\varepsilon, \Omega(\varepsilon))$-spectrum generator for $\mathcal{G}_h$.*

Next we show that derandomized hypercubes are small-set expanders. We say that a regular graph G with n vertices is a (τ, ϕ)-expander if for every set S of at most τn vertices, at least a fraction ϕ of the edges incident to S leave S (i.e. are on the boundary between S and $\overline{S}$).

The following lemma says that if a graph has a (μ, ν) spectrum generator, then it is a (τ, ϕ)-expander for appropriately chosen τ and ϕ. The lemma is proved in [BGH$^+$12]. Since our terminology and notation is different, we present a proof of the Lemma in Appendix C.

Lemma 18. *[BGH$^+$12] Let $G = \mathrm{Cay}(\mathcal{A}, \mathcal{D})$ be a Cayley graph on the group $\mathcal{A} = \mathbb{F}_2^h$. Let $\mathcal{B}^*$ be a d-wise independent set which is a (μ, ν)-spectrum generator for G. Then G is a (τ, ϕ_τ) expander for $\phi_\tau = \nu d/4 - 3^{d/2}\tau^{1/4}$.*

To interpret the expansion bound, think of $\nu d/4 = \Omega(1)$ (we can assume that the graphs obtained from LTCs have this property, since this is analogous to saying that words at distance $d/4$ are rejected with constant probability). So if we take $\tau = \exp(-d)$, then $\phi_\tau = \Omega(1)$. A particularly interesting instantiation of this bound is obtained by combining Corollary 17 and Lemma 18:

Corollary 19. *For $d \geq 3$, suppose there exists an asymptotic family of codes $\{\mathcal{C}_n\}$ where $\mathcal{C}_n$ has parameters $[n, n - c_d \log n, d]_2$ and is $(O(1/d), \Omega(1/d))$-locally testable. There for infinitely many h there exists a Cayley graph $\mathcal{G}_h = \mathrm{Cay}(\mathbb{F}_2^h, \mathcal{D})$ such that $\mathcal{G}_h$ is $(O(9^{-d}), \Omega(1))$-expander and has $2^{h/c_d}$ eigenvalues greater than $1 - O(1/d)$.*

In contrast, Arora et al. [ABS10] showed that if G is an $(\tau, \Omega(1))$-expander, then there are at most $n^{O(\epsilon)}/\tau$ eigenvalues greater than $1 - \epsilon$. Their bound implies that the graph $\mathcal{G}_h$ obtained in Corollary 19 can have at most $2^{O(h/d)}$ eigenvalues greater than $1 - O(1/d)$. If there exist LTCs where $c_d = O(d)$, the resulting graphs $\mathcal{G}_h$ would meet the ABS bound. The only lower bound we know of for c_d is $c_d \geq d/2$ by the Hamming bound.

Acknowledgements

We thank Or Meir for pointing out an error in an earlier version of this paper, regarding the relationship between smoothness and bounded query complexity for locally testable codes. We thank Alex Andoni, Anupam Gupta and Kunal Talwar for useful discussions and pointers to the literature on metric embeddings. And we thank Raghu Meka, Prasad Raghavendra and Madhu Sudan for helpful discussions.

References

[ABS10] Sanjeev Arora, Boaz Barak, and David Steurer. Subexponential algorithms for Unique Games and related problems. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science*, pages 563–572, 2010.

[ALM+98] Sanjeev Arora, Carsten Lund, Rajeev Motwani, Madhu Sudan, and Mario Szegedy. Proof verification and the hardness of approximation problems. *Journal of the ACM*, 45(3):501–555, May 1998.

[AS98] Sanjeev Arora and Shmuel Safra. Probabilistic checking of proofs: A new characterization of NP. *Journal of the ACM*, 45(1):70–122, January 1998.

[Ben10] Eli Ben-Sasson. Property testing. chapter Limitation on the rate of families of locally testable codes, pages 13–31. Springer-Verlag, Berlin, Heidelberg, 2010.

[BFL91] László Babai, Lance Fortnow, and Carsten Lund. Nondeterministic exponential time has two-prover interactive protocols. *Computational Complexity*, 1(1):3–40, 1991.

[BFLS91] László Babai, Lance Fortnow, Leonid A. Levin, and Mario Szegedy. Checking computations in polylogarithmic time. In Cris Koutsougeras and Jeffrey Scott Vitter, editors, *STOC*, pages 21–31. ACM, 1991.

