6.897 Algorithmic Introduction to Coding Theory September 30, 2002

Lecture 7

Lecturer: Madhu Sudan *Scribe: Jonathan Kelner, some sections by Adam Smith*

Today we'll describe a new family of codes, called algebraic-geometry codes. These codes generalize Reed-Solomon codes and yield surprisingly good codes over constant sized alphabets. We'll motivate the codes from two points of view. First we'll show how they challenge the "conjectured converse of the Gilbert-Varshamov bound". We'll then motivate the codes from the algebraic perspective.

The reader is warned that this lecture is not self-contained. It describes the "nature" of the codes without showing how to construct them, or why they even exist.

1 Motivation 1: Getting Better Parameters

q-ary Codes Today we shall concentrate on q-ary codes. Our alphabet will be $\mathbb{F}_q$, the finite field of size q. Our codes will be subsets $C \subseteq \mathbb{F}_q^n$.

The Singleton Bound The Singleton bound that we proved for binary alphabets can be established in this setting as well. Suppose $C \subseteq \mathbb{F}_q^n$ and that $|C| > q^{k-1}$. By the pigeonhole principle, the projection of C onto the first $k-1$ coordinates will have to send some two points to the same value. This shows that $\mathrm{dist}(C) \leq n-(k-1)$. We thus have, for *any* alphabet, a tradeoff between rate and relative distance: For any code,

$$R \leq 1 - \delta.$$

We would like to know how close to this bound we can get.

The q-ary Gilbert-Varshamov Construction One way to approach this bound is to perform a q-ary analogue of the Gilbert-Varshamov construction by greedily constructing random codes. Start with the origin in $\mathbb{F}_q^n$. Choose a second point that has distance greater than d from the origin, choose a third point that has distance greater than d from the first two points, etc. This yields codes with

$$q^k = |C| \approx \frac{q^n}{\mathrm{Vol}_q(d,n)}. \tag{1}$$

By noting that, for large n, most of the volume of $B_q(0,r)$ comes from points at distance exactly r from the origin (as opposed to distance less than r), we obtain

$$\mathrm{Vol}_q(d,n) \approx \binom{n}{d}(q-1)^d 1^{n-d} \approx q^{nH_q(d/n)}, \tag{2}$$

where $H_q(p)$ is the *q-ary entropy*:

$$H_q(p) = p\log_q \frac{q-1}{p} + (1-p)\log_q \frac{1}{1-p}.$$

Combining equations (1) and (2), we obtain

$$q^k \approx q^{n(1-H_q(\delta))},$$

so that there exist codes with

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

The q-ary Gilbert-Varshamov Bound If we fix δ and let q tend to ∞, we get

$$\begin{aligned}
H_q(\delta) &= \delta \log_q(q-1) + \frac{1}{\log q} H_2(\delta) \\
&= \delta + \frac{H_2(\delta)}{\log q} + O(\frac{1}{q \log q}) \\
&\quad \left(\text{using the fact that } \frac{\log(q-1)}{\log q} = 1 - \frac{\log q - \log(q-1)}{\log q} \approx 1 - 1/(q \log q) \right) \\
&\approx \delta + O(\frac{1}{\log q})
\end{aligned}$$

This means that for fixed δ, random codes will more or less achieve

$$R = 1 - \delta - O(1/\log q).$$

(Note: All logarithms are base 2 unless noted otherwise).

Note that the Singleton bound shows that no code can achieve $R > 1 - \delta$. The above bound shows that random (linear) codes approach the Singleton bound with an inverse logarithmic deficit in the alphabet size. This means that we need an alphabet that is exponentially large in $1/(1 - R - \delta)$.

It is thus natural to ask whether we can do better, i.e., whether we can find codes that approach the Singleton bound but use smaller alphabets. We begin by noting that we have already found one such family of codes.

