CS 229r Essential Coding Theory April 13, 2020

Lecture 20

Instructor: Madhu Sudan *Scribe: Louis Golowich*

1 Outline

In this lecture we covered local decodability/correctability[1], as well as local testability. Local decodability refers to decoding the message at given indices, whereas local testability refers to testing if a given string is close to a codeword. These concepts are "local" because they only allow access to the input through a limited number of oracle queries. We present local decodability and local testability results for the Hadamard and Reed-Muller codes. These concepts have applications in PCPs.

2 Local Decodability

In this lecture we think of the space Σ^n as the space of functions $f : [n] \to \Sigma$, so that in particular codewords will be denoted by functions.

Definition 1. *A code $\mathcal{C} \subseteq \Sigma^n$ is (ℓ, ε)-locally correctible if there exists a decoder D such that for all*

- *$g \in \Sigma^n$ such that there exists some $f \in \mathcal{C}$ with $\delta(f, g) < \varepsilon$,*
- *$x \in [n]$,*

$D^g(x)$ makes ℓ queries into g and outputs $f(x)$ with probability greater than $\frac{1}{2}$.

The decoder $D^g(x)$ above takes two inputs: an explicit position x, as well as a random access oracle for g, which should be thought of as a corrupted codeword. The algorithm $D^g(x)$ must be randomized, as the corruption model in Definition 1 is adversarial, so an adversary could corrupt the specific ℓ positions queried by any deterministic algorithm. The definition allows for any success probability greater than $\frac{1}{2}$, as this probability can be amplified by running the algorithm multiple times in the standard way. Note that we will not place any restrictions on running times of algorithms in this lecture, so that complexity is determined just by the locality ℓ.

2.1 Hadamard Code

As a reminder, the Hadamard code $\mathcal{H}_n \subseteq \{f : \mathbb{F}_2^n \to \mathbb{F}_2\}$ is the set of all linear functions, that is,

$$\mathcal{H}_n = \left\{ f : \mathbb{F}_2^n \to \mathbb{F}_2 \;\middle|\; \exists \alpha_1, \ldots, \alpha_n \in \mathbb{F}_2 \text{ s.t. } f(x) = \sum_{i=1}^{n} \alpha_i x_i \; \forall x_1, \ldots, x_n \in \mathbb{F}_2^n \right\}.$$

Locally, $\mathcal{H}_n$ is characterized by the property that $f \in \mathcal{H}_n$ if and only if for all $x, y \in \mathbb{F}_2^n$, it holds that $f(x) + f(y) = f(x + y)$. Therefore $f(x)$ can be recovered from $f(y)$ and $f(x + y)$. To apply this idea, let's first reiterate the local decoding problem: we are given

- oracle access to $g : \mathbb{F}_2^n \to \mathbb{F}_2$ such that there exists some $f \in \mathcal{H}_n$ with $\delta(f, g) < \varepsilon$, and
- a position $x \in \mathbb{F}_2^n$.

[1] We treat the terms local decodability and local correctability as equivalent here, but in the literature there is a distinction, and the concept we discuss is local correctability.

We want to compute $f(x)$. Note that the inputs g and x can be provided by an adversary subject to the above conditions. The local decoder $D^g(x)$ will work as follows:

1. Pick $y \in \mathbb{F}_2^n$ at random.
2. Output $g(x+y) - g(y)$.

Claim 2. $\Pr[D^g(x) \neq f(x)] < 2\varepsilon$.

Proof. It follows by definition that $\Pr_y[g(y) \neq f(y)] < \varepsilon$. Similarly, it holds that $\Pr_y[g(x+y) \neq f(x+y)] < \varepsilon$ because $x+y$ is distributed uniformly over $\mathbb{F}_2^n$. The result follows from a union bound with these two inequalities, as $f(x+y) - f(y) = f(x)$ for any $f \in \mathcal{H}_n$. □

Thus we have shown:

Theorem 3. *$\mathcal{H}_n$ is $(2, \frac{1}{4})$-locally correctible.*

Note that $\frac{1}{4}$ in the theorem above is optimal, as $\mathcal{H}_n$ has distance $\frac{1}{2}$, and unique decoding is at best possible for error rates at half the distance.

