BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 36, Number 3, Pages 399–404
S 0273-0979(99)00783-1
Article electronically published on June 8, 1999

Codes and algebraic curves, by Oliver Pretzel, Oxford Lecture Series in Mathematics and Its Applications, Clarendon Press, Oxford, 1998, xii + 192 pp., \$65.00, ISBN 0-19-850039-4

The year 1948 saw the publication of Andre Weil's monograph *Sur les Courves Algebriques et les Variétés qui s'en Deduisent* and of Claude Shannon's work on *The mathematical theory of communication*. The first gave the details of Weil's celebrated proof of the Riemann Hypothesis for algebraic curves over finite fields, while the second lay down the foundations for information theory, including the proof of Shannon's fundamental theorem which ascertains that a communication channel has a well defined capacity and that by using suitable transmission codes it is possible to transmit information at any rate less than channel capacity with an arbitrarily small probability of errors in transmission.

To the casual observer these two theories would appear to be mathematically distinct, and for many years each would pursue its own line of development. Weil's foundations set in motion developments in algebraic geometry with marvelous consequences for number theory. In the other direction, the probabilistic aspects of Shannon's theory flourished amidst developments in stochastic processes, martingale theory, and ergodic theory, while simultaneously generating a great deal of attention among electrical engineers interested in the construction of error correcting codes. The deployment of space probes for the exploration of deep space as well as the technological advances in the storage and retrieval of digital information are but two significant contributions made possible by advances in the theory of error correcting codes.

A quarter of a century after the publication of Weil's and Shannon's work, a visionary electrical engineer had the audacity to suggest that there was possibly a close connection between the theory of algebraic curves and the theory of error correcting codes. This was the birth of the family of algebraic geometric codes also known as Goppa codes. The monograph under review is an introduction to some of the most interesting developments of the last 25 years in the theory of algebraic geometric codes.

What is an error correcting code?

The best way to answer this question is to quote an example of "efficient coding" from Shannon's monograph ([11], §17, p. 80):

> Let a block of seven symbols be $X_1, X_2, \ldots, X_7$. Of these X_3, X_5, X_6, and X_7 are message symbols and chosen arbitrarily by the source. The other three are redundant and calculated as follows:
>
> $$\begin{array}{lll} X_4 & \text{is chosen to make} & \alpha = X_4 + X_5 + X_6 + X_7 \quad \text{even} \\ X_2 & \text{is chosen to make} & \beta = X_2 + X_3 + X_6 + X_7 \quad \text{even} \\ X_1 & \text{is chosen to make} & \gamma = X_1 + X_3 + X_5 + X_7 \quad \text{even} \end{array}$$

1991 *Mathematics Subject Classification.* Primary 94B27, 11T71, 14H05.

> When a block of seven is received α, β, and γ are calculated and if even called zero, if odd called one. The binary number $\alpha\beta\gamma$ then gives the subscript of the X_i that is incorrect (if 0 there was no error).

This is Shannon's brief description of Hamming's coding algorithm which is capable of detecting and correcting one error. This code has a very simple algebraic structure: the code words $X = (X_1, \ldots, X_7)$ are solutions of the linear system $H \cdot X = 0$, and thus the Hamming code is simply the null space of the matrix

$$H = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}$$

acting on the vector space $\mathbf{F}_2^7$ of binary vectors of length 7. The discovery of Hamming's error correcting code and the subsequent development of bigger and better error correcting codes, including that of Golay with all its beautiful relations to the Leech Lattice and Conway's Moonshine numerology, are landmarks in the mathematical landscape of twentieth century mathematics. (These developments are very clearly recounted in the mathematically interesting monograph by T. Thompson [12].)

Practical considerations related to the implementation of error correcting codes stimulated the development of efficient decoding algorithms for detecting and correcting more than one error, e.g., BCH codes, Reed-Solomon codes. It was soon realised that the fundamental principle underlying the mechanics of decoding was the Euclidean Algorithm. The basic idea is quite old and simple and goes back to Huygens in the seventeenth century; it utilises the fact that the partial quotients of a continued fraction represent the best rational approximations to a rational number. In the hands of coding theorists this principle applied to continued fractions with polynomial entries suggested that a received code word could be transformed into a polynomial approximation from which the original code word could be determined using a majority logic procedure. The highly technical but very practical Berlekamp-Massey algorithm as well as its many ingenious variants, all relatives of the polynomial Euclidean algorithm, became part of the day to day solutions of the problem of decoding.

