6.897 Algorithmic Introduction to Coding Theory September 18, 2002

Lecture 5

Lecturer: Madhu Sudan *Scribe: Jonathan Herzog*

1 Overview

- Existence of asymptotically good codes
- Gilbert codes / Random codes
- Vashamov codes / Random linear codes
- Bounds
- Wozencraft ensemble of codes

2 Random codes

So far, we have seen a number of codes, but we would like an asymptotically good one. To review, an asymptotically good code is a family of codes, indexed by the block length (n) where:

- $\frac{k}{n} \geq R > 0$ and
- $\frac{d}{n} \geq \delta > 0$

and R and δ are both constants. That is, we would like a code with a rate (R) and a minimum distance (δ) constant relative to the block length. Can we show that such codes exist? Yes, by examining *random* codes:

Theorem 1 (Gilbert) *There exists a code with a constant δ and a constant rate $R \geq 1 - H_2(\delta)$.*

Proof Consider a random code with block size n and minimum distance d, constructed according to the following algorithm:

1. Let $S \leftarrow \{0,1\}^n$.
2. Let $C \leftarrow \emptyset$.
3. While $S \neq \emptyset$, do:
   - Pick x randomly (uniformly) from S.
   - Let $C \leftarrow C \cup \{x\}$
   - Let $S \leftarrow S \setminus Ball(x, d-1)$

(Here, $Ball(x, d-1) = \{y \in S | \Delta(x,y) \leq d-1\}$, and $Vol_2(n, d-1)$ is the number of points in $Ball(x, d-1)$.) Let C be a code produced by the above algorithm. How large will C be? Each time you add a codeword, you remove at most $Vol_2(n, d-1)$ elements from S:

$$|C| \geq \frac{2^n}{Vol_2(n, d-1)}$$

If $\delta = \frac{d}{n}$ for a family of codes, then $Vol_2(n, d-1) \approx 2^{H_2(\delta)n}$. In which case

$$|C| \geq 2^{n(1-H_2(\delta))}$$

If $|C| = 2^k$, then $2^k \geq 2^{n(1-H_2(\delta))}$, or $k \geq n(1 - H_2(\delta))$. Hence, there exists a code C so that

$$R \geq 1 - H_2(\delta).$$

■

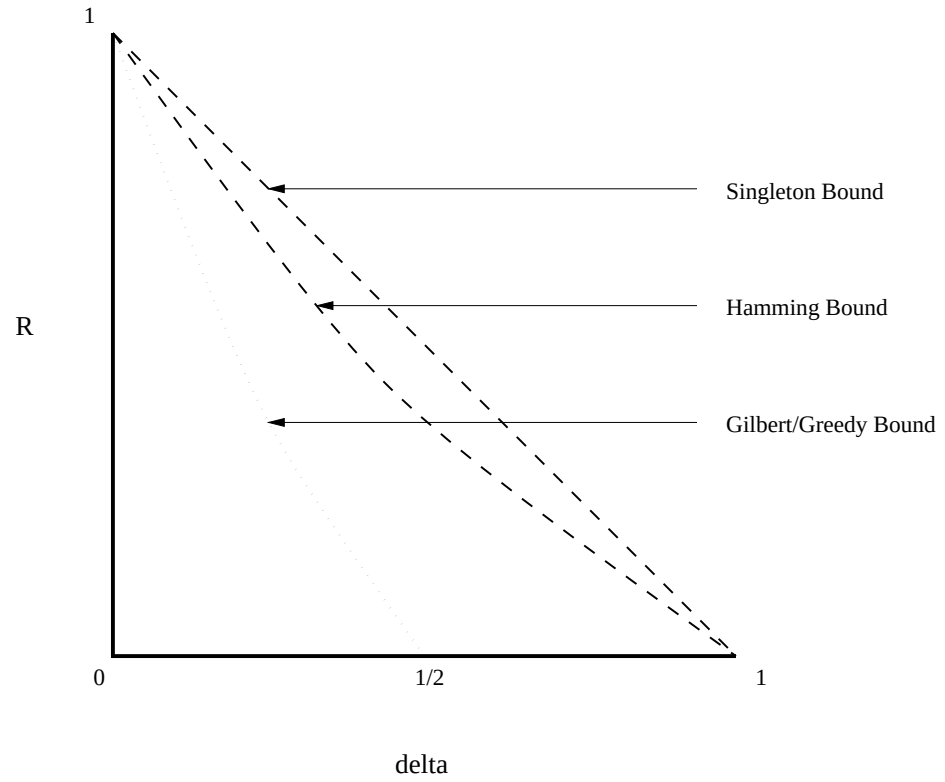

Figure 1: Bounds on R and δ

3 Bounds

How does this relate to the Hamming Bound? The Hamming bound tells us that if C is a code with distance d, then