Reed-Solomon Codes Recall that Reed-Solomon codes met the Singleton bound exactly and did so with an alphabet size of exactly n (for infinitely many choices of n). So Reed-Solomon codes seem to perform much better, although in this case one cannot really talk about q and n separately. With RS codes, we must have $q \geq n$, and so q must go to ∞ with n. Nonetheless, we know that $R = 1 - \delta - O(1/n)$ for RS codes (for any $q \geq n$), and so we can wave our hands and claim that we get

$$R = 1 - \delta - O(1/q).$$

So in effect the difference between $1 - \delta$ and R is growing inversely in q, rather than inversely in the logarithm of q. This motivates the question—can we somehow turn the Reed-Solomon intuition into a formal proof where we actually get to fix q and let n go to infinity and see the behavior of R vs. δ. algebraic geometry (AG) codes turn out to do exactly this.

AG codes The constructions of AG codes in fact yield

$$R = 1 - \delta - \frac{1}{\sqrt{q} - 1} = 1 - \delta - O(1/\sqrt{q}),$$

for every even prime power q (i.e., q must equal p^ℓ for prime p and even integer ℓ). These codes do not require that q scale with n i.e. in our "analysis" we may fix δ and q, and let $n \to \infty$; then we can let q increase and see how the parameters scale with q. While the codes do not achieve a deficit of an inverse in q, they do get a polynomial decay in this deficit as a function of q. So it becomes clear that as q grows this family of codes will outperform the Gilbert-Varshamov bound. Since the deficit functions are quite explicit, it is possible to compare them exactly and note that the function $\delta + 1/(\sqrt{q} - 1)$ is smaller than $H_q(\delta)$ for $\delta = \frac{1}{2}$ and $q \geq\approx 44$. The smallest square larger than this number is 49 and so we get that for $q \geq 49$ the algebraic geometry codes outperform random codes!

2 Motivation 2: Generalizing Previous Constructions

Recall that in previous classes we got codewords by taking multivariate polynomials and evaluating them at all points in $\mathbb{F}_q^m$ (RS codes were the univariate case $m = 1$). Consider the univariate and bivariate cases with degree ℓ:

Univariate case:	Yields $[q^2, \ell^2, q^2 - \ell^2]_{q^2}$-code.
Bivariate case:	Yields $[q^2, \ell^2, (q - \ell)^2]_q$-code.

Thus, reducing the alphabet size from q^2 to q cost us a reduction in the distance of $2\ell(q - \ell)$. Where does this difference come from? Intuitively, this is because in the bivariate plane $\mathbb{F}_q \times \mathbb{F}_q$, there are many small subspaces that encode quite inefficiently. For example, if we take any axis-parallel line in the plane. Knowing that a codeword is 0 on $\ell + 1$ of the points means it must be 0 at all q points on the line. Yet this code may still be non-zero elsewhere. Thus these q zeroes of the codeword only lead to $\ell + 1$ linear constraints on the codeword - a deficit roughly of $q - \ell$.

Another example of such a subspace is the circle $x^2 + y^2 = c$ for some constant c. One can prove that if the polynomial is 0 on 2ℓ of the points, then it must 0 everywhere on that circle—this could be up to $2q$ points, depending on c and on the field size.

The big idea: The main idea of algebraic geometry codes is to not evaluate your polynomials at every point in $\mathbb{F}_q \times \mathbb{F}_q$, but only at some carefully chosen subset. We thus ask, what's a good subset? It will turn out that a good answer is to choose some nice polynomial $R(x, y)$, and evaluate our polynomials at the points of