Goppa's efforts in [5] to come to terms with the diophantine approximation problems connected with the decoding of BCH codes and his gradual realization that his interpretation of BCH codes as "linear algebra constructions" on the projective line over a finite field led him to the beautiful and original idea that the linear spaces constructed on algebraic curves also furnish new error correcting codes whose transmission and decoding rates would surpass the known bounds, provided one could optimize the error term in Weil's formula for counting rational points on an algebraic curve over a finite field. Once this insightful observation was made it became clear that what was needed was to find curves with plenty of rational points. Such curves, also known as modular curves, had been studied extensively by the Japanese school of number theorists, especially by Shimura and Ihara (see [7]).

The spectacular success of Goppa's construction sent all coding theorists back to study algebraic curves, and here they found insurmountable pedagogical blocks. Goppa's original intuition was based on his keen appreciation of the constructive approach of Brill-Noether to the study of Riemann's theory of algebraic curves. The temptation of many coding theorists trying to extend Goppa's work was to approach the subject of algebraic curves from Weil's geometric approach, a goal which for many proved to be unreachable. Weil himself had cultivated the myth that the geometric approach was essential to the deeper understanding of the birational properties of curves, especially with respect to the theory of correspondences needed in the proof of the Lefshetz fixed point formula for curves (see Weil [13], p. 396). It was a remarkable discovery by Bombieri (following some essential ideas of Stepanof) that the older approach of Hasse and Schmidt was sufficient to prove the Riemann Hypothesis [2]. In [8] Manin would later surmise that Bombieri's methods were very much code theoretic in spirit, a remark that would greatly compliment Berlekamp's well known dictum that "*coding theory is linear algebra with basis.*"

What is a geometric Goppa code?

Let us consider on the projective plane $\mathbf{P}^2$ the algebraic curve defined over the binary field $\mathbf{F}_2$ by the equation

$$\mathscr{C} : x_0^3x_1 + x_1^3x_2 + x_2^3x_0 + x_0^2x_1^2x_1^2x_2^2 + x_0^2x_2^2 + x_0^2x_1x_2 + x_0x_1^2x_2 = 0.$$

(See Serre [10].) It is of genus 3 and has seven finite rational points $\mathscr{P} := \{P_1, P_2, \ldots, P_7\}$. (The coordinates of these points correspond to the digits in the binary expansion of the numbers $1, 2, \ldots, 7$.) On this curve we consider the rational functions $\mathscr{L} = \{\varphi_1, \varphi_2, \varphi_2\}$ defined for $i = 0, 1, 2$ by $\varphi_i(x_0, x_1, x_2) = x_i$. Consider the 3×7 matrix defined by

$$H = (\varphi_i(P_j))_{\substack{i=0,1,2 \\ j=1,\ldots,7}}.$$

The geometric Goppa code associated to the data $\{\mathscr{C}, \mathscr{L}, \mathscr{P}\}$ is simply the null space of the matrix H acting on vectors in the space $\mathbf{F}_2^7$. The resulting code in this example is nothing new: it is a geometric construction of the Hamming code.

In general a Goppa code is obtained by starting with a finite field $\mathbf{F}_q$ (e.g. a finite extension of $\mathbf{F}_2$), $\mathscr{C}$ a smooth algebraic curve defined over $\mathbf{F}_q$, and a set $\mathscr{P} = \{P_1, \ldots, P_n\}$ of rational points on $\mathscr{C}$ defined over $\mathbf{F}_q$ (these are fixed points under the coordinate map $x_i \mapsto x_i^q$). With $\mathscr{P}$ we associate the divisor $P := P_1 + P_2 + \cdots + P_n$. Let D be a divisor on $\mathscr{C}$ whose support is disjoint from $\mathscr{P}$:

$$D = \sum_Q n_Q Q, \qquad (n_P = 0 \text{ if } P \in \mathscr{P})$$

with $\deg(D) = \sum_Q n_Q \deg(Q)$, where $q^{\deg(Q)}$ is the cardinality of the field in which the coordinates of Q lie. Let $\mathscr{L} = \mathscr{L}(D)$ be the linear space of functions φ on $\mathscr{C}$ whose associated divisor satisfies $(\varphi) + D \geq 0$, that is to say, the set of rational functions on $\mathscr{C}$ with poles of order at most n_Q when this coefficient is positive and with zeros of multiplicity at least $-n_Q$ in the contrary case. The set of vectors $c_\varphi = (\varphi(P_1), \ldots, \varphi(P_n))$ of length n with φ in $\mathscr{L}$ form what is usually called a **geometric Goppa code**. (In the terminology of Pretzel, these are called Goppa Residue Codes and are denoted by $C_L(\mathscr{P}, D)$.)

