Introduction to Coding Theory **CMU: Spring 2010**

Notes 2: Gilbert-Varshamov bound

January 2010

Lecturer: Venkatesan Guruswami *Scribe: Venkatesan Guruswami*

1 Asymptotically good codes and Gilbert-Varshamov bound

We begin by answering the question raised at the end of the previous notes on the existence of asymptotically good codes.

Suppose we are interested in q-ary codes (not necessarily linear) of block length n and minimum distance d that have many codewords. What is the largest size such a code can have? This is a fundamental quantity for which we define a notation below.

Definition 1 *Let $A_q(n,d)$ be the largest size of a q-ary code of block length n and minimum distance d. The binary case is of special importance, and in this case $A_2(n,d)$ is denoted simply as $A(n,d)$.*

There is a natural greedy approach to construct a code of distance at least d: start with any codeword, and keep on adding codewords which have distance at least d from all previously chosen codewords, until we can proceed no longer. Suppose this procedure halts after picking a code C. Then Hamming balls in $\{0,1,\ldots,q-1\}^n$ of radius $d-1$ centered at the codewords of C must cover the whole space. (Otherwise, we can pick one more codeword which has distance at least d from every element of C, and the process would not have terminated.)

Definition 2 *For integers q, n, ℓ, denote by $\mathrm{Vol}_q(n,\ell)$ the volume of (i.e., the number of strings in) a Hamming ball of radius ℓ in $\{0,1,\ldots,q-1\}$. Note that this number does not depend on where the ball is centered and equals*

$$\mathrm{Vol}_q(n,\ell) = \sum_{j=0}^{\ell} \binom{n}{j} (q-1)^j \ .$$

Therefore, the greedy procedure terminates with a code C satisfying

$$|C| \cdot \mathrm{Vol}_q(n, d-1) \geq q^n \ .$$

We therefore have the following lower bound.

Lemma 3 (Gilbert-Varshamov bound) *The maximal size of a q-ary code of block length n and distance d satisfies*

$$A_q(n,d) \geq \frac{q^n}{\mathrm{Vol}_q(n,d-1)} = \frac{q^n}{\sum_{j=0}^{d-1} \binom{n}{j}(q-1)^j} \ . \tag{1}$$

There also exist linear codes of size given by the Gilbert-Varshamov bound:

Exercise 1 *By a suitable adaptation of the greedy procedure, prove that there also exists a linear code over $\mathbb{F}_q$ of dimension at least $n - \lfloor \log_q \mathrm{Vol}_q(n, d-1) \rfloor$.*

The Gilbert-Varshamov bound was actually proved in two independent works (Gilbert, 1952) and (Varshamov, 1957). The latter actually proved the existence of *linear codes* and in fact got a slightly sharper bound stated below. (You can verify that the Hamming code in fact attains this bound for $d = 3$.)

Exercise 2 *For every prime power q, and integers n, k, d, prove that there exists an $[n, k, d]_q$ linear code with*

$$k \geq n - \lfloor \log_q \Big(\sum_{j=0}^{d-2} \binom{n-1}{j} (q-1)^j \Big) \rfloor - 1 \ .$$

In fact, one can prove that a random linear code almost matches the Gilbert-Varshamov bound with high probability, so such linear codes exist in abundance. But before stating this, we will switch to the asymptotic viewpoint, expressing the lower bound in terms of the rate vs. relative distance trade-off.

1.1 Entropy function and volume of Hamming balls

We now give an asymptotic estimate of the volume $\mathrm{Vol}_q(n, d)$ when $d = pn$ for $p \in [0, 1 - 1/q]$ held fixed and n growing. This volume turns out to be very well approximated by the exponential $q^{h_q(p)n}$ where $h_q()$ is the "entropy function" defined below.

Definition 4 (Entropy function) *For a positive integer $q \geq 2$, define the q-ary entropy function $h_q : [0, 1] \to \mathbb{R}$ as follows:*

$$h_q(x) = x \log_q(q-1) - x \log_q x - (1-x) \log_q (1-x) \ .$$

Of special interest is the binary entropy function

$$h(x) = x \log \frac{1}{x} + (1-x) \log \frac{1}{1-x}$$

where we use the notational convention that $\log = \log_2$.

If X is the $\{0, 1\}$-valued random variable such that $\mathbb{P}[X = 1] = p$ and $\mathbb{P}[X = 0] = 1 - p$, then the Shannon entropy of X, $H(X)$, equals $h(p)$. In other words, $h(p)$ is the uncertainty in the outcome of a p-biased coin toss (which lands heads with probability p and tails with probability $1 - p$). The function h_q is continuous and increasing in the interval $[0, 1 - 1/q]$ with $h_q(0) = 0$ and $h_q(1-1/q) = 1$. The binary entropy function is symmetric around the $x = 1/2$ line: $h(1-x) = h(x)$.