[BGH+06] Eli Ben-Sasson, Oded Goldreich, Prahladh Harsha, Madhu Sudan, and Salil Vadhan. Robust PCPs of proximity, shorter PCPs and applications to coding. *SIAM Journal on Computing*, 36(4):889–974, 2006.

[BGH+12] Boaz Barak, Parikshit Gopalan, Johan Håstad, Raghu Meka, Prasad Raghavendra, and David Steurer. Making the long code shorter, with applications to the Unique Games Conjecture. In *Proceedings of the 53rd Annual IEEE Symposium on Foundations of Computer Science*, 2012.

[BHR05] Eli Ben-Sasson, Prahladh Harsha, and Sofya Raskhodnikova. Some 3CNF properties are hard to test. *SIAM J. Comput.*, 35(1):1–21 (electronic), 2005.

[Big93] Norman Biggs. *Algebraic graph theory*. Cambridge University Press, 1993.

[BKS+10] Arnab Bhattacharyya, Swastik Kopparty, Grant Schoenebeck, Madhu Sudan, and David Zuckerman. Optimal testing of reed-muller codes. In *FOCS*, pages 488–497, 2010.

[BLR93] Manuel Blum, Michael Luby, and Ronitt Rubinfeld. Self-testing/correcting with applications to numerical problems. *Journal of Computer and System Sciences*, 47(3):549–595, 1993.

[BSGK+10] Eli Ben-Sasson, Venkatesan Guruswami, Tali Kaufman, Madhu Sudan, and Michael Viderman. Locally testable codes require redundant testers. *SIAM J. Comput.*, 39(7):3230–3247, 2010.

[Din07] Irit Dinur. The PCP theorem by gap amplification. *Journal of the ACM*, 54(3):article 12, 44 pages (electronic), 2007.

[FGL+96] Uriel Feige, Shafi Goldwasser, Laszlo Lovász, Shmuel Safra, and Mario Szegedy. Interactive proofs and the hardness of approximating cliques. *Journal of the ACM*, 43(2):268–292, 1996.

[GGR98] Oded Goldreich, Shafi Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. *Journal of the ACM*, 45(4):653–750, 1998.

[Gol11] Oded Goldreich. Studies in complexity and cryptography. chapter Short locally testable codes and proofs, pages 333–372. Springer-Verlag, Berlin, Heidelberg, 2011.

[GS06] Oded Goldreich and Madhu Sudan. Locally testable codes and PCPs of almost-linear length. *Journal of the ACM*, 53(4):558–655 (electronic), 2006.

[HLW06] S. Hoory, N. Linial, and A. Wigderson. Expander graphs and their applications. *Bull. Amer. Math Soc.*, 43:439–561, 2006.

[KKL88] Jeff Kahn, Gil Kalai, and Nathan Linial. The influence of variables on Boolean functions (extended abstract). In *29th Annual Symposium on Foundations of Computer Science*, pages 68–80, White Plains, New York, 24–26 October 1988. IEEE.

[KM13] Daniel M. Kane and Raghu Meka. A prg for lipschitz functions of polynomials with applications to sparsest cut. In *STOC*, pages 1–10, 2013.

[KN06] S. Khot and A. Naor. Nonembeddability theorems via Fourier analysis. *Mathematische Annalen*, 334(4):821–852, 2006.

[KT00] Jonathan Katz and Luca Trevisan. On the efficiency of local decoding procedures for error-correcting codes. In *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing*, pages 80–86 (electronic), New York, 2000. ACM.

[Lee05] James Lee. On distance scales, embeddings, and efficient relaxations of the cut cone. In *Proc. 16th ACM-SIAM Symposium on Discrete Algorithms*, pages 92–101, 2005.

[Mat02] Jiri Matousek. *Lectures on Discrete Geoemetry*. Springer, GTM, 2002.

[Mei09] Or Meir. Combinatorial construction of locally testable codes. *SIAM Journal on Computing*, 39(2):491–544, 2009.

[Mei13] Or Meir. Personal communication, 2013.

[NN93] Joseph Naor and Moni Naor. Small-bias probability spaces: Efficient constructions and applications. *SIAM Journal on Computing*, 22(4):838–856, August 1993.

[RS96] Ronitt Rubinfeld and Madhu Sudan. Robust characterizations of polynomials with applications to program testing. *SIAM Journal on Computing*, 25(2):252–271, 1996.

[Tre04] Luca Trevisan. Some applications of coding theory in computational complexity. *Quaderni di Matematica*, 13:347–424, 2004.

[Vid13] Michael Viderman. Strong LTCs with inverse poly-log rate and constant soundness. In *To appear in FOCS 2013*, 2013.

