Today

- Algebraic codes
- Reed-Solomon Codes
- Reed-Muller Codes
- Hadamard Codes as a special case
- The Plotkin Bound

The story so far

- Hamming defines codes.
- Shannon's results: Motivate need for asymptotically good codes (codes with constant relative minimum distance, constant rate and constant alphabet).
- Have only two constructions:
  - Hamming codes: Good Rate but small distance.
  - Random codes: Asymptotically good, but non-constructive.

What next

- Exploit algebra.
- Use it to obtain a family of codes over large alphabet. (Reed-Solomon)
- Will try to reduce alphabet size algebraically. (Reed-Muller).
- Get binary codes - Hadamard codes.
- Plotkin Bound.

Reed-Solomon Codes

- Discovered in the context of coding theory by Reed and Solomon in 1960.
- Discovered earlier in the context of block designs by Bush. (Hmph!)
- Extremely simple codes + analysis.
- But can be easily obscured! (See any text on coding theory!)

Definition

- RS codes specified by:
  - Field F_q.
  - Parameters n, k.
  - Vector $\mathbf{a} = \langle \alpha_1, \ldots, \alpha_n \rangle$ of distinct elements in $\mathbb{F}_q$. (Need $n \leq q$.)
- Encoding as follows:
  - Associate message $\mathbf{m} = \langle m_0, \ldots, m_{k-1} \rangle$ with polynomial $p(x) = m_0 + m_1 x + \cdots + m_{k-1} x^{k-1}$ of degree less than k.
  - Encoding: $p \mapsto \langle p(\alpha_1), \ldots, p(\alpha_n) \rangle$.
- Parameters: $[n, k, n-k+1]_q$ code for $k \leq n \leq q$. Distance follows from: "Non-zero degree $k-1$ polynomial has at most k roots". (Hold over all fields? When else?)

The large alphabet issue

- Why is it reasonable to have large alphabets?
- In practice: CDs/DVDs think of single byte as a single symbol. Why is the Hamming metric right?
- Error often bursty! When single bit of byte is corrupted all nearby symbols also unreliable. So might as well treat them together!
- Even if we don't - RS codes are interesting.
- Let $q = n$ and write element of $\mathbb{F}_q$ as $\log n$ bit string.

- RS code becomes a $[n \log n, k \log n, n-k+1]_2$ code.
- Example: $k = n - 4$, then get approx. $[N, N - 4\log N, 5]_2$ code.
- Hamming/Volume bound: Distance 5 code must have $k \leq N - 2\log N$.
- So our defect is at most factor of two worse than best possible.

Reducing alphabet size: Bivariate polynomials

- Bottleneck in increasing length of code: Need more distinct elements!
- Way around - use more variables.
- Example:
  - Think of message as $\mathbf{m} = \langle m_{ij} \rangle_{i,j<\sqrt{k}}$ as matrix.
  - Associate bivariate polynomial $p(x, y)$ of degree at most $\sqrt{k}$.
  - Evaluate at all points in $S \times S$ where $S \subseteq \mathbb{F}_q$.
  - Using $S = \mathbb{F}_q$ gives $n = q^2$. Longer!
- Distance = ?

Schwartz-Zippel Lemma

Theorem: m-variate polynomial of total degree d is zero on at most $d/|S|$ fraction of the inputs in S^m.

- Will choose $x_1, \ldots, x_m$ at random from S^m and argue that random choice gives zero value with probability at most $d/|S|$.
- Perform induction on m. Base case $m = 1$ clear.
- Write polynomial $p(x_1, \ldots, x_m)$ as $p_1(x_1, \ldots, x_{m-1})x_m^{d_m} +$ lesser degree terms in x_m.
- Pick $a_1, \ldots, a_{m-1}$ at random from S^{m-1}.

- Prob. $p_1(a_1, \ldots, a_{m-1}) = 0$ at most $(d - d_m)/|S|$ by induction.
- Assume above doesn't happen. Let $g(x_m) = p(a_1, \ldots, a_{m-1}, x_m)$. g is a non-zero polynomial of degree d_m. Choice $x_m = a_m$ makes it zero w.p. at most $d_m/|S|$. Else $p(a_1, \ldots, a_m) \neq 0$.
- Union bound: Prob. of being zero at most $d/|S|$.

Schwartz-Zippel Lemma (contd.)

Some myths about the Lemma:

- That it is a Lemma: Actually a theorem.
- That it is due to Schwartz+Zippel: Actually used many times in algebra/algebraic geometry/coding theory before.
- That its discovery in theoretical computer science is due to Schwartz/Zippel alone: Also discovered by DeMillo+Lipton independently!
- Still nice to have a named object and we will perpetuate the myth.

Back to bivariate polynomials

- Bivariate polynomials give $[n, k, d]$ code for $d \geq n - k - (\sqrt{k}(2q - \sqrt{k})$.
- Why this strange way of writing it? Need to see how much worse than $n - k$ it gets.
- Can improve bound to $d \geq n-k-(\sqrt{k}(2q-2\sqrt{k})$ by paying more attention.
- So certainly not as good as RS codes. But do have significantly longer code.

m-variate polynomials

- $n = q^m$, $k = \binom{m+\ell}{m}$ if degree of polynomial ℓ. $d = (1 - \ell/q) \cdot n$.
- Codes called Reed-Muller codes.
- Asymptotically good?
  - Can't be. Need $m = \log_q n$ variables and constant degree $\ell < q$.
  - $k = \binom{m+\ell}{m}$ grows as m^ℓ - polynomial in m, while $n = q^m$ grows exponentially in m.
- Coding theorists try $\ell > q$, but with individual degree per variable at most $q-1$. Gives interesting range of parameters (see exercise), but not asymptotically good.

A special case: Hadamard codes

- Let $q = 2$ and $\ell = 1$. Gives $[2^\ell, \ell+1, 2^{\ell-1}]_2$ code.
- Variants ...
  - $[n, \log_2 n, n/2]$ - equidistant code.
  - $[2n, \log_2 n, n/2]$ - code using all rows and complements.
  - $[n-1, \log_2 n, n/2]$ - code by assuming w.l.o.g. first column is all 1's and deleting this column.
- First is weaker than second and third, but has additional property. Second is what we get from polynomials. Third is the dual of the Hamming code. All essentially same

from our perspective. Give similar flavor of results.

Plotkin Bound

- Given any $(n, k, n/2)_2]$ code, $k \leq 1 + \log_2 n$.
- Projection technique: If an $(n, k, d)_q$ code exists, then so does an $(n-r, k-r, d)_q$ code.
- Putting them together: $k \leq 1 + \log_2 n + n - 2d$. Asymptotically, $R + 2\delta \leq 1$ for binary codes.