We can define the inverse of the entropy function as follows. For $y \in [0, 1]$, the inverse $h_q^{-1}(y)$ is equal to the unique $x \in [0, 1 - 1/q]$ satisfying $h_q(x) = y$.

Lemma 5 *For an integer $q \geq 2$ and $p \in [0, 1-1/q]$,*

$$\mathrm{Vol}_q(n, pn) \leq q^{h_q(p)n} \ .$$

Proof: We have

$$\begin{aligned}
\frac{\mathrm{Vol}_q(n,pn)}{q^{h_q(p)n}} &= \frac{\sum_{j=0}^{pn} \binom{n}{j}(q-1)^j}{(q-1)^{pn}p^{-pn}(1-p)^{-(1-p)n}} \\
&= \sum_{j=0}^{pn} \binom{n}{j}(q-1)^j(q-1)^{-pn}p^{pn}(1-p)^{(1-p)n} \\
&= \sum_{j=0}^{pn} \binom{n}{j}(q-1)^j(1-p)^n\left(\frac{p}{(q-1)(1-p)}\right)^{pn} \ .
\end{aligned}$$

Since $p \leq 1-1/q$, $\frac{p}{q-1} \leq 1-p$, and therefore the above quantity is at most

$$\sum_{j=0}^{pn} \binom{n}{j}(q-1)^j(1-p)^n\left(\frac{p}{(q-1)(1-p)}\right)^{j} = \sum_{j=0}^{pn} \binom{n}{j}(1-p)^{n-j}p^j \ .$$

The latter sum is at most

$$\sum_{j=0}^{n} \binom{n}{j}(1-p)^{n-j}p^j = 1$$

by the binomial theorem. □

The above upper bound is tight up to lower order terms. The quantity $\mathrm{Vol}_q(n, pn)$ is at least as large as $\binom{n}{pn}(q-1)^{pn}$. By Stirling's formula $m! = \sqrt{2\pi m}(m/e)^m(1+o(1))$, it follows that

$$\binom{n}{pn} \geq \left(\frac{1}{p}\right)^{pn}\left(\frac{1}{1-p}\right)^{(1-p)n}\exp(-o(n)) = 2^{h(p)n-o(n)}$$

and therefore

$$\mathrm{Vol}_q(n, pn) \geq \binom{n}{pn}(q-1)^{pn} \geq q^{h_q(p)n-o(n)} \ .$$

For a self-contained derivation of the entropy estimate for the binomial coefficients, we can work with a crude estimate of $m!$ given by the integral estimate

$$\sum_{i=1}^{m-1} \ln i \leq \int_1^m \ln x \leq \sum_{i=2}^{m} \ln i$$

which gives

$$\frac{m^m}{e^{m-1}} \leq m! \leq \frac{m^{m+1}}{e^{m-1}} \ .$$

This immediately gives the lower bound

$$\binom{n}{pn} \geq 2^{h(p)n} \cdot \frac{1}{en\sqrt{p(1-p)}} \geq 2^{h(p)n-o(n)} \ .$$

We summarize the above discussion in the following important estimate.

Lemma 6 *For positive integers $n, q \geq 2$ and real p, $0 \leq p \leq 1 - 1/q$,*

$$q^{(h_q(p)-o(1))n} \leq \mathrm{Vol}_q(n, pn) \leq q^{h_q(p)n} \ .$$

1.2 Asymptotic form of GV bound

Combining the greedy construction of Lemma 3 with the estimate of the Hamming volume from Lemma 6 gives the following asymptotic version of the Gilbert-Varshamov bound.

Theorem 7 (Asymptotic Gilbert-Varshamov bound) *For every q and $\delta \in [0, 1 - 1/q]$, there exists an infinite family $\mathcal{C}$ of q-ary codes with rate*

$$R(\mathcal{C}) \geq 1 - h_q(\delta) - o(1) \ .$$

(In fact, such codes exist for every block length.)

Since $h_q(\delta) < 1$ for $\delta < 1 - 1/q$, the above implies that for every $\delta < 1 - 1/q$ there exists an asymptotically good family of q-ary codes of rate at least $R_0(\delta) > 0$ and relative distance at least δ. By Exercises 1 and 2 this also holds for linear codes over $\mathbb{F}_q$. We now give an alternate proof based on the probabilistic method.