This definition would not have any merit were it not for the fact that the Riemann-Roch theorem allows one to determine the parameters of the code. It is clear that the length of each code word c_φ is $n = \deg(P)$. The dimension (also called the rank of the code) is $m = \dim_{\mathbf{F}_q} \mathscr{L}(D) - \dim_{\mathbf{F}_q} \mathscr{L}(D - P)$. The minimum Hamming distance is $d \geq n - \deg(D)$ (recall that this is the minimum number of nonzero entries in a nontrivial vector). If $\deg(D) < n$, then the dimension of the code is at least $\deg(D) + 1 - g$, where g is the genus of $\mathscr{C}$, and the minimum distance is at least $n - \deg(D)$.

What is the decoding problem?

Usually a code word c (more precisely a block of symbols) is sent through a communications channel and the received word f may have changed. If the transmitted word and the received word are the same, there is no error and the message is accepted as is. Under normal circumstances f differs from c, and one is interested in computing the difference $e = f - c$, called the error word, from information carried by f. To formulate the decoding problem for a Goppa code associated to the data $(\mathscr{C}, \mathscr{P}, \mathscr{L})$, we introduce a slight modification of the notion of syndrome. Let $f = (f_1, \dots, f_n)$ be a vector in $\mathbf{F}_q^n$, and let φ be a rational function on $\mathscr{C}$. We form the value

$$\varphi \cdot f = \sum_{P_i \in \mathscr{P}} \varphi(P_i) f_i$$

and call it the syndrome of f with respect to φ. If φ has a pole in $\mathscr{P}$, we put $\varphi \cdot f = \infty$. The points in $\mathscr{P}$ which are indexed by the non-zero coordinates in the error word e are called the error locations. A function θ on $\mathscr{C}$ with no poles in $\mathscr{P}$ and vanishing at all the error locations for e is called an error locator.

The significance of the above definition is that if the Hamming weight of the error word is at most t and if A and X are divisors disjoint from the support of $\mathscr{P}$ with $\deg(A) \leq t + r$ and $\deg(X) \geq t + r + 2g - 1$, then the linear space $\mathscr{L}(A)$ contains an error locator whenever $\dim \mathscr{L}(A) > t$. Furthermore the error word e is uniquely determined by any error locator $\theta \in \mathscr{L}(A)$ and the values of the syndromes $\varphi \cdot f$ of f with respect to the functions φ in $\mathscr{L}(X)$. In fact if $\mathscr{M}$ is a set of not more than $t + r$ points including the error locations of e and if $\varphi_1, \dots, \varphi_u$ is a basis for $\mathscr{L}(X)$, where $\deg(X) \geq t + r + 2g - 1$, then e is the unique solution of the linear system

$$\varphi_i \cdot f = \sum_{P_i \in \mathscr{M}} \varphi_i(P_j) e_j, \qquad 1 \leq i \leq u.$$

In this way the error correcting process becomes a problem of linear algebra, and for its solution several ingenious algorithms have been invented.

Pretzel's book

The book comprises fifteen chapters and is physically divided into two separate parts. Part 1 (chapters 1–7) is a pleasant and lucid introduction to the theory of Goppa codes and the associated decoding problems. Included in this part is a discussion of the bare essentials of the theory of curves; the basic notions are amply illustrated with a number of significant examples, among which the Klein quartic plays a dominant role. (The student of this part who wishes to gain a deeper understanding of the many number theoretic properties of the Klein quartic can consult the beautiful article by Mazur in [9].) After some curious historical remarks on the Greek origin of the study of algebraic curves, the author moves on in chapter 2 to the study of affine plane curves utilizing the Euclidean algorithm for univariate polynomials. The birational point of view, that is to say, the study of the properties of curves which are reflected in the field of functions on the curve, is the theme of chapter 3, while the geometric properties of curves is dealt with in chapter 4. Chapter 5 gives the details of the elementary theory of Goppa codes as described earlier. In the opinion of the reviewer, the high point of the book is found in chapters 6 and 7, where the reader is exposed to the technical aspects of the well known decoding algorithm due to Skorobogatov-Vladut and its strengthening by Duursma.

These algorithms follow in principle the ideas which are manifest in the practical Berlekamp-Massey decoding of BCH codes as well as the original ideas implicit in Goppa's diophantine approximation interpretation. The key technical improvement in Duursma's algorithm is the possibility of using certain refined filtrations in the linear spaces of functions associated to the base curve. These filtrations serve as a suitable substitute for the Euclidean algorithm.

In Part 2 (chapters 8–14) a very economical discussion of the theory of function fields is given which is tailored to the minimalist needs of electrical engineers who want to understand the developments arising from the study of Goppa codes. This part can be thought of as a layman's introduction to Chevalley's theory of algebraic functions. It is a far cry from Chevalley's rigid and austere but solid treatment, which Weil described (not without reason!) as a "...severely dehumanized book." (Weil [14], p. 397.) Pretzel draws heavily on the underlying classical ideas about algebraic functions and their associated Riemann surfaces. The development of the connections between closed points on algebraic curves over finite fields and discrete valuations on the field of functions follows very closely the Artin-Whaples approach. The proof given in chapter 12 for the Riemann-Roch theorem is preceded by a discussion of the theory of repartitions and differentials and is followed by the proof of the important Residue Theorem; both of these results are key tools in the study of the parameters of Goppa codes and their duals.

