CS294-226: Advances in Error-Correcting Codes **UC Berkeley, Fall 2022**
Lecture #1: Asymptotically Good Codes August 31, 2022
Lecturer: Venkatesan Guruswami Scribe: Chang-Han Chen

Today's Outline:

1. Asymptotically good codes
2. Gilbert-Varshamov bound
3. ~~Read-Solomon Codes~~

Last time:

1. Linear codes
2. Parity Check/ Generator matrix
3. Dual Code
4. Hamming & Simplex code

1 Reminder

Recall that the Hamming code is a $[2^r - 1, 2^r - 1 - r, 3]$ code[1]. It's associated parity-check matrix H is an $r \times (2^r - 1)$ matrix, where the ith column corresponds to the binary representation of i. For example, when $r = 3$, the parity-check matrix H will look as follows:

$$\begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}$$

Furthermore, we have also seen that the Simplex code is a $[2^r - 1, r, 2^{r-1}]$ and is the dual of the Hamming code.

Both the Hamming code and the Simplex code are two extreme ends of optimal codes: the Hamming code has optimal size among all binary codes of distance 3 ($|C| \leq \frac{2^n}{n+1}$), while the Simplex code has optimal size among all codes of distance $\frac{n}{2}$.

Lemma 1. *Given any code $C \subseteq \{0,1\}^n$ of distance d such that $d > \frac{n}{2}$, we have*

$$|C| \leq \frac{2d}{2d - n} \leq 2d$$

Notice that the Simplex code saturates the bound: $|C| = 2^n, d = 2^{r-1}, n = 2^r - 1$, then $\frac{2d}{2d-n} = 2^r$, optimal.

[1] We say that a code C is an $[n, k, d]$ code if it has blocklength n, dimension $\dim(C) = k$, and distance $\Delta(C) = d$.

Proof. There are two components in this proof. First, let's map a codeword $c = (c_1, ..., c_n)$ to $v_c := \frac{1}{\sqrt{n}}((-1)^{c_1}, \cdots, (-1)^{c_n})$. Observe that

$$v_{c_1} \cdot v_{c_2} = 1 - \frac{2\Delta(c_1, c_2)}{n}. \tag{1.1}$$

This is because whenever c_1 and c_2 differ at one site, the bit-wise product of $(-1)^{c_1}$ and $(-1)^{c_2}$ on that site is -1 instead of 1, *i.e.,* the penalty of each differing site is -2.

Second, it is a geometric fact that if $v_1, .., v_m \in \mathbb{R}^n$ such that $|v_i|| = 1$ and $v_i \cdot v_j \leq -\alpha$ for all $1 \leq i < j \leq m$ where $\alpha > 0$ is some positive constant, then [2]

$$m \leq \frac{1}{\alpha} + 1.$$

This fact follows by inspecting $0 \leq ||\sum v_i||^2$.

Finally, we can plug Equation (1.1) in place of α above. That completes the proof. □

As a side note, when $\alpha = 0$, the second fact can be modified.

Fact 1. *If $v_i \cdot v_j \leq 0$ for all $1 \leq i < j \leq m$, then $m \leq 2n$.*

For non-binary codes, Lemma 1 can be generalized analogously:

Lemma 2. *Given any positive integer $q \geq 2$ and code $C \subseteq \{0, 1, \ldots, q-1\}^n$ of distance d such that $d > (1 - \frac{1}{q})n$, we have*

$$|C| \leq \frac{d}{d - (\frac{q-1}{q})n}.$$

2 Asymptotically good codes

Goal: Given a family of codes over fixed alphabet $[q]$: $\{C_1, \cdots, C_n\}$ such that the block length n_i of each C_i goes to infinity as $i \to \infty$ and

$$R(C_i) \geq R_0 \tag{1.2}$$
$$\delta(C_i) \geq \delta_0 \tag{1.3}$$

for some non-zero $R_0, \delta_0 > 0$.

We want to understand the trade-off between R and δ.

Theorem 1.1. *For all $q, n, d \leq n$, there exists $C \subseteq [q]^n$ with $d(C) \geq d$ and*

$$|C| \geq \frac{q^n}{\sum_{i=0}^{d-1} \binom{n}{i}(q-1)^i}.$$

Proof. Greedily pick codewords such that each codeword is not within distance $(d-1)$ of any previously chosen one. Stop when running out of room. Thus,

$$|C| \cdot |B_q(0, d-1)| \geq q^n.$$

□

[2] Note that the RHS is independent of n.

