Lecture Notes for Mathematics 601
Error Correcting Codes and Algebraic Curves

Patrick J. Morandi

Fall 2001

Contents

1 Introduction to Coding Theory

1.1 Introduction to the Course

This course will discuss error correcting codes and connections with algebraic geometry. When coding theory began in the 1940s, the main mathematical technique was linear algebra. Later, ring theory was used, notably the theory of polynomial rings and quotient rings. In the 1970s, Goppa discovered a method for producing codes from algebraic curves, and his class of curves gave some nice theoretical results in the theory of error correcting codes. After spending some time on the basics of coding theory, we will develop the theory of algebraic curves and algebraic function fields in order to define Goppa codes and prove some results about their error correction capability. This will require spending some time on the basics of algebraic geometry, allowing non-algebraically closed base fields. However, since we will concentrate on curves, we will be able to bypass much of the difficult machinery of algebraic geometry. In fact, it is possible to work purely field theoretically, as does the book of Stichtenoth. However, I believe that this is too narrow a viewpoint, and that thinking geometrically gives better insight.

One aspect where Goppa codes are beneficial is in finding asymptotic bounds on codes. Without describing what this means, it is necessary to be able to produce long codes. One of the most important class of codes, BCH codes, can produce long codes but at the expense of requiring the use of increasingly large finite fields. Goppa codes, which can be viewed as a generalization of BCH codes, get around this problem.

Finding long Goppa codes reduces the problem of finding algebraic curves with many rational points. We will see that, for any algebraic curve over a finite field, there is an analogue of the Riemann Zeta function. (There is also such an analogue for every algebraic number field.) The Riemann hypothesis, which states that the Riemann Zeta's nontrivial zeros lie on the line $\mathrm{Re}(s) = \frac{1}{2}$, and which remains unproven, was proved by André Weil for the Zeta functions associated to curves over a finite field. We shall see how this fact leads to a bound on the number of rational points of a curve, and what implications this has for coding theory.

There is a website for this course. The URL is math.nmsu.edu/∼pmorandi/math601.

1.2 Definition of Error Correcting Codes

Let F be a finite field and let F^n be the collection of all n-tuples over F. This is an n-dimensional F-vector space. A *code* over F is a (nonempty) subset of F^n. It does not have to be a subspace although all of our examples will be subspaces. We will sometimes write elements of F^n as strings of elements without using commas or parentheses. The elements of a code are called *codewords*, and the elements of F^n are called *words*.

Before we give some notation, we will follow the normal practice of writing $\mathbb{F}_q$ for the unique, up to isomorphism, field with q elements.

Example 1.1. Let $F = \mathbb{F}_2$. Then $\{0, 1\}$ is a code in F and $\{00, 11\}$ is a code in F^2. Likewise, $\{000, 111\}$ is a code in F^3 and $\{00000, 11111\}$ is a code in F^5.

Example 1.2. Let

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

Then the kernel of H is a code in F^7. This is called the *Hamming code*, and was the first real example of an error correcting code. We will investigate this code further in a little while.

The important property of a code is its ability to correct errors. Let us discuss this idea starting with an example. In 1979 the Mariner 9 spacecraft took black and white pictures of Mars. These pictures were created by using 64 shades of grey. Each pixel of a photograph was assigned a shade of grey. Each picture consisted of a 600 by 600 grid of pixels. Thus, to transmit one photograph, the spacecraft had to send 360,000 pieces of data, each piece representing the color of a pixel. Suppose that the shades of grey were represented by a number from 1 to 64 in binary. Thus, we could represent any color with a string of six binary digits. If, in transmission, an error was made, then NASA would incorrectly color the corresponding pixel. Since electromagnetic activity can easily cause such errors, this would be a problem. NASA wanted an encoding system that would take received data, perform some sort of test, and determine if the received data was the same as that send, and, if not, determine what was the actual transmitted data. What they did was to encode each color as a string of 32 binary digits. Of the 2^{32} possible strings, 64 of them were valid codewords.