2.2 Reed-Muller Codes

Recall that Reed-Muller codes are defined by

$$\mathrm{RM}(q, r, m) = \{f : \mathbb{F}_q^m \to \mathbb{F}_q | f \text{ polynomial, } \deg(f) \leq r\}.$$

In this lecture we assume $r < q$, so that no additional constraints on degrees are needed. We will show that Reed-Muller codes are locally decodable, and we will be happy to get locality bounded above by q; note that q is often significantly smaller than $|\mathbb{F}_q^m|$.

We take advantage of the local properties of Reed-Muller codes by restricting the domain $\mathbb{F}_q^m$ to lines. Specifically, for $a, b \in \mathbb{F}_q^m$, let $\ell_{a,b} = \{at + b | t \in \mathbb{F}_q\}$, so that the restriction of $f : \mathbb{F}_q^m \to \mathbb{F}_q$ to $\ell_{a,b}$ is given by $f|_{\ell_{a,b}}(t) = f(at+b)$. The restriction of an m-variate polynomial of degree r to a line is a 1-variate polynomial of degree r, or more formally,

$$f \in \mathrm{RM}(q, r, m) \implies f|_{\ell_{a,b}} \in \mathrm{RM}(q, r, 1).$$

This observation is useful because elements of $\mathrm{RM}(q, r, 1)$ are $(r+1)$-local: the values at $r+1$ locations determine the values everywhere.

As an aside, if $r < \frac{q}{2}$, then

$$f \in \mathrm{RM}(q, r, m) \impliedby f|_{\ell_{a,b}} \in \mathrm{RM}(q, r, 1)\ \forall a, b. \tag{1}$$

Exercise 4. *Prove the above statement using induction on m.*

Proof. It is sufficient to show that for every $f \in \mathrm{RM}(q, r, m)$, there exists some line $\ell_{a,b}$ such that $\deg f|_{\ell_{a,b}} = r$, as then the right hand side of (1) will not hold for any $r' < r$. This result holds for $m = 1$ using $a = 1$, $b = 0$. For $m = 2$, without loss of generality let $f(x, y) = x^{e_1} y^{e_2}(c_s x^s + c_{s-1} x^{s-1} y + \cdots + c_0 y^s) + g(x, y)$, where $e_1 + e_2 + s = r$ and $c_s \neq 0$, and $g(x, y)$ is a polynomial of degree at most $r - 1$. Set $a = (a_1, 1)$ and $b = (0, 0)$. Then $f|_{\ell_{a,b}} = (c_s a_1^s + c_{s-1} a_1^{s-1} + \cdots + c_0) a_1^{e_1} t^r + g(a_1 t, t)$, which has degree r as long as $(c_s a_1^s + c_{s-1} a_1^{s-1} + \cdots + c_0) a_1^{e_1} \neq 0$. But because $c_s \neq 0$, this polynomial has degree $s + e_1 \leq r < q$, so there exists some $a_1 \in \mathbb{F}_q$ that is not a root of the polynomial. Then $\deg f|_{\ell_{a,b}} = r$ for this choice of a_1. The proof can now proceed using induction on m, with $m = 1, 2$ as the base cases. For the inductive hypothesis, assume that for all $m' < m$, every r', and every $f' \in \mathrm{RM}(q, r', m')$, there exists some line $\ell_{a',b'}$ such that $\deg f'|_{\ell_{a',b'}} = r'$. For a given $f \in \mathrm{RM}(q, r, m)$, assume without loss of generality that x_1 lies in some degree-r term in f, so that f can be written in the form

$$f(x_1, \ldots, x_m) = x_1^{e_1}(c_s x_1^s g_s(x_2, \ldots, x_m) + c_{s-1} x_1^{s-1} g_{s-1}(x_2, \ldots, x_m) + \cdots + c_0 g_0(x_2, \ldots, x_m)) + h(x_1, \ldots, x_m),$$

