What is the maximum number of points on a curve over F_2?

By Yu. I. MANIN

To the memory of Takuro Shintani

1. Introduction. Let X be a smooth irreducible projective curve of genus g over F_2. An upper bound for the number of F_2-points on it is furnished by the classical Weil inequality $N(X)=\text{card}\, X(F_2)<3+2\sqrt{2}\,g$. It turned out that for large g this bound is far from being perfect. In fact

$$N(X)<(2+o(1))g$$

and even this inequality which follows from the simpler results on the linear error-correcting codes can be slightly improved (cf. below).

On the other hand if one could find curves with $N(X)>1.7096g$ this would result in the existence of binary linear codes with quite good parameters (*transmission rate, number of corrected errors*).

For $q=p^2$, $p\geqq 7$ modular curves have sufficiently many F_q-points and furnish asymptotically very good linear error-correcting codes over F_q.

In this contribution I wish to describe for algebraic geometers a remarkable connection between algebraic curves and codes, discovered recently by V. D. Goppa, and to draw their attention to the interesting unsolved problems. The new results expounded here are due to V. D. Goppa and M. Tsfasman.

I am grateful to S. I. Gelfand for the very helpful discussion.

2. Codes. Let F_q be a finite field with q elements. A linear (κ, n)-code over F_q is a κ-dimensional linear subspace $C\subset F_q^n$. The weight $d=d(C)$ of this code is the minimum weight of a vector $c\in C\smallsetminus\{0\}$ i.e. the number of its non-zero coordinates. The main problems of the coding theory are optimization problems. In particular one wants to maximize the parameters (κ, d) and to find good (κ, d, n)-codes with simple decoding algorithms. Here we will only consider the problem of asymptotic (large n) optimization of (κ, d). To be precise, fix q and set

$$x(C)=\frac{d}{n}, \qquad R(C)=\frac{\kappa}{n};$$

V_q=the family of points $(x(C), R(C))\subset[0, 1]^2$;

U_q=the set of limit points of V_q.

The simplest constructions of the coding theory lead to the following result.

3. Theorem. *There exists a continuous function* $\alpha_q(x)$, $x\in[0, 1]$, *such that*

$$U_q=\{(x, R)\,|\,0\leqq R\leqq\alpha_q(x)\}.$$

Moreover,

$$\alpha_q(0)=1;\ \alpha_q(x)\leqq\max\left\{1-\frac{q}{q-1}x, 0\right\}.$$

PROOF. Let C be a (κ, d, n)-code. For every $l\leqq\kappa$ there exists a subspace $C_l\subset C$ consisting of vectors with vanishing last l coordinates and such that $\dim C_l=\kappa-l$. Then $C_l\subset\boldsymbol{F}_q^{n-l}$ is a $(\kappa-l, d, n-l)$-code. On the (x, R)-plane the points corresponding to C_l's lie to the south-east of the point $\left(\frac{d}{n}, \frac{\kappa}{n}\right)$ on the line connecting $\left(\frac{d}{n}, \frac{\kappa}{n}\right)$ with (0, 1) and fill the whole segment down to the x-axis arbitrarily densely when $n\to\infty$. It follows that if $(x, R)\in U_q$, $x>0$, $R<1$, then the segment of the line from (x, R) to $\left(\frac{x}{1-R}, 0\right)$ entirely lies in U_q. To see this consider the sequence of codes $C^{(i)}$ with $\lim(x(C^{(i)}), R(C^{(i)}))=(x, R)$, $n(C^{(i)})\to\infty$, and look at the $(x(C_l^{(i)}), R(C_l^{(i)}))$.

In the same vein one can derive from a (κ, d, n)-code a family of $(\kappa-l, d-l, n)$ codes using quotients instead of subspaces (choose l places where a minimum weight vector of C has nonzero coordinates and divide them out). It follows as earlier that if $(x_0, R_0)\in U_q$ then the whole segment of the line $x-R=x_0-R_0$ to the south-west of this point lies in U_q.

