CS 6815 Pseudorandomness and Combinatorial Constructions **Fall 2018**

Lecture 8: September 20

Lecturer: Eshan Chattopadhyay *Scribe: Jesse Goodman*

In which we obtain r-wise independence from linear codes, obtain ϵ-balanced codes from ϵ-biased spaces, and explore Reed-Solomon and Reed-Muller (polynomial) codes.

8.1 Recap

We briefly recall a few definitions and observations from previous lectures:

- C is a **linear** $[n, k, d]_q$ **code** if C is a linear subspace of $\mathbb{F}_q^n$ with dimension $k := \log_q(|C|)$ and distance $d := \min_{x \neq y \in C} |\{i \in [n] : x_i \neq y_i\}| = \min_{x \in C} |\{i \in [n] : x_i \neq 0\}|$ (because it is linear).
- $G \in \mathbb{F}_q^{n \times k}$ is a **generator matrix of** C if $C = \{Gx : x \in \mathbb{F}_q^k\}$.
- $C^\perp := \{y \in \mathbb{F}_q^n : \forall c \in C, \langle c, y\rangle = 0\}$ is called the **dual code of** C. Basic linear algebra tells us that $C^\perp$ has dimension $n - k$.
- $H := (G^\perp)^T$ is the **parity check matrix of** C if $G^\perp \in \mathbb{F}_q^{n \times (n-k)}$ is the generator matrix of $C^\perp$. Using the definitions above, it is straightforward to show that $C = \{y \in \mathbb{F}_q^n : Hy = \mathbf{0}\}$. In other words, when $q = 2$, C contains exactly the strings that evaluate to 0 under every *parity function* induced by a row in H.

8.2 r-wise independence from linear codes

We show that generator matrices for certain linear codes can also generate r-wise independent distributions.

Claim 8.1 *Let C be an $[n, k, d]_2$ linear code such that $C^\perp$ is an $[n, n-k, r+1]_2$ linear code. Then, any r rows in the generator matrix G of C are linearly independent.*

Proof: Note that the parity check matrix of $C^\perp$ is G^T, and thus $C^\perp = \{y \in \mathbb{F}_2^n : G^T y = \mathbf{0}\}$. Let r^+ be the smallest number of rows in G that are linearly dependent. Then, there are r^+ columns in G^T, labeled $v_{i_1}, v_{i_2}, \ldots, v_{i_{r^+}}$, such that $\sum_{j \in [r^+]} v_{i_j} = \mathbf{0}$ (recall we are working over $\mathbb{F}_2$). Now, define a vector $y \in \mathbb{F}_2^n$ that equals 0 everywhere except at coordinates $i_1, i_2, \ldots, i_{r^+}$, where it equals 1. Then $G^T y = \sum_{j \in [r^+]} v_{i_j} = \mathbf{0}$, and thus $y \in C^\perp$. Because $C^\perp$ is a linear code with distance $r + 1$, every vector it contains must have Hamming weight at least $r + 1$. Because the Hamming weight of y, defined above, is r^+, we must have $r^+ \geq r + 1$. Thus, by definition of r^+, any set of r rows in G must be linearly independent. ∎

The corollary below follows from the observation that taking any r rows of G produces a matrix of rank r that therefore outputs a vector uniformly from $\mathbb{F}_2^r$ when applied to a vector sampled uniformly from $\mathbb{F}_2^k$.

Corollary 8.2 *The distribution $y = Gx$, where x is sampled uniformly from $\mathbb{F}_2^k$, is r-wise independent on $\mathbb{F}_2^n$.*

8.3 ϵ-balanced codes from ϵ-biased spaces

Recall that an **ϵ-biased space**, or **ϵ-biased distribution**, can be thought of as the output of a pseudorandom generator for the class of parity functions. In particular, it is a distribution D on $\{0,1\}^k$ such that for all nonempty $T \subseteq [k]$,

$$\frac{1}{2} - \epsilon \leq \Pr_{x \sim D}\big[\bigoplus_{i \in T} x_i = 1\big] \leq \frac{1}{2} + \epsilon.$$

It is equivalent to think of an ϵ-biased space as a uniform distribution D over subset $S \subseteq \{0,1\}^k$; that is, $\Pr[D = x] = 1/|S|$ for all $x \in S$, and 0 otherwise.

