A Crash Course on Coding Theory

Madhu Sudan
MIT

Topic: Bounds on Codes

This lecture will focus on limitations on the performance of codes. I.e., Upper bounds on rate/distance, or Lower bounds on block length.

Singleton bound

Thm: \( n \geq k + d - 1 \)

- Note: Independent of \( q \).
- Codes meeting the Singleton bound are called MDS codes (Max. Dist. Seperable). (Only) example: Reed-Solomon codes.

Proof (of Thm):
- Pick (any) \( k - 1 \) coordinates and project code.
- Two codewords collide (PHP).
- Implies distance \( \leq n - k + 1 \).

Greismer bound

Thm: For linear codes, \( n \geq \sum_{i=0}^{k-1} \left\lceil \frac{d}{q^i} \right\rceil \).

In particular, \( n \geq \frac{q}{q-1}d + k - \log_q d \).

Note: Strictly improves Singleton bound.

Proof: (for binary case)

Let \( G = \begin{bmatrix} n-d & d \\ 00\ldots0 & 11\ldots1 \\ G' & G'' \end{bmatrix} \)

- Every row of \( G'' \) has \( \geq \left\lceil \frac{d}{q} \right\rceil \) zeroes.
- \( G' \) generates \( [n-d, k-1, \left\lceil \frac{d}{q} \right\rceil]_q \) code.
- Theorem follows.
Recall Hamming Balls

- \( V(n, r, q) \) = “volume” of \( B(\cdot, r) \) in \( \Sigma^n \).
- Let \( H_q(p) \) be \( q \)-ary entropy function.

\[
H_q(p) = p \log_q \left( \frac{q-1}{p} \right) + (1-p) \log_q \left( \frac{1}{1-p} \right)
\]
- Fact:

\[
V(n, pn, q) \approx q^{H_q(p)n}
\]

Packing (Hamming) Bound

Thm: \( k \leq \left( 1 - H_q \left( \frac{1}{2} \cdot \frac{d}{n} \right) \right) n \).

Proof: Consider balls of radius \( \frac{d-1}{2} \) around codewords.
- Balls don’t intersect.
- Thus: \( V(n, d/2, q)^k \leq q^n \)
- Using approximation, get theorem.

Note: Codes meeting the inequality in proof tightly are called Perfect codes. e.g. Hamming codes (and only few others).

Compare with random linear codes:
- (Letting \( \delta = d/n \) and \( R = k/n \))

\[
1 - H_q(\delta) \leq R \leq 1 - H_q \left( \frac{\delta}{2} \right).
\]

Intermission

- Have met Singleton, Griesmer and Hamming.
- Will soon meet Plotkin, Elias-Bassalygo, and Johnson.
- Will view MacWilliams and LP from afar.
- Why?

Comparing Bounds

- Obviously want the best bound for a given choice of parameters.
- Say fixed \( q, R = k/n \), what is the best distance \( \delta = d/n \)?
- But relationship is not yet known!
- Further known relationships involve complicated functions - even if one is better, can verify this only by calculations?
**Broad Issues**

- Behavior at high rate? Hamming bound is good enough.
- Behavior at low-rate? Codes can’t have $\delta > 1 - 1/q$, but Hamming bound can’t prove this! Griesmer bound does, but only good for linear codes. Plotkin bound will work.
- Asymptotic behavior? Given $k, \epsilon$, How does $n$ behave is we want $\delta = 1 - 1/q - \epsilon$. Elias-Bassalygo bound will give a decent bound: $n = \Omega(k/\epsilon)$. LP bound gives the correct result $n = \Omega(k/\epsilon^2)$.

**Proof Idea**

- Will omit proof of Plotkin bound.
- Will start with Elias-Bassalygo and this will motivate the Johnson bound.
- Johnson bound: Proven via a geometric argument. (Proof + improved bound from [Guruswami+S.’01].)

**Bounds - Round II**

**Plotkin Bound:**
If $d \geq (1 + \epsilon) \cdot (1 - \frac{1}{q}) \cdot n$ then
$\# \text{ codewords} \leq 1 + \frac{1}{\epsilon}$.

**Elias-Bassalygo Bound:**
$$R \leq 1 - H_q \left( (1 - \frac{1}{q}) \cdot (1 - \sqrt{1 - \frac{q - 1}{q-\delta}}) \right).$$

**Johnson Bound:** If $C$ is an $(n, ?, d)_q$ code then any Hamming ball of radius at most $e$ contains at most $nq$ codewords, provided
$$e/n < (1 - \frac{1}{q}) \cdot \left( 1 - \sqrt{1 - \frac{q - 1}{q-\delta}} \right).$$

(Never mind the actual numbers for now.)