A Smoothness versus Bounded Query Complexity

In this section, we describe an equivalence between constructing LTCs with smooth testers and constructing LTCs with bounded query complexity. The equivalence is only for *weak soundness*, so we define that more formally now.

Definition 20. *Let $\mathcal{C} \subseteq \mathbb{F}_2^n$ be a linear code, and $\mathcal{D}$ a distribution on $\mathcal{C}^\perp$ (which we view as a tester for $\mathcal{C}$). For $\delta > 0$, we say that $\mathcal{D}$ has* weak soundness *at distance d', if for every received word $r \in \mathbb{F}_2^n$ such that $d(r, \mathcal{C}) \geq d'$, we have $\Pr_{\alpha \leftarrow \mathcal{D}}[\alpha \cdot r = 1] \geq 1/2$.*

Thus, a tester with strong soundness δ has weak soundness at distance $d' = 1/2\delta$, but the converse does not hold. If d is the minimimum distance of the code, we typically want strong soundness at distance $d' = \beta d$ for an arbitrarily small constant β. (If β is too large, e.g. $\beta \geq 1/2$, then there may be no received words at distance at least βd from the code, and weak soundness becomes vacuous.)

Now we note that it is easy to convert an ε-smooth tester into one with bounded query complexity. Since an ε-smooth tester for a code queries at most εn coordinates in expectation, we can get a tester with query complexity $q = O(\varepsilon n)$ by discarding tests that query more than q coordinates. However, this transformation costs an *additive* constant (namely $\varepsilon n/q$) in the soundness probability, and hence it only preserves weak soundness. (We can amplify the soundness probability back to $1/2$ by repeating the tester a constant number of times.)

The other direction is more involved and requires a modification to the underlying code.

Theorem 21. *Let $\mathcal{C}$ be an $[n, k, d]_2$ code that has a q-query tester $\mathcal{D}$ giving weak soundness at distance d'. Then for every $d'' \in \mathbb{N}$ there exists an $[n', k, d]_2$ code $\mathcal{C}'$ that has a q/d''-smooth tester $\mathcal{D}'$ giving weak soundness at distance $d' + d''$, with blocklength $n' \leq n + d''$.*

Thus taking $d'' = \Theta(d')$, we only increase the distance at which we have soundness by a constant factor, and obtain smoothness $\varepsilon = O(q/d')$. In an asymptotically good LTC, we have $d' = \Omega(n)$, so smoothness $\varepsilon = O(q/n)$ is indeed equivalent to query complexity $O(q)$. When $d' = o(n)$, the relationship is not tight, and it would be nice to improve Theorem 21 to yield smoothness $O(q/n')$ in general.

Proof: Let $\mathcal{C}$ be an $[n, k, d]_2$ code that has a q-query tester $\mathcal{D}$ with weak soundness at distance d'. Our goal is to construct from it a smooth LTC with comparable parameters. Let ε_i denote the probability that $\mathcal{D}$ queries coordinate i. We define a new code $\mathcal{C}'$ by replicating coordinate i

$$m_i = \left\lceil \frac{\varepsilon_i d''}{q} \right\rceil$$

times. This operation cannot decrease the minimum distance of the code.

The tester $\mathcal{D}'$ for $\mathcal{C}'$ is obtained by first sampling a test α according to $\mathcal{D}$ and then for each i in the support of α, choosing one of the m_i copies at random. Now each copy of coordinate i is queried with probability $\varepsilon_i/m_i \leq q/d''$, so the resulting tester $\mathcal{D}'$ is q/d''-smooth.

We bound the blocklength of $\mathcal{C}'$ as follows:

$$\begin{aligned} n' &= \sum_{i=1}^{n} m_i \\ &\leq \sum_{i=1}^{n} \left(1 + \frac{\varepsilon_i d''}{q}\right) \\ &= n + \frac{d''}{q} \cdot \sum_{i=1}^{n} \varepsilon_i \\ &\leq n + d'', \end{aligned}$$

where the last inequality follows from the fact that $\mathcal{D}$ is a q-query tester. We now analyze the soundness. Running a test according to $\mathcal{D}'$ on a received word r' is the same as running a test from $\mathcal{D}$ on a distribution $\mathcal{R}(r')$ of recieved words in $\mathcal{C}$, where we sample $r \in \mathcal{R}(r')$ by setting the i-th coordinate in r randomly to one of the m_i coordinates of r' which are meant to be copies of coordinate i. Note that if r' is actually obtained by replicating coordinates of some received word r, then $\mathcal{R}(r')$ is supported entirely on r.