$$|C| \leq \frac{2^n}{Vol_2(n, \frac{d-1}{2})}$$

(In the asymptotic case, $R \leq 1 - H_2(\frac{\delta}{2})$.) In other words, we know we can achieve a code with $Vol_2(n, d-1)$ in the denominator, and the Hamming bound tells us that we can do no better than $Vol_2(n, \frac{d-1}{2})$ in the denominator. Halving the radius reduces the volume of the sphere dramatically, leading to a large gap between this bound and the bound from the Gilbert proof. Figure 1 plots these bounds as a function of R and δ. The Gilbert proof above shows that there are codes on and to the left of the Gilbert bound. (Any code *on* the Gilbert bound which is not also on an axis is an asymptotically good code.) Hamming showed that there do not exist codes to the right of the Hamming bound. The area in between is still open.

4 Random Linear Codes

Theorem 2 (Vashamov) *Suppose that* $2^k - 1 \leq \frac{2^n}{Vol_2(n,d-1)}$. *Then there exists a linear code* C *with a minimum distance* d *and so that*

$$|C| \geq \frac{2^n}{Vol_2(n, d-1)}$$

Proof Pick $G \in \{0,1\}^{k \times n}$ at random. Then let $C = \left\{xG | x \in \{0,1\}^k\right\}$. To show that C has minimum distance d, it suffices to show that $\text{wt}(xG) \geq d$ for all $x \neq \mathbf{0}$.

Fix an $x \neq \mathbf{0}$. Then

$$\Pr_G[\mathrm{wt}(xG) < d] = \frac{Vol_2(n, d-1)}{2^n}$$

Hence, via union bound:

$$\begin{aligned} \Pr_G[\exists x \neq \mathbf{0} \text{ s.t. } \mathrm{wt}(xG) < d] &= \frac{\left(2^k - 1\right) Vol_2(n, d-1)}{2^n} \\ &< 1 \end{aligned}$$

Hence, there exists a G so that $\mathrm{wt}(xG) < d$) for all non-zero x, and so C is a linear code with minimum distance at least d. ■

5 Bounds, revisted

Can we do better than the Gilbert-Vashamov bound for, at least for specific n, k and d? Yes, we've already seen codes that do better. However, asymptotically, they exist on the *axises* of the diagram in Figure 1.

Hamming codes If we let $d = 3$, then the Gilbert-Vashamov construction gives a lower bound on the number of codewords as:

$$|C| \geq \frac{2^n}{Vol_2(n,2)} = \frac{2^n}{1 + n + \binom{n}{2}} \approx \Theta\left(\frac{2^n}{n^2}\right)$$

We know that Hamming codes actually do better than this: $|C| = \frac{2^n}{n+1}$. Asymptotically, however, d is constant and so δ goes to 0 as n gets large. Hence, Hamming codes are not asymptotically good.

Hadamard codes With these codes, $n = 2^k$ and $d = \frac{n}{2}$. So asymptotically, $R = \frac{k}{n}$ goes to 0 as n gets large and Hadamard codes are not asymptotically good. However,

$$Vol_2(n, \frac{n}{2}) \approx 2^{n-1},$$

and hence the Gilbert-Vashamov bound gives

$$|C| \geq \frac{2^n}{Vol_2(n, \frac{n}{2})} \approx 2$$

and we know that Hadamard codes do *much* better than this.

BCH codes If you fix d and let $n \to \infty$, then BCH codes have $\Theta\left(\frac{2^n}{n^{\frac{d-1}{2}}}\right)$ codewords. The Gilbert-Vashammov bound gives that it will have at least $\Theta\left(\frac{2^n}{n^{d-1}}\right)$ codewords, a very loose bound.

Reed-Solomon Codes In a q-ary construction, the greedy Gilbert-Vashamov bound implies that there are codes such that

$$k \geq n - d - \Theta\left(\frac{n}{\log q}\right).$$

Reed-Solomon codes, it turns out, are such that

$$k \geq n - d - 1,$$

a much better result. (Also, there are codes from algebraic geometry such that

$$k \geq n - d - \Theta\left(\frac{n}{\sqrt{q}}\right),$$

also a better result.)

Hence, is the Gilbert-Vashamov result tight? Madhu doesn't think so. However, all the "counterexamples" above lie on the R- or δ-axis of Figure 1. Hence, they are not asymptotically good codes, and are not *direct* counterexamples to the bound.

6 Wozencraft Ensamble of Codes

We would like to make the Gilbert-Vashamov "constructions" more deterministic, if possible. So much, in fact, that we are willing to accept an exponential ($2^{0(n)}$-time) algorithm to create a good linear code.