We say that an $[n,k,d]_2$ code C is **ϵ-balanced** if, for all nonzero $c \in C$,

$$(\frac{1}{2} - \epsilon) \cdot n \leq |c| \leq (\frac{1}{2} + \epsilon) \cdot n,$$

where $|\cdot|$ denotes the Hamming weight. Observe that if C is linear, then $d \geq (\frac{1}{2} - \epsilon) \cdot n$, because distance in a linear code is equal to the smallest Hamming weight of any vector within it. We now see that we can easily obtain an ϵ-balanced code from an ϵ-biased space.

Claim 8.3 *Let D be an ϵ-biased space on $\mathbb{F}_2^k$ that is supported (uniformly) on S. Let $n := |S|$, and denote the elements of S by $\{s_1, s_2, \ldots, s_n\}$. Define a matrix $G \in \mathbb{F}_2^{n \times k}$ such that row i of G is s_i. Then, $C := \{Gy : y \in \mathbb{F}_2^k\}$ is an ϵ-balanced code.*

Proof: This follows almost immediately from the definitions. Fix any nonzero $y \in \mathbb{F}_2^k$. Define $T := \{i \in [k] : y_i = 1\} \subseteq [k]$. (Notice T is not empty.) Then, element j of vector $Gy \in \mathbb{F}_2^n$ is simply $\langle s_j, y\rangle = \bigoplus_{i \in T} s_{j,i}$, where $s_{j,i}$ is the i^{th} coordinate of the j^{th} vector in S. Thus, letting $|\cdot|$ denote Hamming weight,

$$|Gy| = \#\big\{j \in [n] : \bigoplus_{i \in T} s_{j,i} = 1\big\} = n \cdot \Pr_{j \sim [n]}\big[\bigoplus_{i \in T} s_{j,i} = 1\big] = n \cdot \Pr_{x \sim D}\big[\bigoplus_{i \in T} x_i = 1\big],$$

which completes the proof, because we assumed that D is an ϵ-biased space. ■

8.4 Polynomial codes

8.4.1 Reed-Solomon codes

A **Reed-Solomon code** is constructed as follows: consider a field $\mathbb{F}_q$, any subset $S \subseteq \mathbb{F}_q$ with n distinct elements $\alpha_1, \alpha_2, \ldots, \alpha_n$ (typically, $S = \mathbb{F}_q$ or $S = \mathbb{F}_q \setminus \{0\}$ is used), and some $k < n$. Then, the code is defined as

$$C := \{(p(\alpha_i))_{i \in [n]} : p \in \mathbb{F}_q[x], \deg(p) \leq k-1\}.$$

To encode a message into C, the following protocol is used: given a message $m = (m_0, m_1, \ldots, m_{k-1}) \in \mathbb{F}_q^k$, define a corresponding polynomial $p_m \in \mathbb{F}_q[x]$ as $\sum_{i=0}^{k-1} m_i x^i$. Clearly it has degree $\leq k-1$, so $(p_m(\alpha_i))_{i \in [n]}$ is a codeword. Using this encoding scheme, we see that the generator for this code is, in fact, the Vandermonde matrix:

$$\begin{pmatrix} 1 & \alpha_1 & \alpha_1^2 & \ldots & \alpha_1^{k-1} \\ 1 & \alpha_2 & \alpha_2^2 & \ldots & \alpha_2^{k-1} \\ \vdots & & & \ddots & \vdots \\ 1 & \alpha_n & \alpha_n^2 & \ldots & \alpha_n^{k-1} \end{pmatrix}$$

Next, we relate this code to the types of codes we've already seen.