$$S = \{(\alpha, \beta) | R(\alpha, \beta) = 0\}.$$

3 History of AG codes

- AG codes were conceived by V.D. Goppa, a Russian coding theorist around 1975. When he published his first paper on this topic [3], it was not clear that the resulting codes would lead to new asymptotic results — in particular, the necessary algebra had not been studied yet. His paper motivated the study of the associated algebraic questions and eventually led to the breakthrough results.
- The first family of AG codes meeting the bound $R \geq 1 - \delta - \frac{1}{\sqrt{q}-1}$ were discovered by Tsfasman, Vladuts and Zink [7]. There underlying algebra was quite involved, and the constructions were very complicated. Manin and Vladuts [4] put some effort into showing that these codes were actually polynomial time constructible (with an $O(n^{30})$ construction time!).
- In a sequence of works Garcia and Stichtenoth [1, 2] simplified both the constructions and the proofs significantly. The resulting codes were built on curves that were completely explicit in the specification. The proofs involved in showing some of the properties are also significantly simpler. (One could even say these are "elementary", as works on algebraic geometry go.)
- Recent works by Shum et al. [6, 5] clarifies the Garcia-Stichtenoth papers further, eventually getting some codes with $\tilde{O}(n^3)$ construction time (the notation $\tilde{O}(\cdot)$ means ignoring polylog factors). The eventual hope is that these families will become completely explicit.

4 An Example in 2-D

We now return towards the task of describing algebraic-geometry codes. We will start by giving an example of a very *concrete* algebraic-geometry code — specifically a $[19, 6, 13]_{13}$ code. We will then attempt to show how the construction generalizes.

Our example will be a code based on the "plane" $\mathbb{F}_{13} \times \mathbb{F}_{13}$. We want to choose a subset of the plane with lots of points on which to evaluate low-degree bivariate polynomials in order to get codewords. We know that if we choose something that intersects too much with lines or circles, then we will have the same problem that we did with the whole plane — there will be subspaces don't contain much information.

Goppa's insight was to use an algebraic curve: pick a polynomial $R(x, y)$ of small degree, and consider the subset

$$S = V(R) = \{(x, y) \ : \ R(x, y) = 0\}.$$

In order to avoid intersecting too much with lines, circles and their other small-degree friends, we can choose R so that it's *irreducible* (see Bézout's theorem below). This, together with a judicious choice of which polynomials to use, will yield the desired properties.

[In lecture, we started doing the example that follows, and we then changed over to doing a different example. There was some interest in how the argument for the first example could be completed, so the proof is included. The example worked to completion in lecture will be discussed a little later.] In our example, we will use:

- $q = 13$, i.e. $\mathbb{F} = \mathbb{Z}_{13}$
- $S = V(R)$ given by $R(x, y) = y^2 - 2(x - 1)x(x + 1)$.
- The polynomials we will use as codewords are linear combinations of the 6 basis polynomials $\{1, x, y, x^2, xy, x^3\}$. Notice that we aren't taking all polynomials of a given degree, but a carefully chosen subspace.

The parameters given by this code are described below:

- $q = 13$: By choice.
- $n = 19$: This parameter is typically verified by exhaustive search. In this specific case, it maybe verified that $S = \{(0, 0), (\pm1, 0), (2, \pm5), (3, \pm3), (4, \pm4), (6, \pm2), (7, \pm3), (9, \pm6), (10, \pm2), (11, \pm1)\}$.
- $k = 6$: A message is a polynomial of the form $a_0 + a_1x + a_2x^2 + a_3x^3 + b_0y + b_1xy$ which is given by six coefficients, thus giving a message length of 6.
- $d = 13$, as we will argue below.

In general finding the block length (n) is non-trivial task, however the distance can be argued algebraically. In this special case, we do so by an ad-hoc argument tuned to give the best possible result. Later we will mention a slightly more general argument that is more illustrative of the principle behind the construction.

Claim 1 *Any non-zero polynomial $f(x, y) = a_0 + a_1x + a_2x^2 + a_3x^3 + b_0y + b_1xy$ is zero on at most six points in S.*

Proof We divide the analysis into two cases:
Case 1: $b_0 + b_1x$ does not divide $a_0 + a_1x + a_2x^2 + a_3x^3$.