Chapter 13 is a brief discussion of the algebraic extensions of function fields. The behavior of the genus, differentials, discrete valuations, etc., under such extensions are discussed particularly as they apply to constant field extensions. In his reliance on the intuitions arising in the study of classical algebraic curves and Riemann surfaces to explain fundamental notions, the author has not neglected to temper his discussion by giving a brief introduction to the delicate problems in positive characteristic which may arise from inseparable field extensions and the lack of suitable trace functions in such situations. A much deeper and more complete treatment of the topics of this chapter would require the theory of ramification groups, a topic which is beyond the scope and level of Pretzel's book.

Chapter 14 initiates an elementary discussion of the bridge between the geometry on a not necessarily smooth curve and the (birational) properties of the associated function field of a smooth model. These include some elementary bounds on the genus of a curve. (The reader who wants to see a fuller discussion of the computation of the genus would do well to consult Abhyankar's beautiful expository treatment [1].)

The final chapter in Pretzel's book serves to place the family of Goppa codes within the context of the larger theory of error correcting codes. The Gilbert bound is proved, and the significance of Goppa codes as providing families of codes with rates that surpass the Gilbert bound is discussed.

A brief but critical remark on page 10 describes the role of Gröbner bases in calculations where the Euclidean algorithm is not available. This is certainly the most promising line to study at present, and the reader who has reached the final chapter would be amply rewarded by consulting the excellent survey article by Hoholdt, van Lint and Pellikan [6].

The overall organization of the material together with the ample supply of examples makes Pretzel's book quite suitable as an introduction to the study of Goppa codes for both pure and applied mathematicians.

Bibliography

[1] S. S. Abhyankar, *Historical ramblings in algebraic geometry and related algebra*, Amer. Math. Soc. Monthly, vol. 83 (1976), 409–448. MR **53:**5581

[2] E. Bombieri, *Counting points on curves over finite fields* (d'aprés S. A. Stepanov), Sem. Bourbaki, No. 430 (1972/73). MR **55:**2912

[3] C. Chevalley, *Introduction to the Theory of Functions of One Variable*, Amer. Math. Soc. Math Surveys, New York, 1951. MR **13:**64a

[4] V. D. Goppa, *Geometry and Codes*, Mathematics and its Applications, vol. 24, Kluwer, Dordrecht, 1991. MR **91a:**14013

[5] V. D. Goppa, *Decoding and diophantine approximations*, Problems of Control and Information Theory, vol. 5 (1976), 195–206. MR **56:**11511

[6] T. Hoholdt, J. van Lint, and R. Pellikan, *Algebraic geometry codes*, Handbook of Coding Theory (V. S. Pless, W. C. Huffman, Eds.), Elsevier, 1998. CMP 99:07

[7] Y. Ihara, *Some remarks on the number of rational points of algebraic curves over finite fields*, J. Fac. Sci. Univ. Tokyo, vol. 28 (1981), 721–724. MR **84c:**14016

[8] Y. I. Manin, *What is the maximum number of points on a curve over F*? J. Fac. Sci. Univ. Tokyo, vol. 28 (1981), 715–720. MR **84c:**14015

[9] B. Mazur, *Arithmetic on curves*, A.M.S. Colloquium Lectures, Bull. Amer. Math. Soc., vol. 14, no. 2 (1986), 206–259. MR **88e:**11050

[10] J.-P. Serre, *Sur le nombre des points rationnels d'une courbe algébrique sur un corps fini*, C. R. Acad. Sci. Paris Sér. I, vol. 296 (1983), 397–402. MR **85b:**14027

[11] C. Shannon, *A mathematical theory of communication*, Bell. System Tech. J., vol. 27 (1948), 379–423, 623–656. MR **10:**133e

[12] T. M. Thompson, *From Error Correcting Codes through Sphere Packings to Simple Groups*, The Carus Mathematical Monographs, vol. 21, Mathematical Association of America, Washington, DC, 1983. MR **86j:**94002

[13] A. Weil, *Sur les Courbes Algebriques et les Variétés qui s'en Deduisent*, Hermann, Paris, 1948.

[14] A. Weil, *Review of "Introduction to the theory of algebraic functions, by C. Chevalley"*, Bull. Amer. Math. Soc., vol. 57 (1951), 384–398.

Carlos Moreno
City University of New York
E-mail address: carlos@kepler.baruch.cuny.edu