where $e_1 + s \leq r$, $c_s \neq 0$, each g_i is a polynomial of degree $r - i - e_1$, and h is a polynomial of degree at most $r - 1$. For any line $\ell_{a,b}$ it holds that $\deg x_1^{e_1}|_{\ell_{a,b}} = e_1$ and $\deg h|_{\ell_{a,b}} \leq r - 1$, so in order for $\deg f|_{\ell_{a,b}} = r$, it is sufficient to choose a, b so that $\deg(c_s x_1^s g_s + \cdots + c_0 g_0)|_{\ell_{a,b}} = r - e_1$. By the inductive hypothesis, there exist choices of $a' = (a'_2, \ldots, a'_m)$, $b' = (b'_2, \ldots, b'_m)$ such that $\deg g_s|_{\ell_{a',b'}} = r - s - e_1$. Then along this line $\ell_{a',b'}$, the polynomial $c_s x_1^s g_s + \cdots + c_0 g_0$ restricts to a degree-$(r - e_1)$ polynomial in two variables x_1, t. Thus again applying the inductive hypothesis gives that there is some choice of a, b, such that $\deg(c_s x_1^s g_s + \cdots + c_0 g_0)|_{\ell_{a,b}} = r - e_1$, and thus $\deg f|_{\ell_{a,b}} = r$, completing the inductive step. □

Theorem 5. *For all m, $r < q - 1$, RM(q, r, m) is $(r + 1, O(\frac{1}{r+1}))$-locally correctible.*

Proof. Given $g : \mathbb{F}_q^m \to \mathbb{F}_q$ such that $\delta(f, g) < \varepsilon$ for some f, and given $x \in \mathbb{F}_q^m$, the local decoding algorithm proceeds as follows:

1. Pick a random $y \in \mathbb{F}_q^m$.
2. Find the unique polynomial $p \in \mathbb{F}_q^{<r+1}[X]$ such that $p(i) = g|_{\ell_{y,x}}(i)$ for all $i = 1, \ldots, r + 1$.
3. Output $p(0)$.

The algorithm above uses $r + 1$ calls to the oracle for g to compute $g|_{\ell_{y,x}}(i)$ for $1 \leq i \leq r + 1$. Because $x + iy$ is uniformly distributed over $\mathbb{F}_q^m$ for each i, it follows that $\Pr_y[f(x + iy) \neq g(x + iy)] < \varepsilon$. Applying a union bound over all i gives that $\Pr_y[\exists i : f(x + iy) \neq g(x + iy)] < (r + 1)\varepsilon$, which implies that $\Pr_y[p \neq f|_{\ell_{y,x}}] < (r + 1)\varepsilon$. Thus the decoding algorithm fails with probability over y less than $(r + 1)\varepsilon$, which is less than $\frac{1}{2}$ for $\varepsilon = O(\frac{1}{r+1})$. □

In fact a stronger result holds:

Theorem 6. *If $r = o(q)$, then RM(q, r, m) is $(O(r), \frac{1}{4} - o(1))$-locally correctible.*

Exercise 7. *Prove this result.*

Sketch of Proof Consider the case where $\frac{1}{4} - \varepsilon$ fraction errors are introduced for some $\varepsilon > 0$. Therefore for any codeword $f \in \mathrm{RM}(q, r, m)$, let $g : \mathbb{F}_q^m \to \mathbb{F}_q$ satisfy $\delta(f, g) < \frac{1}{4} - \varepsilon$. Consider the following local decoding algorithm, which takes an input parameter C, as well as $x \in \mathbb{F}_q^m$:

1. Pick a random $y \in \mathbb{F}_q^m$ and a random subset $S \subset \mathbb{F}_q$ of size C.
2. Find the polynomial $p \in \mathbb{F}_q^{<r+1}[X]$ that agrees with $g|_{\ell_{y,x}}$ for as many $t \in S$ as possible.
3. Output $p(0)$.