As a first step to that end, we consider *Wozencraft ensambles* of linear codes:

Definition 3 *The space* $\{0,1\}^n$ *is* packed *with linear codes* C_1, C_2, $\ldots C_t$ *(each having* 2^k *elements) if:*

1. *For all* $i \neq j$, $C_i \cap C_j = \{\mathbf{0}\}$, *and*
2. $\bigcup_i C_i = \{0,1\}^n$.

Note that if C_1, C_2, $\ldots C_t$ pack $\{0,1\}^n$, then $t = \frac{2^n-1}{2^k-1}$.

Theorem 4 *If* C_1, C_2, $\ldots C_t$ *pack* $\{0,1\}^n$ *and* $\epsilon t \geq Vol_2(n, d-1)$, *then more than an* $(1-\epsilon)$ *fraction of the* C_i*'s have distance* d.

Proof For all C_i of distance $\Delta(C_i) < d$, there exists a representative $v_i \in C_i$ in the set $Ball(\mathbf{0}, d-1) \setminus \{\mathbf{0}\}$. If $i \neq j$, then $v_i \neq v_j$. Hence, there can be only $|Ball(\mathbf{0}, d-1)| - 1$ codes with a representative that close to $\mathbf{0}$, and so only that many codes of distance less than d. Since $\epsilon t \geq Vol_2(n, d-1)$, the number of codes with a representative with in d of $\mathbf{0}$ must be less than ϵt. Hence, $t - \epsilon t = (1-\epsilon)t$ of the codes must have a distance larger than d. ■

Do packings exist? Let's build one. We need

$$t = \frac{2^n - 1}{2^k - 1}$$

codes. So, we first need $2^k - 1$ to divide $2^n - 1$, which happens if k divides n.

Now, to construct a packing, let $n = ck$, and construct C_1, C_2, $\ldots C_t$ as follows:

- Work over the field $\mathbb{F}_2^k$, interpreted as $\mathbb{F}_{2^k}$. Call this field K for convenience.
- In this interpretation, a message (usually an element of $\mathbb{F}_2^k$) will be a single field element of K.
- The space $\{0,1\}^n$ can now be viewed as $\mathbb{F}_{2^k}^c = K^c$. Hence, an codeword is now c field elements.

Consider vectors $(\alpha_1, \alpha_2, \ldots \alpha_c) \in K^c$ such that

1. $\alpha_i = 0$ for not all i, and
2. for the first i so that $\alpha_i \neq 0$, $\alpha_i = 1$

How many such vectors are there? If α_1 is the first non-zero entry, it must be 1 and there are $(2^k)^{c-1}$ ways to chose the $c-1$ remaining entries. If α_2 is the first non-zero entry, it must be 1 and there are $(2^k)^{c-2}$ ways to chose the $c-2$ remaining entries. So, the number of such vectors is:

$$\left(2^k\right)^{c-1} + \left(2^k\right)^{c-2} + \ldots \left(2^k\right) + 1 = \frac{\left(2^k\right)^c - 1}{2^k - 1} = \frac{2^{ck} - 1}{2^k - 1} = \frac{2^n - 1}{2^k - 1} = t.$$

Hence, we can associate each code with such a vector.

So, let $C'_{(\alpha_1, \alpha_2, \ldots \alpha_c)} : x \to (\alpha_1 x, \alpha_2 x, \ldots \alpha_c x)$ be a function from message to codeword. Then we can let the code

$$C_{(\alpha_1, \alpha_2, \ldots \alpha_c)} = \left\{ C'_{(\alpha_1, \alpha_2, \ldots \alpha_c)}(x) | x \in K \right\}$$

Have we packed the space K^c with these codes? To show that, we will give a function from $K^c \setminus \{\mathbf{0}\}$ to codes. Given $y = (y_1, y_2 \ldots y_c) \in K^c$, we can compute the index $(\alpha_1, \alpha_2, \ldots \alpha_c)$ of the code that contains it via:

- Suppose y_j is the first non-zero entry in y. Then it must be that $\alpha_1 = 0$, $\alpha_2 = 0$,$\ldots \alpha_{j-1} = 0$.

- Since the first non-zero element of $(\alpha_1, \alpha_2, \ldots \alpha_c)$ must be 1, we know that $\alpha_j = 1$ and $x = y_j$.
- $\alpha_{j+1} = \frac{y_{j+1}}{x}$,
- $\alpha_{j+2} = \frac{y_{j+2}}{x}$,and so on until
- $\alpha_c = \frac{y_c}{x}$.

Hence, each element of $K^c \setminus \{\mathbf{0}\}$ can belong to exactly one code, and so the codes pack the space K^c.