**Elias-Bassalygo Bound**

- Pushes the packing bound.
- Go to larger radius.
- Suppose: Can prove that at most 4 balls of radius $e = 2d/3$ contain any one given point.
- Previous argument gives:
$$V(n, 2d/3, q)q^k \leq 4q^n.$$
- Lose almost nothing on RHS.
- Improve LHS (significantly).
Motivates the Johnson question.
Johnson Bound

Question: Given \( \bar{r} \in \Sigma^n \), \((n, k, d)\) code \( C \).
How many codewords in \( B(\bar{r}, e) \)?

Motivation: (for binary alphabet)
How to pick a bad configuration?
I.e. many codewords in small ball.
W.l.o.g. set \( \bar{r} = \bar{0} \).
Pick \( c_i \)'s at random from \( B(\bar{0}, e) \).

Expected’ dist. between codewords \( = ? \)
Let \( e = e/n \).
Codewords simultaneously non-zero on
\( e^2 \) fraction of coordinates;
Thus distance \( \approx (2e - 2e^2)n \).

Johnson bound shows you can’t do better!

Hamming to Euclid

\begin{itemize}
  \item Map \( \Sigma \to \mathcal{R}^q \): \( \text{ith element } \mapsto 0^{i-1}10^{q-i} \).
  \item Induces natural map \( \Sigma^n \to \mathcal{R}^{mn} \):
    \begin{itemize}
      \item Maps vectors into Euclidean space.
      \item Hamming distance large implies Euclidean distance large.
    \end{itemize}
  \item Argue: Can’t have many large vectors with pairwise small inner products.
\end{itemize}

Hamming to Euclid (contd).

In our case:

Given: \( c_1, \ldots, c_m \) codewords in \( \Sigma^n \) and \( \bar{r} \in \Sigma^n \), s.t.
\begin{itemize}
  \item \( \Delta(c_i, \bar{r}) \leq e \)
  \item \( \Delta(c_i, c_j) \geq d \)
\end{itemize}
Want: Upper bound on \( m \).

After mapping to \( \mathcal{R}^{mq} \)
(and abusing notation)

Given: \( c_1, \ldots, c_m \in \mathcal{R}^{mq} \) and \( \bar{r} \in \mathcal{R}^{mq} \), s.t.
\begin{itemize}
  \item \( \langle \bar{r}, \bar{r} \rangle = n \).
  \item \( \langle c_i, c_i \rangle = n \).
  \item \( \langle c_i, \bar{r} \rangle \geq n - e \)
  \item \( \langle c_i, c_j \rangle \leq n - d \)
\end{itemize}
Want: Upper bound on \( m \).

Main idea: Find a new point \( O' \) to set as origin, such that the angle subtended by \( C_i \) and \( C_j \) at \( O' \) is at least \( 90^\circ \).

Conclude: \# vectors \( \leq \) dimension \( = nq \).
Johnson bound (contd).

How to pick the new origin?

Idea 1: Try some point of the form \( \alpha \vec{r} \).

Then \( \langle c_i - \alpha \vec{r}, c_j - \alpha \vec{r} \rangle \)
\[= \langle c_i, c_j \rangle - \alpha \langle c_i, \vec{r} \rangle + \alpha^2 \langle \vec{r}, \vec{r} \rangle \]
\[\leq (1 - \alpha)^2 n + 2\alpha e - d \]

Setting \( \alpha = 1 \), says: Need \( e \leq d/2 \).

Setting \( \alpha = 1 - e/n \) yields:
Need \( e/n \leq 1 - \sqrt{1 - \delta} \).

(Not quite what was promised.)

Back to Elias Bound

Plugging Johnson bound into earlier argument:

\[ k \leq (1 - H_q(\epsilon))n + o(n), \]

where \( \epsilon \) such that the Johnson bound holds for \( e = \epsilon n \).

Importance:

- Proves e.g. No codes of exponential growth with distance \((1 - 1/q)n\).
- Decently comparable with existential lower bound on rate from random code.

Johnson bound (contd).

A better choice for origin.

Idea 2: Try some point of the form
\[ \alpha \vec{r} + (1 - \alpha) \vec{Q}, \]
where \( \vec{Q} = (\frac{1}{q})^{\frac{m}{n}} \).

Appropriate setting of \( \alpha = 1 - e/n \) yields, the desired bound.

MacWilliams Identities

Defn: Weight distribution of code is \( \langle A_0, \ldots, A_n \rangle \), where \( A_i \) is \# codewords of weight \( i \).