Because the lines $\ell_{y,x}$ for $y \in \mathbb{F}_q^m$ form a partition of $\mathbb{F}_q^m \setminus \{x\}$, it follows that with probability $\geq \frac{1}{2} + \frac{\varepsilon}{2}$, the choice of y is such that $g|_{\ell_{y,x}}$ agrees with $f|_{\ell_{y,x}}$ on $\geq \frac{1}{2} + \frac{\varepsilon}{2}$ fraction of points in $\ell_{y,x}$. Therefore as long as ε is chosen such that $\frac{\varepsilon}{2} q > r$, it follows that no polynomial $p \in \mathbb{F}_q^{<r+1}[X]$ such that $p \neq f|_{\ell_{y,x}}$ can agree with $g|_{\ell_{y,x}}$ at $\geq \frac{1}{2}$ fraction points in $\ell_{y,x}$, as any two distinct polynomials in $\mathbb{F}_q^{<r+1}[X]$ agree in at most r locations. Therefore for a sufficiently large choice of $\varepsilon = \varepsilon(q) = o(1)$ and of $C = C(r) = O(r)$, the probability that the decoding algorithm above chooses $p \neq f|_{\ell_{y,x}}$ is exponentially small in r, as having $p \neq f|_{\ell_{y,x}}$ would require the set S to contain a significantly smaller than expected fraction of points $t \in \mathbb{F}_q$ for which $f|_{\ell_{y,x}}(t) = g|_{\ell_{y,x}}(t)$. Thus the overall probability that $p = f|_{\ell_{y,x}}$ and therefore that $p(0) = f(x)$ is at least $\frac{1}{2} + \frac{\varepsilon}{2} - o(\varepsilon)$, which is greater than $\frac{1}{2}$, as desired. □

The $\frac{1}{4}$ in the above theorem is not theoretically optimal, as the alphabet can be large. However, it is at least a constant unlike in Theorem 5.

Note that both the Hadamard and Reed-Muller local decoding algorithms involve restricting to a low-dimensional subspace that preserves digree. For the Hadamard code the restriction was to a 2-dimensional linear subspace, while for Reed-Muller it was to a 1-dimensional affine subspace.

3 Local Testability

Definition 8. *A code $\mathcal{C} \subseteq \Sigma^n$ is (ℓ, α)-locally-testable if there exists a tester T such that:*

- *T^g accepts with probability 1 if $g \in \mathcal{C}$,*
- *T^g rejects with probability at least $\alpha\delta(g, \mathcal{C})$, where $\delta(g, \mathcal{C}) = \min_{f \in \mathcal{C}}\{\delta(f, g)\}$,*
- *T^g makes ℓ queries to the random acces oracle for g.*

Theorem 9. *$\mathcal{H}_n$ is $(3, \frac{1}{10})$-locally-testable, and $RM(q, r, m)$ is $(r + 2, \frac{1}{r^2})$-locally testable.*

For both the Hadamard and Reed-Muller codes, the testing algorithm is as follows:

- Pick x at random.
- Accept g if $g(x) = D^g(x)$.

Therefore in both cases, local testing uses one more query than local decoding.

Proof that $\mathcal{H}_n$ is $(3, \frac{1}{10})$-locally-testable. Fix $g : \mathbb{F}_2^n \to \mathbb{F}_2$, and let $\varepsilon(g) = \Pr_{x,y}[g(x) + g(y) \neq g(x+y)]$ be the probability of the test returning a rejection. By definition $\varepsilon(g) = 0$ if $g \in \mathcal{H}_n$. Therefore we need to show that $\varepsilon(g) \geq \frac{1}{10}\delta(g, \mathcal{H}_n)$. For $r \in \mathbb{F}_2^n$, let $D^g(x; r) = g(x + r) - g(r)$. Define $f(x)$ to be the mode of the set $\{D^g(x; r) | r \in \mathbb{F}_2^n\}$, with ties broken arbitrarily.

Lemma 10. $\delta(f, g) \leq 2\varepsilon(g)$.

Proof. If $f(x) \neq g(x)$, then it follows by the definition of f that $\Pr_y[g(x) \neq g(x+y) - g(y)] \geq \frac{1}{2}$. Therefore if there were more than $2\varepsilon(g)$ values of x for which $f(x) \neq g(x)$, then it would follow that $\varepsilon(g) = \Pr_{x,y}[g(x) \neq g(x + y) - g(y)] > 2\varepsilon(g) \cdot \frac{1}{2} = \varepsilon(g)$, a contradiction. □

The following lemma, which we will not prove, provides the key idea in the proof.