Consider any common zero (α, β) of $f(x, y)$ and $R(x, y)$. Such a zero must also be a zero of any polynomial of the form $f \cdot g + R \cdot T$, for any polynomials $g(x, y)$ and $T(x, y)$. If we choose $g(x, y) = y(b_0 + b_1x) - (a_0 + a_1x + a_2x^2 + a_3x^3)$ and $T(x, y) = -(b_0 + b_1x)^2$, then the resulting polynomial $f \cdot g + R \cdot T$ is independent of y and equals $U(x) = 2(b_0 + b_1x)^2(x - 1)x(x + 1) - (a_0 + a_1x + a_2x^2 + a_3x^3)^2$, a polynomial of degree 6 is x. Since (α, β) should be a root of any such polynomial we conclude that α is a root of $U(x)$ and thus there are at most six possible choices for α.

Next we note that $b_0 + \alpha b_1 \neq 0$, since in such a case $a_0 + a_1\alpha + a_2\alpha^2 + a_3\alpha^3$ would also have to be zero, which contradicts the assumption for this case. So we can now use the relation $f(x, y) = 0$ to conclude that $\beta = -(a_0 + a_1\alpha + a_2\alpha^2 + a_3\alpha^3)/(b_0 + b_1\alpha)$ and thus the number of pairs (α, β) that satisfy both f and R is at most six.

Case 2: $b_0 + b_1x$ divides $a_0 + a_1x + a_2x^2 + a_3x^3$.

In this case $f(x, y) = (b_0 + b_1x)(y + c_0 + c_1x + c_2x^2) = f_1(x)f_2(x, y)$. Since every zero of f is a zero of f_1 or of f_2, we can divide this analysis into two parts.

Note first that $f_1(x)$ and $R(x, y)$ have at most two common zeros (α, β), with $\alpha = -b_0/b_1$ and β satisfying $\beta^2 = 2(\alpha - 1)\alpha(\alpha + 1)$.

Next eliminating y from $f_2(x, y)$ and $R(x, y)$ we find that any common zero (α, β) must satisfy

$$(c_0 + c_1\alpha + c_2\alpha^2)^2 = 2(\alpha - 1)\alpha(\alpha + 1),$$

$$\text{and } \beta = -(c_0 + c_1\alpha + c_2\alpha^2).$$

Again, we conclude that there are at most four choices of α satisfying the first condition, every such choice leads to one β satisfying the second condition. Thus f_2 and R have at most four common zeroes.

Putting the two parts together, we see that in this case also f and R have at most six common zeroes. ■

Hence, we get a $[19, 6, 13]_{13}$ code. In contrast a Reed-Solomon code could give a slight increase in the distance, to 14, for a big increase in the alphabet size, to 19. This demonstrates, non-asymptotically, some of the tradeoffs that become possible with AG codes.

Bézout's Theorem The example done in lecture used a general theorem from algebraic geometry known as "Bézout's theorem." Using it considerably simplifies the derivation of the distance of an AG code in the plane, and, in fact, is quite useful in choosing the appropriate R and C.

Theorem 2 (Bézout) *If $A(x, y), B(x, y)$ are polynomials of degree d_1, d_2 respectively, then if they share more than d_1d_2 zeroes, they must share a common factor.*

A proof of this fact can be found in most texts on algebraic geometry or algebraic curves (cf. [8, Theorem 3.1]). It is also available on the web at `http://theory.lcs.mit.edu/~madhu/FT98/lect18.ps`. The rough idea behind one proof of the theorem is to eliminate one of the variables y by finding polynomials $C(x, y)$ and $D(x, y)$ such that $A \cdot C + B \cdot D$ is a function of x alone. The fact that such a polynomial exists is not trivial, but not too hard to prove either. Once one gets this polynomial, it limits the number of choices in x and in turn one can limit the number of y's for every such x.