- MacWilliams Identity determines weight distribution of code from weight distribution of its dual.
- Quite magical.
- Many nice consequences.
MacWilliams Identities

Thm:
• Let $A_0, \ldots, A_n$ wt. dist. of $C$.
• Let $A'_0, \ldots, A'_n$ wt. dist. of $C^\perp$.
• Let $W(y) = \sum_i A_i y^i$.
• Let $W'(y) = \sum_i A'_i y^i$.
• Then $W'(y) = \frac{(1+qy-q^{-1})W(1-y)}{|C|^{1+q^{-1}y}}$.

Implications:
- Equating coefficients of $y^i$, get $n+1$ linear equations in $2(n+1)$ variables.
- Natural use, gives weight distribution of primal given dual or vice-versa.
- Interesting use: Can compute weight distribution of MDS codes!

MacWilliams Identities (contd.)

Trivial Claim: Given $W_C$, can compute $W_{C^\perp}$.

Explicit version: (non-trivial)
$W_C(x_1 + y_1, x_1 - y_1, \ldots, x_n + y_n, x_n - y_n)
= |C| \cdot W_{C^\perp}(x_1, y_1, \ldots, x_n, y_n)$

Proof steps:

Bit case:
$W(y; x+y, x-y) = \sum_{b \in \{0,1\}} (-1)^{b,b'} W_b(x, y)$

Vector case:
$W_C(x_1 + y_1, x_1 - y_1, \ldots, x_n + y_n, x_n - y_n)
= \sum_{b \in \{0,1\}} (-1)^{b,c} W_b(x_1, y_1, \ldots, x_n, y_n)$

Code case:
$W_C(x_1 + y_1, x_1 - y_1, \ldots, x_n + y_n, x_n - y_n)
= \sum_{c \in C} \sum_{b \in \{0,1\}^n} (-1)^{b,c} W_b(x_1, y_1, \ldots, x_n, y_n)
= \sum_{b \in \{0,1\}^n} W_b(x_1, y_1, \ldots, x_n, y_n) \sum_{c \in C} (-1)^{b,c}
= |C| \cdot W_{C^\perp}(x_1, y_1, \ldots, x_n, y_n)$

MacWilliams Identity follows using:
$(1+y)^n W\left(\frac{1-y}{1+y}\right) = W_C(1+y, 1-y, \ldots, 1+y, 1-y)
\quad \text{and } W'(y) = W_{C^\perp}(1, y, \ldots, 1, y)$

MacWilliams Identities: Proof

(Will only do the Binary case)

Defn: The verbose generating function

(a) The generating function of a bit:
$W_b(x, y) = (1 - b)x + by$

(b) The generating function of a word:
$W_C(x_1, y_1, \ldots, x_n, y_n) = \prod_{i=1}^b W_c(x_i, y_i)$

(c) The generating function of a code:
$W_C(x_1, y_1, \ldots, x_n, y_n)
= \sum_{c \in C} W_c(x_1, y_1, \ldots, x_n, y_n)
= x_1 x_2 x_3 + x_1 y_2 y_3 + y_1 x_2 y_3 + y_1 y_2 x_3$
MDS Codes

Fact: Dual of MDS code is MDS.
Proof: Along lines of Singleton bound.
Fact: MDS code of dim $k$ has $(q - 1)\binom{n}{k}$ codewords of minimum weight.
Proof: By inspection.
Consequence: Have values for $n + 1$ variables out of $2(n+1)$ used in M.I. System turns out to have full rank.
Thm: # poly of degree $< k$ with $w$ non-zero evaluations at $n$ points is:
\[
\binom{n}{w} \sum_{j=0}^{w+k-n} (-1)^j \binom{w}{j} (q)^{w+k-n-j-1}.
\]

LP bound

- One more bound in literature.
- Strongest known bound.
- Analysis hard.
- So hard, one only has upper bounds on the LP bound.
- Current upper bound on LP bound is still far from random code or AG-code (so may not be optimal either).
- Will see LP later.
- However (only) bound proving that if $d = (\frac{1}{2} - \epsilon)n$, then $n = O(k/\epsilon^2)$. (Matches random code for small $\epsilon$.)

Alon’s proof for $\epsilon$-biased spaces

Thm: Suppose have binary code with $K$ codewords of length $n$ s.t. no two are have distance less than $(\frac{1}{2} - \epsilon)n$ or greater than $(\frac{1}{2} + \epsilon)n$: Then $K \leq 2n$, provided $\epsilon \leq \frac{1}{2\sqrt{n}}$.
Proof:
- Map 0 to 1 and 1 to $-1$, and normalize so that vectors have unit norm.
- Then inner products lie between $-2\epsilon$ and $2\epsilon$.
- Let $M$ be $K \times K$ matrix of inner products.
- $M$ close to identity matrix and hence has rank close to that of identity matrix. Specifically: rank $\geq \frac{K}{1+4(K-1)\epsilon^2}$.
- On the other hand, rank$(M) \leq n$. 