1.3 Random linear codes attain the GV bound

Theorem 8 *For every prime power q, $\delta \in [0, 1 - 1/q)$, $0 < \epsilon < 1 - h_q(p)$, and sufficiently large positive integer n, the following holds for $k = \lceil (1 - h_q(\delta) - \epsilon)n \rceil$. If $G \in \mathbb{F}_q^{n \times k}$ is chosen uniformly at random, then the linear code with G as generator matrix has rate at least $(1 - h_q(\delta) - \epsilon)$ and relative distance at least δ with probability at least $1 - e^{-\Omega(n)}$.*

Proof: The claim about rate follows whenever G has full column rank. The probability that the i'th column is in the span of the first $(i-1)$ columns is at most q^{i-1}/q^n. By a union bound, G has rank k with probability at least $1 - \frac{k}{q^{n-k}} \geq 1 - e^{-\Omega(n)}$.

For each nonzero $x \in \mathbb{F}_q^k$, the vector Gx is a uniformly random element of $\mathbb{F}_q^k$. (Indeed, say that $x_k \neq 0$, then conditioned on the choice of the first $k-1$ columns G' of G, $Gx = G'x + g_k x_k$ is uniformly distributed since the k'th column g_k is chosen uniformly at random from $\mathbb{F}_q^n$.) Therefore the probability that $\mathrm{wt}(Gx) \leq \delta n$ is at most

$$\frac{\mathrm{Vol}_q(n, \delta n)}{q^n} \leq q^{(h_q(\delta)-1)n} \ .$$

Now a union bound over all nonzero x implies that the probability that the code generated by the columns of G has distance at most δn is bounded from above by

$$q^k q^{(h_q(\delta)-1)n} \leq q^{(1-h_q(\delta)-\epsilon)n+1} q^{(h_q(\delta)-1)n} = q \cdot q^{-\epsilon n} \leq e^{-\Omega(n)} \ .$$

We conclude that with probability at least $1 - e^{-\Omega(n)}$, the code generated by G has relative distance at least δ and rate at least $1 - h_q(\delta) - \epsilon$. □

Exercise 3 *Establish a similar result by picking a random $(n-k) \times n$ parity check matrix for the code.*

1.4 Some comments on attaining/beating the GV bound

We have seen that there exist binary linear codes that meet the Gilbert-Varshamov bound, and thus have rate approaching $1-h(\delta)$ for a target relative distance of δ, $0 < \delta < 1/2$. The proof of this was non-constructive, based on an exponential time algorithm to construct such a code (by a greedy algorithm), or by picking a generator matrix (or a parity check matrix) at random. The latter leads to a polynomial time randomized Monte Carlo construction. If there were a way to ascertain if a randomly chosen linear code has the claimed relative distance, then this would be a practical method to construct codes of good distance; we will have a Las Vegas construction that picks a random linear code and then checks that it has good minimum distance. Unfortunately, given a linear code, computing (or even approximating) the value of its minimum distance is NP-hard.

A natural challenge therefore is to give an explicit (i.e., deterministic polynomial time) construction of a code that meets the Gilbert-Varshamov bound (i.e., has rate R and relative distance close to $h_q^{-1}(1-R)$). Giving such a construction of binary codes (even non-linear ones) remains an outstanding open question.

For prime powers $q = p^{2k}$ for $q \geq 49$, explicit constructions of q-ary linear codes that not only attain but surpass the GV bound are known! These are based on algebraic geometry and a beautiful construction of algebraic curves with many rational points and small genus. This is also one of the rare examples in combinatorics where we know an explicit construction that beats the parameters obtained by the probabilistic method. (Another notable example is the Lubotzky-Phillips-Sarnak construction of Ramanujan graphs whose girth surpasses the probabilistic bound.)

What about codes over smaller alphabets, and in particular binary codes? The Hamming upper bound on size of codes (Lemma 13 in Notes 1) leads to the asymptotic upper bound $R \leq 1-h(\delta/2)$ on the rate. This is off by a factor of 2 in the coefficient of δ compared to the achievable $1-h(\delta)$ rate. We will later see improvements to the Hamming bound, but the best bound will still be far from the Gilbert-Varshamov bound. Determining the largest rate possible for binary codes of relative distance $\delta \in (0, 1/2)$ is another fundamental open problem in the subject. The popular conjecture seems to be that the Gilbert-Varshamov bound on rate is asymptotically tight (i.e., a binary code of relative distance δ must have rate $1-h(\delta)+o(1)$), but arguably there is no strong evidence that this must be the case.

While we do not know explicit constructions of binary codes approaching the GV bound, it is still interesting to construct codes which achieve good trade-offs. This leads to the following questions, which are the central questions in coding theory for any noise model (once some existential bounds are established on the trade-offs, the questions below pertaining to the worst-case or adversarial noise model where we impose no restriction on the channel other than a limit on the total number of errors caused):

1. Can one explicitly construct an asymptotically good family of binary codes with a "good" rate vs. relative distance trade-off?
2. Can one construct such codes together with an efficient algorithm to correct a fraction of errors approaching half-the-relative distance (or even beyond)?

We will answer both the questions in the affirmative in this course.