Now consider a received word r' where $d(r', \mathcal{C}') \geq d' + d''$. We claim that every r in the support of $\mathcal{R}(r')$ satisfies

$$d(r, \mathcal{C}) \geq d'.$$

Suppose there exists an $r \in \mathcal{R}(r')$ and $c \in \mathcal{C}$ so that $d(r, c) < d'$; equivalently r and c agree in more than $n - d'$ coordinates. Then consider the codeword $c' \in \mathcal{C}'$ that is obtained by replicating coordinates of c. Then r' and c' agree in at least as many coordinates as r and c, so

$$d(r', c') < n' - (n - d') \leq (n + d'') - (n - d') = d' + d'',$$

which contradicts our assumption that $d(d', \mathcal{C}') \geq d' + d''$.

Thus every every $r \in \mathcal{R}(r')$ is at distance at least d' from $\mathcal{C}$ and hence is rejected with probability at least $1/2$ by $\mathcal{D}$. Hence $\mathcal{D}'$ rejects words at distance $d' + d''$ from $\mathcal{C}'$ with probability at least $1/2$, and thus has weak soundness at distance $d' + d''$. ∎

B Proof of Theorem 10

Let $\mathcal{D}^{\oplus \ell}$ denote the distribution on $\mathcal{C}^{\perp}$ where we sample ℓ independent codewords according to $\mathcal{D}$ and add them. We claim that for suitable ℓ, both ε and δ scale by roughly a factor of ℓ.

Lemma 22. *Let ℓ be such that*

$$\ell \leq \frac{1}{4\mathrm{Rej}(\bar{x}, \mathcal{D})} \text{ for all } \bar{x} \in \mathbb{F}_2^n / \mathcal{C}.$$

Then $\mathcal{D}^{\oplus \ell}$ is an $(\varepsilon\ell, \delta\ell/2)$-tester for $\mathcal{C}$.

Proof: The tester $\mathcal{D}^{\oplus \ell}$ is an $\ell\varepsilon$-smooth tester by the union bound. Its soundness can be analyzed by noting that

$$1 - 2\mathrm{Rej}(\bar{x}, \mathcal{D}^{\oplus \ell}) = (1 - 2\mathrm{Rej}(\bar{x}, \mathcal{D}))^{\ell}$$

Using the bound on ℓ to truncate the RHS, we get

$$\mathrm{Rej}(\bar{x}, \mathcal{D}^{\oplus \ell}) \geq \frac{\ell}{2}\mathrm{Rej}(\bar{x}, \mathcal{D}) \geq \frac{\ell\delta}{2} d(\bar{x}, 0). \tag{10}$$

■

We use this to prove Theorem 10.

Proof of Theorem 10. We start with an (ε, δ)-tester where $\delta \geq \varepsilon / c_1(\mathcal{G})$. If δ exceeds the claimed bound, we are already done. Assume this is not true, so

$$\delta \leq \frac{1}{16 c_1(\mathcal{G}) t},\ \varepsilon \leq c_1(\mathcal{G})\delta \leq \frac{1}{16t}$$

Since the covering radius is t, we have that for every $\bar{x} \in \mathbb{F}_2^n$,

$$\mathrm{Rej}(\bar{x}, \mathcal{D}) \leq t\varepsilon \leq 1/16$$

Let $\ell = \lfloor 1/(4t\varepsilon) \rfloor$ so that

$$\frac{1}{8t\varepsilon} \leq \ell \leq \frac{1}{4t\varepsilon}.$$

By Lemma 22, $\mathcal{D}^{\oplus \ell}$ has smoothness ε' where

$$\varepsilon' \leq \ell\varepsilon \leq \frac{1}{4t}$$

and soundness δ' where

$$\delta' \geq \frac{1}{2}\ell\delta \geq \frac{1}{2}\frac{1}{8t\varepsilon}\frac{\varepsilon}{c_1(G)} \geq \frac{1}{16 t c_1(G)}.$$

□

C Proof of Lemma 18

We first show the follow hypercontractive inequality:

