CAS CS 599 B: Mathematical Methods for TCS

Lecturer: Mark Bun | Spring 2022

Lecture Notes 20:

Existential Bounds, Reed-Solomon Codes

Reading.

- Guruswami-Rudra-Sudan §4, 5, 10

Here's a reminder of some definitions:

- An $(n, k, d)_q$ code is a subset $C \subseteq \Sigma^n$ where $|\Sigma| = q$. Here, n is the block length, $k = \log_q |C|$ is the dimension, and $d = \min_{v,w \in C} \Delta(v, w)$ is the distance of the code.
- The rate of a code is $R = k/n$. The relative distance is $\delta = d/n$.
- An $[n, k, d]_q$ linear code is one for which C is a k-dimensional subspace of $\Sigma^n = \mathbb{F}_q^n$.

The $[2^r - 1, 2^r - r - 1, 3]_2$-Hamming and $[2^r, r, 2^{r-1}]_2$-Hadamard codes achieve rates and distances at the extreme ends of a spectrum. Hamming gets excellent rate $\approx (n - \log n)/n$, but terrible relative distance $3/n$. Meanwhile, Hadamard gets terrible rate $\log n/n$ but excellent relative distance $1/2$. What kinds of tradeoffs between rate and relative distance are achievable in general?

We're often interested in the asymptotic behavior of families of codes, one for every block length. Can we construct families of codes $\{C_n\}$, where $C_n \subseteq \Sigma^n$, such that both

$$R := \liminf_{n\to\infty} \frac{k_n}{n}, \qquad \delta := \liminf_{n\to\infty} \frac{d_n}{n}$$

are both constant? Such codes are called "asymptotically good." What kinds of tradeoffs, i.e., pairs $(R, \delta) \in [0, 1] \times [0, 1]$ are achievable?

1 Hamming Bound

An impossibility result (rate upper bound) follows from the packing interpretation of codes. If a code C has distance d, then we can construct disjoint Hamming balls of radius $\lfloor (d - 1)/2 \rfloor$ around the points in C. Each such ball contains

$$\mathrm{Vol}_2(\lfloor (d-1)/2 \rfloor, n) = \sum_{i=0}^{\lfloor (d-1)/2 \rfloor} \binom{n}{i}$$

points in $\{0, 1\}^n$. On the other hand, there is only a total of 2^n points in then entire space $\{0, 1\}^n$. So we conclude

Theorem 1 (Binary Hamming bound). *Let $C \subseteq \{0, 1\}^n$ be a code with distance d. Then*

$$|C| \le \frac{2^n}{\mathrm{Vol}_2(\lfloor (d-1)/2 \rfloor, n)}.$$

To interpret this bound, let's estimate $\mathrm{Vol}_2(\lfloor (d-1)/2 \rfloor, n)$. I think the most enlightening thing to do is to estimate it from above, even though this isn't technically useful for proving a rate upper bound. Let $p \leq 1/2$. Then

$$\begin{aligned}
\mathrm{Vol}_2(pn, n) &= \sum_{i=0}^{pn} \binom{n}{i} \\
&\leq \frac{1}{p^{pn}(1-p)^{n-pn}} \sum_{i=0}^{pn} \binom{n}{i} p^i (1-p)^{n-i} \\
&\leq 2^{H_2(p)n} \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} \\
&= 2^{H_2(p)n}
\end{aligned}$$

where $H_2(p) := -p \log_2(p) - (1-p)\log_2(1-p)$ is the binary entropy function, which captures the Shannon entropy of a single bit that comes up 1 with probability p. Using Stirling's approximation (see Proposition 3.3.3 in GRS), one can show that this upper bound is basically tight: $\mathrm{Vol}_2(pn, n) \geq 2^{H_2(p)n - o(n)}$. Plugging this into Theorem 1 gives

$$|C| \leq 2^{n - H_2(\delta/2)n + o(n)} \implies k \leq n(1 - H_2(\delta/2)) + o(n).$$

Thus, asymptotically, the best rate of code with relative distance δ is at most $1 - H_2(\delta/2)$. These argument all work over larger alphabets as well, giving an asymptotic rate upper bound of $1 - H_q(\delta/2)$ for q-ary codes.

2 Gilbert-Varshamov Bound

Let us now see what rate vs. distance tradeoffs are actually achievable by codes. The following positive result is known as the GV bound:

Theorem 2 (Gilbert-Varshamov Bound)**.** *For $\delta < 1/2$, there exists a code family with relative distance δ and rate $R = 1 - H_2(\delta)$. Moreover, a random linear code achieves rate $R = 1 - H_2(\delta) - \varepsilon$ with probability at least $1 - 2^{-\varepsilon n}$.*

Proof of Part 1. We construct a code of distance d greedily as follows:

- Initialize $C = \emptyset$
- While there exists $w \in \{0,1\}^n$ such that $\Delta(w, v) \geq d$ for every $v \in C$, add w to C.

By construction, this code has distance d. Moreover, the balls of radius $d-1$ around codewords must cover the entire space:

$$\bigcup_{v \in C} B(v, d-1) = \{0,1\}^n.$$

This implies

$$|C| \geq \frac{2^n}{\mathrm{Vol}_2(d-1, n)} \geq 2^{n - H_2((d-1)/n)n}.$$

Taking logs shows that the rate of this code is at least $1 - H_2((d-1)/n) \geq 1 - H_2(\delta)$. □

Proof of Part 2. Let $k = (1 - H_2(\delta) - \varepsilon)n$ be our target dimension. Choose a random linear code by taking its generator matrix $G \in \mathbb{F}_2^{k \times n}$ to have independent, uniform entries. We need to show that with high probability, G has full rank k and induces a code with minimum distance at least $d = \delta n$. Both properties follow as long as for every nonzero message $x \in \mathbb{F}_2^k$, we have $wt(xG) \geq d$, where wt denotes Hamming weight.