Of course, this is a crude estimate. One can refine the prefactor of the bound by doing a more careful combinatorial analysis. Nonetheless, over binary alphabets, this is asymptotically the best known existential rate-distance tradeoff.

As an exercise, let's try to modify this argument to produce a linear code (over $\mathbb{F}_q$). For more details, see 4.2 of the textbook. The key idea is to fill in the columns of a parity check matrix greedily, such that no $i \leq d-1$ columns are linearly dependent, and thereby by a similar argument as above, we have

$$\sum_{i=0}^{d-2} \binom{n-1}{i} (q-1)^i < q^m,$$

where m is the rank of the parity-check matrix. Recall that $m = n - k$. Hence, there exist codes of dimension larger than or equal to $n - \lceil \log_q |B_q(0, d-2)| \rceil$. If q^m is very large, then a random parity-check matrix would work: this is the idea of a "Random Linear Code," whose parameters $[n, k, d]_q$ satisfy

$$q^k << \frac{q^n}{\sum_{i=0}^{d-1} \binom{n}{i}(q-1)^i}.$$

See 4.2.2 of the textbook for the proof.

3 Asymptotic form of GV bound

For binary, fix $\delta \in [0, 1/2)$, then as $n \to \infty$,

$$\frac{2^{h(\delta)n}}{\text{poly}(n)} \leq \binom{n}{\delta n} \leq 2^{h(\delta)n}, \tag{1.4}$$

where $h(x)$ is the binary entropy.

With the following lemma,

Lemma 3. *Suppose, $\delta \in [0, 1/2)$, then*

$$\sum_{i=0}^{\delta n} \binom{n}{i} \leq 2^{h(\delta)n}. \tag{1.5}$$

Therefore, from the GV bound,

$$|C| \geq 2^{(1-h(\delta))n}. \tag{1.6}$$

Finally, the asymptotic form of the GV bound becomes:

Theorem 1.2. *Fix $\delta \in [0, 1/2)$, there exists a binary linear code C with relative distance $\delta(C) = \delta$ and the rate*

$$R(C) \geq R_{GV}(\delta) = 1 - h(\delta). \tag{1.7}$$

Comments :

1. The GV bound approaches the Singleton bound very slowly: something like $h_q(\delta) \leq \delta - \frac{1}{\log q}$ as far as Venkat can remember.

2. For the binary codes, we don't know how to beat the GV bound in full generality. It is known for $q \geq 49$ though: "Algebraic Geometry Code," where the rate is

$$R \geq 1 - \delta - \frac{1}{\sqrt{q} - 1}. \tag{1.8}$$

3. The asymptotically good codes guaranteed by the GV bound is not efficiently constructible:
   (a) The greedy algorithm is exponentially costly.
   (b) A Random Linear Code is efficiently samplable, but cannot be certified in poly-time; in other words, there exists NO known decoding algorithm. This leads to the hardness assumption in cryptography, "Learning Parity with Noise".

4. When $\delta \to 0$, $R_{GV}(\delta) = 1 - O(\delta \log \frac{1}{\delta})$. When $\delta = 1/2 - \epsilon$, $R_{GV} = \Theta(\epsilon^2)$.

5. Nonetheless, there are some nearly-optimal constructible good codes:
   (a) As $\delta \to 0$, there are explicit codes of rate

$$R_{exp}(\delta) \geq 1 - O(\delta \operatorname{poly}(\log \frac{1}{\delta})).$$

   Examples include
      i. "Expander Codes" [2015ish]
      ii. Deletion correction codes: $R(\delta) \geq 1 - O(\delta \log^2 \frac{1}{\delta})$ [2018]
   (b) As $\delta \to \frac{1}{2} - \epsilon$, there are explicit codes of rate

$$R_{exp}(\delta) \geq \Omega(\epsilon^{2+o(1)}).$$

   Examples include
      i. "Expander walks" [**This is the same as the one below, I think.** –omar]
      ii. "Pseudorandomness" [Ta-17]

References

[Ta-17] Amnon Ta-Shma. Explicit, almost optimal, epsilon-balanced codes. In Hamed Hatami, Pierre McKenzie, and Valerie King, editors, *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Montreal, QC, Canada, June 19-23, 2017*, pages 238–251. ACM, 2017. 5(b)ii