Now set $\alpha_q(x)=\sup\{R\,|\,(x, R)\in U_q\}$. From the remarks above it follows that for $0\leqq x<y\leqq 1$ the point $(x, \alpha_q(x))$ lies in the "west light cone" of the point $(y, \alpha_q(y))$, whereas $(y, \alpha_q(y))$ lies in the "east light cone" of $(x, \alpha_q(x))$, the light cones being generated by two pencils of lines, $x-R=$const and lines going through (0, 1) respectively. Hence α_q is continuous.

The Plotkin bound ([1], 1.4.2) asserts that for any (κ, d, n)-code we have $\frac{d}{n}\leqq\frac{q-1}{q}\frac{q^\kappa}{q^\kappa-1}$. It follows that if $(x, R)\in U_q$, $R>0$, then $x<\frac{q-1}{q}$. Hence $(x, \alpha_q(x))$ cannot lie above the line connecting (0, 1) with $\left(\frac{q-1}{q}, 0\right)$. On the other hand, clearly points $0\leqq x\leqq 1$, $R=0$ and $x=0$, $0\leqq R\leqq 1$, lie in U_q. Finally, every point (x, R), $x>0$ below the graph of α_q belongs to U_q since it lies on the line connecting it with (0, 1) lower than the intersection point with this graph, whose existence follows from the Varshamov-Gilbert and Elias bounds (cf. below).

It seems unknown whether α_q is differentiable. From the bounds given below it follows that the graph of α_q has the vertical tangent at (0, 1) and the horizontal one at $\left(\frac{q-1}{q}, 0\right)$. The calculation of α_q or at least of good upper and lower bounds is one of the most important problems of the coding theory. We give below a list of some of the best known bounds.

4. Bounds for the codes. a) The Varshamov-Gilbert lower bound ([1], 3.3):

$$\alpha_q(x) \geqq 1-x\log_q(q-1)+x\log_q x+(1-x)\log_q(1-x) \overset{\text{def}}{=} \beta_q(x).$$

We have $\beta_q(0)=1$, $\beta_q\left(\frac{q-1}{q}\right)=0$; function β_q is convex, its graph lies below the line $R=1-\frac{q}{q-1}x$ and is tangent to coordinate axes at the ends. To compare this bound with the estimates for the number of points on curves it is helpful to know the tangent line to β_q with the equation $R+x=\text{const}$. This is

$$R+x=1-\log_q\frac{2q-1}{q} \qquad (R+x=1-0.584\cdots \text{ for } q=2)$$

and the point of contact is $x_0=\frac{q-1}{2q-1}$, $R_0=\frac{q}{2q-1}-\log_q\frac{2q-1}{q}$.

b) The Elias upper bound ([1], 1.4.3). Let $\beta_q^t(R)$ be the inverse function for β_q, similarly define α_q^t. Then

$$\alpha_q^t(R) \leqq 2\beta_q^t(R)-\frac{q}{q-1}(\beta_q^t(R))^2.$$

Near the point (1, 0) this bound comes to the R-axis at approximately double distance in comparison with the curve β_q. Hence α_q is smooth at this point. But at the point $\left(0, \frac{q-1}{q}\right)$ the Elias curve does not touch the x-axis and near this point lies even higher than the Plotkin line. The following bound is better in this respect.

c) The McEliece-Rodemich-Rumsey-Welch upper bound for $q=2$ ([2], 17.7):

$$z(x)=\frac{1}{2}-\sqrt{x(1-x)};$$

$$\alpha_2(x) \leqq -[z(x)\log_2 z(x)+(1-z(x))\log_2(1-z(x))] \overset{\text{def}}{=} \eta_2(x).$$

The graph of $\eta_2(x)$ quantitatively behaves in the same way as that of $\beta_2(x)$. The line

$$R+x=1-0.525\cdots$$

touches η_2.

We will see that for applications to curves we need also the corresponding line for the best existing bound for $q=3$: cf. [7].

5. Curves. We now introduce the number $\mathfrak{S}_q$ which gives the best asymptotically attainable upper estimate for the number of points on smooth projective irreducible curves over $\boldsymbol{F}_q$:

$$\mathfrak{S}_q(X)=(g-1)/(\operatorname{card} X(\boldsymbol{F}_q)-1), \quad g=\text{genus of } X;$$

$$\mathfrak{S}_q=\liminf \mathfrak{S}_q(X)\geqq \frac{1}{2\sqrt{q}} \qquad \text{(A. Weil)}.$$

The exact value of $\mathfrak{S}_q$ is unknown, some estimates are given below. The following theorem due to M. Tsfasman connects $\alpha_q(x)$ and $\mathfrak{S}_q$.