Suppose that a codeword is transmitted, but errors are made. One can recognize that an error is made if one sees that the received word is not a codeword. The main principle of error detection, Maximum Likelihood Detection (MLD) assumes that few errors are more likely than many errors. Therefore, the codeword that differs from the received word in the fewest number of components is the most likely transmitted codeword. All error correcting schemes use this principle. For example, with the code $\{000, 111\}$, if 101 is received, then MLD would assume that 111 was transmitted. However, with $\{00, 11\}$, if 01 or 10 was transmitted, then MLD would not distinguish between 00 and 11. Even worse, with the code $\{0, 1\}$, if a codeword is transmitted but an error is made, the error cannot be detected because all elements of F are codewords.

To make more precise the notion of closeness of words, we define a metric on F^n. The function $d : F^n \times F^n \to \mathbb{Z}$ defined by $d(u, v)$ is equal to the number of components in which u and v vary. In other words, if x_i is the i-th component of a vector x, then

$$d(u, v) = |\{i : u_i \neq v_i\}|.$$

The function d is in fact a metric. To see this, note that the property $d(u, v) \geq 0$ and $d(u, v) = 0$ if and only if $u = v$ is clear. Similarly, $d(u, v) = d(v, u)$ is clear. The triangle inequality is not hard, but is not completely obvious. We need to show that if $u, v, w \in F^n$,

then $d(u,w) \leq d(u,v) + d(v,w)$. Then

$$d(u,w) = |\{i : u_i \neq w_i\}|.$$

We note that $\{i : u_i \neq w_i\}$ is a subset of $\{i : u_i \neq v_i\} \cup \{i : v_i \neq w_i\}$ since if $u_i = v_i$ and $v_i = w_i$, then $u_i = w_i$. Thus,

$$\begin{aligned} d(u,w) = |\{i : u_i \neq w_i\}| &\leq |\{i : u_i \neq v_i\} \cup \{i : v_i \neq w_i\}| \\ &\leq |\{i : u_i \neq v_i\}| + |\{i : v_i \neq w_i\}| = d(u,v) + d(v,w). \end{aligned}$$

This metric will play a quiet but important role in coding theory.

1.3 Parameters of a Code

There are some important parameters attached to a code C over $F = \mathbb{F}_q$. The first, called the *length* of the code, is the value of n for which C is a subset of F^n. The next, which we will label k, is defined as

$$k = \log_q(|C|).$$

If C is a subspace of F^n, then $k = \dim_F(C)$, and so k is the *dimension* of the code. To define the final parameter, we need some preliminary concepts. The third invariant is labeled d and is the *distance* of the code C. It is defined as

$$d := \min\{d(u,v) : u \neq v \in C\}.$$

It should not be a problem that we are using d both for the metric and for the distance of a code. These three parameters are then positive integers. The parameters, in order n, k, d, of the codes of the first example are $(2,1,2)$, $(3,1,2)$, and $(5,1,5)$, respectively. With a little calculation, we see that the parameters of the Hamming code are $(7,4,3)$. Occasionally we will refer to a code with parameters n, k, and d as an (n,k,d)-code.

1.4 Linear Codes

A code that is a subspace of F^n is said to be a *linear code*. All of the codes we will consider in this course will be linear codes. We will view the elements of F^n as row matrices. There are some useful matrices attached to a linear code $C \subseteq F^n$. The first is called a *generator matrix*. It is a matrix G whose rows form a basis for C. This is an $k \times n$ matrix, and its row space is equal to C. The code C is then given, in terms of G, by

$$C = \left\{vG : v \in F^k\right\}.$$

The second matrix is called a *parity check matrix*. This is a matrix H of full rank for which C is the right nullspace of H^T. That is, $u \in C$ if and only if $uH^T = 0$. Alternatively, $u \in C$

if $Hu^T = 0$. Thus, by rethinking about codewords as column matrices, C is the nullspace of H. It is then an $(n-k) \times n$ matrix. Moreover, $GH^T = 0$ since $xGH^T = 0$ for all $x \in F^k$.