Claim 8.4 *C is an $[n, k, n-k+1]_q$ linear code.*

Proof: To see that C is linear, simply observe that it is closed under linear combinations: for polynomials p, q of degree $\leq k-1$ and scalars $\beta, \gamma \in \mathbb{F}_q$, $\beta p + \gamma q$ clearly has degree $\leq k-1$. Because C is over $\mathbb{F}_q$, the alphabet size of C is q. Because each codeword C has n coordinates, its block length is n. C has dimension k because there are q^k polynomials of the specified type (i.e., think of the correspondence between polynomials and the coefficients that can be attached to each power of x). To see that C has distance $n-k+1$, observe that because C is a linear code, it suffices to show that this is a lower bound for the minimum Hamming weight of any nonzero codeword. So, consider any message m that is encoded as a nonzero polynomial p_m. Because $\deg(p_m) \leq k-1$ (by our encoding protocol), the Fundamental Theorem of Algebra tells us that p_m has at most $k-1$ roots, and thus at least $n-k+1$ elements of $(p_m(\alpha_i))_{i \in [n]}$ are nonzero. Note also that this distance is tight, because there exist degree $k-1$ polynomials that evaluate to 0 on $k-1$ points. For example, $\prod_{i \in [k-1]}(x - \alpha_i)$. ■

8.4.2 Reed-Muller codes

Reed-Muller codes strictly generalize Reed-Solomon codes. To construct a Reed-Muller code, fix some field $\mathbb{F}_q$, along with numbers m and r (which will correspond to *number of variables* and *bound on total degree*, see below). A Reed-Muller code over these parameters is defined as:

$$C := \{(p(y))_{y \in \mathbb{F}_q^m} : p \in \mathbb{F}_q[x_1, x_2, \ldots, x_m], \deg(p) \leq r\},$$

where $\deg(p)$ denotes the total degree of p. Observe that each polynomial p over which this code is defined may be represented as the sum $\sum_T c_T x^T$, where $T = (t_1, \ldots, t_m)$ is a string of powers that sum to at most r, c_T is some coefficient from $\mathbb{F}_q$, and the notation x^T denotes $x_1^{t_1} x_2^{t_2} \cdots x_m^{t_m}$.

Next, we relate Reed-Muller codes on $\mathbb{F}_2$ to the types of code that we've already seen.

Claim 8.5 *The Reed-Muller code $\mathsf{RM}(m, r)$ on $\mathbb{F}_2$ is a $[2^m, \binom{m}{\leq r}, 2^{m-r}]_2$ linear code.*

Proof: $\mathsf{RM}(m, r)$ is a linear code because linear combinations of polynomials preserve degree. The block length of this code is clearly 2^m, from the definitions above. The dimension of $\mathsf{RM}(m, r)$ over $\mathbb{F}_2$ is $\binom{m}{\leq r} := \sum_{i=0}^{r} \binom{m}{i}$, and can be seen by counting the number of (multilinear) monomials on m variables with degree at most r. Now, because $\mathsf{RM}(m, r)$ is linear, to see that the distance is 2^{m-r}, we just need to show that all nonzero vectors in the code have hamming weight at least 2^{m-r}; i.e., that for all $p \in \mathbb{F}_2[x_1, \ldots, x_m]$ of total degree at most r, $|\{x \in \mathbb{F}_2^m : p(x) \neq 0\}| \geq 2^{m-r}$. To see this, observe that we can write every nonzero multilinear polynomial $p(x_1, \ldots, x_m)$ of max total degree r as $x_{i_1} x_{i_2} \cdots x_{i_l} + q(x_1, \ldots, x_m)$, where $l \leq r$, each $i_j \in [m]$, and $q(x_1, \ldots, x_m)$ is a multilinear polynomial of max total degree $\leq l$. Now, notice that for any $\{0, 1\}$ assignment to each variable $x_j, j \notin \{i_1, \ldots, i_l\}$, polynomial q turns into a polynomial of max degree $\leq r-1$, and thus p becomes a nonzero multilinear polynomial p' over variables $x_{i_1}, x_{i_2}, \ldots, x_{i_l}$. Notice that there is always some assignment to these variables such that p' evaluates to 1: simply take the lowest degree monomial in p', set its variables to 1, and set all other variables in p' to 0.

Since we may achieve this result for any $\{0, 1\}$ assignment to the variables $\{x_j\}_{j \notin \{i_1, \ldots, i_l\}}$, we know that $|\{x \in \mathbb{F}_2^m : p(x) \neq 0\}| \geq 2^{m-l} \geq 2^{m-r}$, as desired. Note that this result is tight, considering the polynomial $p = x_1 x_2 \cdots x_r$. ■