Lemma 11. *For all $x \in \mathbb{F}_2^n$,*

$$\Pr_{r_1, r_2}[D^g(x; r_1) \neq D^g(x; r_2)] \leq 2\varepsilon(g).$$

Corollary 12. *For all x,*

$$\Pr_r[f(x) \neq D^g(x; r)] \leq 2\varepsilon(g).$$

Exercise 13. *Prove this corollary.*

Proof. It is sufficient to show the following set-theoretic statement: for any multiset S of size N with mode r^*,

$$\Pr_{r \in S}[r \neq r^*] \leq \Pr_{r_1, r_2 \in S}[r_1 \neq r_2],$$

where r, r_1, r_2 are drawn uniformly from S. Let $m_1 > m_2 > \cdots > m_d$ be the multiplicies of the distinct elements of S, so that m_1 is the multiplicity of r^*. Then

$$\begin{aligned}
\Pr_{r \in S}[r \neq r^*] &= \frac{N - m_1}{N} \\
&= \sum_{i=1}^{d} \frac{m_i}{N} \cdot \frac{N - m_1}{N} \\
&\leq \sum_{i=1}^{d} \frac{m_i}{N} \cdot \frac{N - m_i}{N} \\
&= \Pr_{r_1, r_2 \in S}[r_1 \neq r_2].
\end{aligned}$$

Now the corollary follows letting $S = \{D^g(x; r) | r \in \mathbb{F}_2^n\}$ with $r^* = f(x)$. □

Lemma 14. *If $\varepsilon(g) < \frac{1}{10}$, then for all $x, y \in \mathbb{F}_2^n$ it holds that $f(x) + f(y) = f(x+y)$, so $f \in \mathcal{H}_n$.*

Now Lemmas 10 and 14 imply that if $\varepsilon(g) < \frac{1}{10}$ then $\varepsilon(g) \geq \frac{1}{2}\delta(g, f) \geq \frac{1}{2}\delta(g, \mathcal{H}_n)$, while if $\varepsilon(g) \geq \frac{1}{10}$ then $\varepsilon(g) \geq \frac{1}{10} \cdot 1 \geq \frac{1}{10}\delta(g, \mathcal{H}_n)$. Thus in either case $\varepsilon(g) \geq \frac{1}{10}\delta(g, \mathcal{H}_n)$, so $\mathcal{H}_n$ is $(3, \frac{1}{10})$-locally-testable. □

The statement regarding Reed-Muller codes in Theorem 9 has a similar but more complex proof.

4 Applying Reed-Muller Locally Testable and Locally Decodable Codes to Probabilistically Checkable Proofs (PCPs)

As an example, we define a PCP for graph 3-coloring:

Example 15 (Graph 3-Coloring)**.** 3-COL $\in$ *PCP*(ℓ) *if there exists a polynomial time verifier V accessing a proof $\pi \in \{0,1\}^*$ such that for all graphs G,*

- *if G is 3-colorable then there exists some $\pi \in \{0,1\}^{\text{poly}(|G|)}$ such that $V^\pi(G) = 1$ with probability 1,*
- *if G is not 3-colorable then for all π, it holds that $\Pr[V^\pi(G) = 1] \leq \frac{1}{2}$,*
- *V makes ℓ queries to the proof π.*

As an intermediate step, we consider a PCP for Reed-Solomon codes. The verifier V is now given oracle access to $g : \mathbb{F}_q \to \mathbb{F}_q$ as well as to a proof π, and must satisfy:

- There exists some π such that $\Pr[V^{g,\pi}(k) = 1] = 1$ if g is a polynomial with $\deg(g) \leq k$.
- It holds for all π that $\Pr[V^{g,\pi}(k) = 1] \leq \frac{1}{2}$ if $\delta(f, g) \geq .1$ for all f such that $\deg(f) \leq 2k$.
- The total number of queries to g and π is $\ell = \text{polylog}(|H|)$.