If, in our previous example, we chose our codewords to be linear combinations of the 6 basis polynomials $\{1, x, y, x^2, xy, y^2\}$, then Claim 1 would have been an immediate corollary of Bézout's theorem. Any linear combination f of the basis polynomials would be a polynomial of degree 2, and S is the zero set of R, a polynomial of degree 3. It is not difficult to see that R is irreducible (i.e., cannot be nontrivially factored), so f and R do not have any common factors. Bézout's theorem thus says that f and R do not have more than six common zeros, which is precisely the content of the claim.

5 A General Result

Generalizing the idea of the previous example to more than two variables and one polynomial relation among them, one builds AG codes as follows:

1. Pick $m - 1$ polynomial constraints on m variables:

$$\begin{aligned} P_1(x_1, ..., x_m) &= 0 \\ &\vdots \\ P_{m-1}(x_1, ..., x_m) &= 0 \end{aligned}$$

2. Let $S = V(P_1, ..., P_{m-1}) = \{\mathbf{x} : P_1(\mathbf{x}) = \cdots = P_{m-1}(\mathbf{x}) = 0\}$ be the set of common zeroes of $P_1, \ldots, P_{m-1}$.

3. Choose a linear subspace of polynomials which can't agree too often when restricted to S.

Of course, once again everything depends on how one chooses the polynomials $P_1, \ldots, P_{m-1}$ and then the basis of polynomials to evaluate at the set S. Specifically, one tries to pick polynomials $P_1, \ldots, P_{m-1}$ so that $|S|$ is large, while there exists a large collection of polynomials which don't agree too often on S.

Somewhat surprisingly, algebraic-geometers had been considering exactly this problem for a long time. A collection of polynomials is associated with a "curve" that consists of all zeroes of the polynomials over the algebraic closure, $\overline{\mathbb{F}_q}$, of $\mathbb{F}_q$. Such a curve consists of infinitely many points, but only finitely many are rational, i.e., from $\mathbb{F}_q^m$ (not surprising, since $\mathbb{F}_q^m$ is finite). To every curve they associate two integer parameters - its "genus" and the number of "rational points" lying on the curve. Both concepts are algebraic abstractions of analogous topological terms. Genus of a curve is a non-negative integer indicating the "twistedness" of the curve - the higher the genus, the more twisted the curve. From the point of view of establishing distance of codes, the best curves are the least twisted ones. However to get many rational points one needs twisted curves. This follows from a fundamental result in algebraic-geometry, first due to Hasse and Weil, and then improved by Drinfeld and Vladuts. The latter bound says that the number of rational points is at most the genus times $(\sqrt{q} - 1)$. The curves used by the AG codes are the "examples" showing the tightness of this bound. Once one finds curves matching this bound, a second "fundamental" result of algebra, known as the Riemann-Roch theorem, is invoked to show that a large basis of "polynomials" exists over this curve. We'll get to this part later. First we'll say something about curves of small genus with many rational points.

The first family of curves meeting the Drinfeld-Vladuts bound were found by Tsfasman, Vladuts and Zink [7]. Analyzing these curves was significantly hard. Subsequently much more elementary families were discovered by Garcia and Stichtenoth [1, 2]. We describe a family developed by them below.

Example ([1])

1. We assume $q = r^2$ for some prime power r.

2. The curves are described by $2m - 1$ polynomial equations over $2m$ variables $x_1, \ldots, x_m, y_1, \ldots, y_m$.

3. The polynomial equations are the following:

$$\begin{aligned} x_i^{r+1} &= y_i^r + y_i \qquad (i = 1, ..., m) \\ x_i x_{i+1} &= y_i \qquad (i = 1, ..., m-1) \end{aligned}$$

A relatively simple inductive argument shows that there are roughly q^m rational points (in this $2m$ dimensional space) giving the set S. The genus of this curve, determined by a so-called "Hurwitz's genus formula" is then established to be at most $|S|/(\sqrt{q} - 1)$. To choose the right space of polynomials for encoding, one then uses the notion of "order" of a polynomial. We'll omit its definition (along with so many others) but explain what properties it satisfies, since that will be useful in understanding how to work with the codes (for solving some algorithmic tasks).