6. Theorem. *The segment of the line* $R+x=1-\mathfrak{S}_q$, $0\leqq R$, $x\leqq 1$ *lies entirely in the code domain* U_q.

PROOF. The points of this segment are the limit points for the Goppa codes, constructed by means of triples (X, D, G), where X is a smooth projective curve of genus g, $D=\sum_{i=1}^{n} P_i$, $P_i\in X(\boldsymbol{F}_q)$ are pairwise distinct points, G is an $\boldsymbol{F}_q$-rational divisor of degree a whose support is disjoint with the support of D. We will now define Goppa's codes.

Let $2g-2<a\leqq n+g-1$. Set $C=H^0(X, \Omega_x^1(D-G))$. The map

$$\operatorname{res}_D : C \longrightarrow \boldsymbol{F}_q^n, \qquad \operatorname{res}_D(\omega)=(\operatorname{res}_{P_1}\omega, \cdots, \operatorname{res}_{P_n}\omega)$$

is injective since $H^0(X, \Omega_x^1(-G))=0$. Hence the space C is endowed with the code structure. The beauty of this Goppa's structure consists in the invariant description of the weight function. The weight of the differential ω is now simply the order of its divisor of poles, which is at least $-(2g-2)+\deg G=-(2g-2)+a$. The dimension of C is $n+g-1-a$ by Riemann-Roch.

Now take all $\operatorname{card} X(\boldsymbol{F}_q)=n+1$ points $P_0, \cdots, P_n$ on X and set $D=\sum_{i=1}^{n} P_i$, $G=aP_0$. Setting $\alpha=\dfrac{a}{n}$ we get points in V_q, corresponding to the Goppa codes with

$$x\geqq -2\mathfrak{S}_q(X)+\alpha, \qquad R=1+\mathfrak{S}_q(X)-\alpha,$$

α running through the rational numbers between $2\mathfrak{S}_q(X)$ and $1+\mathfrak{S}_q(X)$ with the denominator n. Letting X run through the sequence of curves with $\lim \mathfrak{S}_q(X)=\mathfrak{S}_q$ we get the result.

7. Codes applied to curves. Every upper bound $\alpha_q(x)\leqq\bar{\alpha}_q(x)$ furnishes a

lower estimate for $\mathfrak{S}_q$ in view of the theorem 6. Namely, $\mathfrak{S}_q \geqq \overline{\mathfrak{S}}_q$, where $R+x=1-\overline{\mathfrak{S}}_q$ is the supporting line for the graph of $\bar{\alpha}_q$.

The bound of the theorem 4, $\bar{\alpha}_q(x)=\max\left(1-\frac{q}{q-1}x, 0\right)$ for $q=2, 3$ ameliorates Weil's estimates:

$$\mathfrak{S}_2 \geqq \frac{1}{2} > \frac{1}{2\sqrt{2}}, \qquad \mathfrak{S}_3 \geqq \frac{1}{3} > \frac{1}{2\sqrt{3}},$$

but for $q>3$ it does not give anything new. The McEliece et al. bound (n° 4c)) shows that

$$\mathfrak{S}_2 \geqq 0.525\cdots \quad \text{i.e.} \quad \text{card } X(\boldsymbol{F}_2) \leqq 1.901\cdots g$$

It would be nice to know all $\overline{\mathfrak{S}}_q$ which are better than Weil's bounds.

8. **Curves applied to codes.** Every upper estimate $\mathfrak{S}_q \leqq \widetilde{\mathfrak{S}}_q$ furnishes a lower bound for $\alpha_q(x)$ in view of the theorem 6. Namely, $\alpha_q(x) \geqq 1-\widetilde{\mathfrak{S}}_q-x$. If $\widetilde{\mathfrak{S}}_q < \log_q \frac{2q-1}{q}$, then this lower bound in a certain interval is better than the best known Varshamov-Gilbert bound (see 4a)).