To see this, we use the fact that for a random G, the vector xG is uniformly random to bound for each individual message x:

$$\Pr_G[wt(xG) < d] \leq \frac{\mathrm{Vol}_2(d-1, n)}{2^n} \leq 2^{H_2(\delta)n - n}.$$

Union bounding over all 2^k messages gives

$$\Pr_G[\exists x \neq \mathbf{0} : wt(xG) < d] \leq 2^{k - H_2(\delta)n - n} \leq 2^{-\varepsilon n}.$$

□

3 Reed-Solomon Codes

Definition 3. Let $k \leq n \leq q$ and take $\Sigma = \mathbb{F}_q$. Fix a set $A = \{\alpha_1, \ldots, \alpha_n\} \subseteq \mathbb{F}_q$ of distinct "evaluation points." Define the Reed-Solomon code $RS_A[n,k]_q$ by its encoding map:

$$\mathrm{Enc}(x_0, \ldots, x_{k-1}) = (p_x(\alpha_1), \ldots, p_x(\alpha_n))$$

where

$$p_x(\alpha) = x_0 + x_1\alpha + x_2\alpha^2 + \cdots + x_{k-1}\alpha^{k-1}.$$

The Reed-Solomon code is a linear code of dimension k. Its generator matrix is

$$G_A = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ \alpha_1 & \alpha_2 & \ldots & \alpha_n \\ \vdots & & & \\ \alpha_1^{k-1} & \alpha_2^{k-1} & \ldots & \alpha_n^{k-1} \end{pmatrix}.$$

Claim 4. *The Reed-Solomon code $RS_A[n,k]_q$ has distance $d = n - k + 1$.*

Proof. Let $(p_x(\alpha_1), \ldots, p_x(\alpha_n))$ be a non-zero codeword. Since p_x is a non-zero polynomial of degree k, it must have at most $k - 1$ zeroes. Therefore, the Hamming weight of this codeword is at least $n - k + 1$. □

Reed-Solomon codes actually match a simple impossibility result called the *Singleton bound*: Every code of distance d and dimension k has $d \leq n - k + 1$. (The proof of this is by the pigeonhole principle. Given a code of distance d, delete the first $d-1$ symbols from each codeword. This gives $|C|$ distinct strings in Σ^{n-d+1}, so $q^k \leq |C| \leq q^{n-d+1}$.)

The downside is that the alphabet is very large – larger than the block length. This turns out to be inherent, as there are strictly stronger impossibility results than the Singleton bound (e.g., the "Plotkin bound") that hold over small alphabets. So it's a natural question to ask whether one can trade the large alphabet size of the explicit Reed-Solomon code for a slightly worse rate vs. distance tradeoff.

A simple way to turn the Reed-Solomon code into a binary code is to just further encode each symbol $p_x(\alpha_i)$ as a binary string of length $\log q$. This produces a code with block length $N = n \log q$, but the same dimension and distance. So the rate and relative distance of the new code are now both at most $1/\log q$.

4 Concatenated Codes

The idea to get an improved binary code is to perform each encoding of Reed-Solomon symbol $p_x(\alpha_i)$ using another error-correcting code. The upshot is that the number of codewords we need for the second step is small, only q, so we can afford to do things like brute force search for such codes.

A concatenated code combines an outer code $C_{out} \subseteq \Sigma^N$ with an inner code $C_{in} \subseteq \sigma^n$ where $|C_{in}| \geq |\Sigma|$. The result is a concatenated code $C = C_{out} \circ C_{in}$, for which encoding is performed via

$$\mathrm{Enc}(x) = (\mathrm{Enc}_{in}((\mathrm{Enc}_{out}(x))_1), \ldots, \mathrm{Enc}_{in}((\mathrm{Enc}_{out}(x))_N))$$

If C_{out} is an $(N, K, D)_{q^k}$-code and C_{in} is an $(n, k, d)_q$-code, then the concatenated code C is an (Nn, Kk, Dd)-code. So the rate and relative distance of the concatenated code are the products of the rates and relative distances of the constituent codes.

4.1 Zyablov Bound

Suppose we take

- The outer code C_{out} is a Reed-Solomon code with rate R_{out} and relative distance $\delta_{out} = 1 - R_{out}$.
- The inner code C_{in} meets the Gilbert-Varshamov bound. So it has rate $R_{in} = 1 - H_2(\delta_{in})$ for relative distance δ_{in}.

One can optimize over the choice of R_{in} to get the Zyablov bound:

$$R \geq \max_{0 \leq r \leq 1 - H_2(\delta)} r \cdot \left(1 - \frac{\delta}{H_2^{-1}(1-r)}\right).$$

Codes asymptotically matching the Zyablov bound can be found in polynomial time. For example, if C_{out} is an $[N = Q - 1, K]_Q$ Reed-Solomon code, then we need the inner code to have dimension $\log Q = O(\log N)$. An explicit linear code with this dimension matching the Gilbert-Varshamov bound can be found in $\mathrm{poly}(N)$ time, e.g., by greedily constructing the parity check matrix.