The order of a polynomial behaves similarly to degree:

- $\text{ord}(\alpha f + \beta h) \leq \max\{\text{ord}(f), \text{ord}(h)\}$ where $\alpha, \beta \in \mathbb{F}_q$. Furthermore, $\text{ord}(\alpha f + \beta h) = \text{ord}(h)\}$ if $\beta \neq 0$ and $\text{ord}(f) < \text{ord}(h)$.

- $\text{ord}(f \cdot h) = \text{ord}(f) + \text{ord}(h)$

- If f is zero on $\text{ord}(f) + 1$ points on S, then $f \equiv 0$ on S.

By the properties above, it is clear that the set of polynomials of order at most t for a vector space. However unlike the case of univariate polynomials over $\mathbb{F}_q$, one need not have polynomials of every order.

The Riemann-Roch theorem shows, however that there do exist polynomials of all but g values of the order, where g is the genus of the curve. (This is why we like curves of small genus). Applying this theorem to the curves of Garcia and Stichtenoth one now gets the family of AG codes as claimed.

Specifically, let $n = |S|$ be the number of rational points on the curve S (fixed once q and m are fixed). By the fact that these curves meet the Drinfeld-Vladuts bound, we get that its genus $g \leq n/(\sqrt{q}-1)$. For any distance parameter d, let $\mathcal{P}_d$ be the set of all polynomials of order $n-d$. Notice that the evaluations of polynomials in L gives a linear code of distance d. By the Riemann-Roch theorem, we get that this space has dimension at least $n-d-g+1 \geq n-d-n/(\sqrt{q}-1)$. We obtain:

Theorem 3 (Very Good AG Codes Exist) *For every even power of a prime q, and every parameter $\delta < 1 - \frac{1}{\sqrt{q}-1}$, there exists an infinite family of q-ary linear codes of relative distance δ and rate $R \geq 1 - \delta - \frac{1}{\sqrt{q}-1}$. Further a generator matrix for such a code can be constructed in $\tilde{O}(n^3)$ time.*

We stress that this means that there are codes that outperform random codes. This is an important moral to take away—random objects are not always the best, even when they're very good.

References

[1] Arnaldo Garcia and Henning Stichtenoth. A tower of Artin-Schreier extensions of function fields attaining the Drinfeld-Vlǎdut bound. *Inventiones Mathematicae*, 121:211–222, 1995.

[2] Arnaldo Garcia and Henning Stichtenoth. On the asymptotic behavior of some towers of function fields over finite fields. *Journal of Number Theory*, 61(2):248–273, December 1996.

[3] V. D. Goppa. Codes associated with divisors. *Problems of Information Transmission*, 13(1):22–26, 1977.

[4] Y. I. Manin and Serge G. Vlǎdut. Linear codes and modular curves. *J. Soviet. Math.*, 30:2611–2643, 1985.

[5] Kenneth Shum. *A Low-Complexity Construction of Algebraic Geometric Codes Better Than the Gilbert-Varshamov Bound.* PhD thesis, University of Southern California, December 2000.

[6] Kenneth W. Shum, Ilia Aleshnikov, P. Vijay Kumar, Henning Stichtenoth, and Vinay Deolalikar. A low-complexity algorithm for the construction of algebraic geometric codes better than the Gilbert-Varshamov bound. *IEEE Transactions on Information Theory*, 47(6):2225–2241, September 2001.

[7] Michael A. Tsfasman, Serge G. Vlǎdut, and Thomas Zink. Modular curves, Shimura curves, and codes better than the Varshamov-Gilbert bound. *Math. Nachrichten*, 109:21–28, 1982.

[8] Robert J. Walker. *Algebraic Curves.* Springer-Verlag, 1978.