Claim 23. *Let $B = \{b_1, \ldots, b_n\} \subseteq \mathbb{F}_2^h$ be $(4d+1)$-wise independent. For every function $f : \mathbb{F}_2^h \to \mathbb{R}$ defined as*

$$f(x) = \sum_{S \subseteq [n], |S| \leq d} \hat{f}(S) \prod_{i \in S} \chi_{b_i}(x),$$

we have

$$\mathbf{E}_x[f(x)^4] \leq 9^d \left(\mathbf{E}_x[f(x)^2]\right)^2.$$

Proof: Let $g : \mathbb{F}_2^n \to \mathbb{R}$ be

$$g(y) = \sum_{S \subseteq [n], |S| \leq d} \hat{f}(S) \prod_{i \in S} y_i.$$

The statement is proved by the standard $(2,4)$-hypercontractive inequality (applied to g function) and the observation that

$$\mathbf{E}_x[f(x)^2] = \mathbf{E}_y[g(y)]^2 = \sum_{S \subseteq [n], |S| \leq d} \hat{f}(S)^2,$$

and

$$\mathbf{E}_x[f(x)^4] = \mathbf{E}_y[g(y)^4] = \sum_{\substack{|S_1|,|S_2|,|S_3|,|S_4|\leq d \\ S_1\Delta S_2\Delta S_3\Delta S_4=\emptyset}} \hat{f}(S_1)\hat{f}(S_2)\hat{f}(S_3)\hat{f}(S_4).$$

■

We now proceed to prove Lemma 18.

Proof of Lemma 18. For any two functions $f, g : \mathbb{F}_2^h \to \mathbb{R}$, define their inner-product as

$$\langle f, g\rangle = \mathbf{E}_{x\in\mathbb{F}_2^h}[f(x)g(x)]$$

and the p-norm of f to be

$$\|f\|_p = \left(\mathbf{E}_{x\in\mathbb{F}_2^h} f(x)^p\right)^{1/p}.$$

For every function $f : \mathbb{F}_2^h \to \mathbb{R}$ with Fourier expansion

$$f(x) = \sum_{a\in\mathbb{F}_2^h} \hat{f}_a\chi_a(x)$$

let

$$f^{<d/4}(x) = \sum_{a:\mathrm{rank}_{\mathcal{B}^*}(a)<d/4} \hat{f}_a\chi_a(x),$$
$$f^{\geq d/4}(x) = \sum_{a:\mathrm{rank}_{\mathcal{B}^*}(a)\geq d/4} \hat{f}_a\chi_a(x).$$

Fix a set $\mathcal{S} \subseteq \mathbb{F}_2^h$. Let $\tau = \mu(\mathcal{S})$ be the volume of $\mathcal{S}$. Let $\mathbf{1}_\mathcal{S}(x) = \mathbf{1}_{x\in\mathcal{S}}$ be the indicator function $\mathcal{S}$. Note that $\|\mathbf{1}_\mathcal{S}\|_p^p = \tau$ for every $p \geq 1$. We will lower bound the expansion $\Phi(\mathcal{S}) = 1 - \langle\mathbf{1}_\mathcal{S}, G\mathbf{1}_\mathcal{S}\rangle/\tau$, which is the fraction of the edges incident to $\mathcal{S}$ leaving $\mathcal{S}$. Observe that

$$\langle\mathbf{1}_\mathcal{S}, G\mathbf{1}_\mathcal{S}\rangle = \langle\mathbf{1}_\mathcal{S}, G\mathbf{1}_\mathcal{S}^{<d/4}\rangle + \langle\mathbf{1}_\mathcal{S}, G\mathbf{1}_\mathcal{S}^{\geq d/4}\rangle. \tag{11}$$

The first term in the RHS of (11) is upper bounded as

$$\langle\mathbf{1}_\mathcal{S}, G\mathbf{1}_\mathcal{S}^{<d/4}\rangle = \|\mathbf{1}_\mathcal{S}\|_{4/3}\|G\mathbf{1}_\mathcal{S}^{<d/4}\|_4 \leq \|\mathbf{1}_\mathcal{S}\|_{4/3}\cdot\sqrt{3}^d\|\mathbf{1}_\mathcal{S}\|_2 = \sqrt{3}^d\tau^{5/4}$$

by Hölder's inequality and Claim 23. The second term in the RHS of (11) is upper bounded as

$$\langle\mathbf{1}_\mathcal{S}, G\mathbf{1}_\mathcal{S}^{\geq d/4}\rangle \leq (1-\nu d/4)\|\mathbf{1}_\mathcal{S}\|_2^2 = (1-\nu d/4)\tau.$$

In all, we have

$$\Phi(\mathcal{S}) = 1 - \frac{\langle\mathbf{1}_\mathcal{S}, G\mathbf{1}_\mathcal{S}\rangle}{\tau} \geq 1 - \sqrt{3}^d\tau^{1/4} - (1-\nu d/4) = \nu d/4 - \sqrt{3}^d\tau^{1/4}.$$

□

ECCC ISSN 1433-8092

http://eccc.hpi-web.de