In the articles [3], [4], Y. Ihara has proved deep results on the towers of modular curves over $\boldsymbol{F}_q$. These curves in particular furnish the infinite families with $\lim \mathfrak{S}_q(X)=\widetilde{\mathfrak{S}}_q=\frac{1}{\sqrt{q}-1}$ for $q=p^2$, p prime. Hence they ameliorate the Varshamov-Gilbert bound for $q \geqq 7^2$.

After the theorem 6 was proved Th. Zink and S. Vlăduţ, independently of the earlier Ihara's work, have indicated that $\widetilde{\mathfrak{S}}_q=\frac{1}{\sqrt{q}-1}$ for the classical modular curves and the Shimura modular curves.

9. **Questions and remarks.** a) Is it possible to find curves over $\boldsymbol{F}_2$ and $\boldsymbol{F}_3$ with

$$\text{card } X(\boldsymbol{F}_2) > 1.71\cdots g(X), \qquad \text{card } X(\boldsymbol{F}_3) > 2.16\cdots g(X)\,?$$

This would garantee the existence of good codes over $\boldsymbol{F}_2$ and $\boldsymbol{F}_3$, those fields being widely used in applications. Of course the problem of good decoding algorithms for Goppa's codes based on curves of high genus should be investigated separately.

b) How can one construct curves with many points? Consider an n-dimensional abelian variety A over $\boldsymbol{F}_{q^2}$ whose all Frobenius roots are $-q$. Let $X \subset A$ be an irreducible curve of genus $g \geqq n$ generating A. Then among its Frobenius roots there are $2n$ equal to $-q$, hence

$$\text{card } X(\boldsymbol{F}_{q^2}) \geqq q^2+2nq+1-2(g-n)q\,.$$

If A is (isogenous to) the jacobian, i.e. one can take $g=n$, we get the curve with the maximum possible number of points. The curve might be still useful if g is not much larger than n. On the other hand, if $\mathfrak{S}_{q^2}>\frac{1}{2q}$, this shows that such "very supersingular" abelian varieties are not jacobians.

c) One could try to seek curves with many points among the Drinfeld modular curves [5].

d) As an afterthought, one realizes that Bombieri's proof [6] of the Weil upper bound, using only Riemann-Roch, is quite code-theoretic in spirit. It might be interesting to systematically investigate linear systems on curves as codes, in particular to fumble with their weight structure. Last but not least, the problems of the efficient computability of complex Goppa codes are wide open.

Note added October 30, 1981.

Professor Ihara kindly let me know the results of his paper [8] before the publication. Refining his method of proof, V. G. Drinfeld and S. Vlăduţ recently succeeded to establish the fundamental inequality $\mathfrak{S}_q^{-1}\geqq(\sqrt{q}-1)$, or, in Ihara's notation, $A(q)\leqq\sqrt{q}-1$, for all q. In view of the Ihara's examples this bound is in fact exact for $q=p^{2m}$. The odd power case $q=p^{2m+1}$ remains unsettled as yet and thus the question posed at the title of my article unresolved. One can conjecture that $A(p^{2m+1})=p^m-1$.

References

[1] Kasami, T., Tokura, N., Iwadari, Y. and Y. Inagaki, The coding theory (in Japanese), 1975.

[2] MacWilliams, F. J. and N. J. A. Sloane, The theory of error-correcting codes, North Holland, 1976.

[3] Ihara, Y., On congruence monodromy problems, vols. 1, 2, Lecture Notes, Univ. Tokyo, 1968-69.

[4] Ihara, Y., On modular curves over finite fields, in: Discrete subgroups of Lie groups and applications to moduli, Oxford University Press, 1973.

[5] Drinfeld, V. G., Elliptic modules, Mat. Sb. 94 (136) (1974), N4, 594-627 (in Russian).

[6] Bombieri, E., Counting points on curves over finite fields, Sém. Bourbaki, 430, 1973.

[7] Gabidullin, E. and V. Sidorenko, A general bound for the volume of codes, Problemy Peredači Informacii, 4 (1976), 31-35 (in Russian).

[8] Ihara, Y., Some remarks on the number of rational points of algebraic curves over finite fields, in this volume; i.e., J. Fac. Sci., Univ. Tokyo, Sect. IA 28 (1981). 721-724.

(Received July 1, 1981)

Steklov Institute of Mathematics
42, ul. Vavilova
Moscow V-333
USSR