To see the symmetry of these matrices, we first give some notation. We will use $\cdot$ for the "usual dot product" on F^n; that is, $u \cdot v = \sum_i u_i v_i$. This is no longer an inner product since $u \cdot u = 0$ can occur with $u \neq 0$, for instance, if $F = \mathbb{F}_2$ and u has an even number of components equal to 1. In terms of matrix multiplication, $u \cdot v = uv^T$. Mimicking what is done for inner product spaces, we define the *dual code* $C^\perp$ of a code C by

$$C^\perp = \{u \in F : u \cdot v = 0 \text{ for all } v \in C\}.$$

We claim that H is a generator matrix for $C^\perp$ and G is a parity check matrix for $C^\perp$. To prove this, we first note the following matrix properties: if A is an $r \times s$ matrix, then (i) if $Ax = 0$ for all $x \in F^s$, then $A = 0$, and (ii) if $yA = 0$ for all $y \in F^r$, then $A = 0$. These are both straightforward to prove. As for the claim, first note that the row space of H is contained in $C^\perp$. For, if $x \in F^{n-k}$, then $xH \cdot v = (xH)v^T = x(vH^T)^T = x \cdot 0 = 0$ for all $v \in C$. This implies that $\dim(C^\perp) \geq n - k = n - \dim(C)$. However, if $u \in C^\perp$, then $0 = u \cdot xG = u(xG)^T = uG^T x^T$ for all $x \in F^k$. Therefore, $uG^T = 0$, and so $C^\perp$ is contained in the right nullspace of G. Because G has rank k and is a $k \times n$ matrix, its nullspace has dimension $n - k$. Thus, $\dim(C^\perp) \leq n - k$, and so both inequalities yield $\dim(C^\perp) = n - k$. Furthermore, this equality shows that the row space of H is $C^\perp$ and that $C^\perp$ is the right nullspace of G. This finishes the proof of the claim.

Example 1.3. Let C be the Hamming code. Then C has parity check matrix H, as defined earlier. A simple calculation shows that $\{1110000, 0101010, 1001100, 1101001\}$ form a basis for C. Thus, we may take

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

as a generator matrix for C. The code $C^\perp$ is then the nullspace of G, which has basis $\{0111100, 1101001, 1011010\}$. Thus,

$$C^\perp = \{0000000, 0111100, 1101001, 1011010, 1010101, 1100110, 0001111\}.$$

Having a linear code allows us to give an alternative description of the distance of a code. First of all, we define the weight of a word to be $w(u) = d(u, 0)$, the number of nonzero components of u. Then since $d(u, v) = w(u - v)$, we see that

$$d = \min\{w(x) : x \in C, x \neq 0\}.$$

By listing out the elements of the Hamming code, it is easy to see that the smallest weight

of a codeword is 3. For the dual code to the Hamming code, the listing of its elements shows that $C^{\perp}$ has distance 4.

Example 1.4. The extended Golay code, discovered by Golay, the codiscoverer of the Hamming code, has generator matrix

$$G = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1
\end{pmatrix}.$$

This is a 12×24 matrix. It can be viewed in the form $[B \mid A]$ with A and B both 12×12 matrices. In fact, $G = [I_{12} \mid A]$, where

$$A = \begin{pmatrix}
0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\
1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\
1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\
1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\
1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\
1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \\
1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 \\
1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 \\
1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\
1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1
\end{pmatrix}.$$

The Golay code is the code C, where C is the row space of G. The matrix G has rank 12, so the code has dimension 12. Thus, there are $2^{12} = 4096$ codewords. This code has distance 8, which can be verified in a hopelessly tedious manner by calculating the weight of all 4095 nonzero codewords, or by proving some facts, by induction, about the rows of G. One sees that the distance is at most 8 since the last row has weight exactly 8.

The Voyager spacecrafts visited Jupiter, Saturn, Uranus, and Neptune, taking pictures of each planet and their moons. The following picture was taken by Voyager 2. The photographs

taken by these spacecrafts utilized the Golay code to encode the data representing the pictures. A photograph consists of a rectangular grid, and at each grid point, or pixel, a color is given. For the Voyager spacecrafts, the use of the Golay code allowed the photos to be made with up to 4096 colors. Each color was represented as a codeword in the Golay code. To send the information representing one picture, each pixel was described by the codeword representing the color of that pixel. This codeword was a 24-tuple of binary digits. Since the Golay code has distance 8, it can correct up to three errors. Thus, each codeword transmitted could have up to three errors without losing any data.

1.5 Error Correction

We make formal the meaning of an error correcting code, and we shall shortly see the connection between the distance and the error correction capabilities of the code.

Definition 1.5. *A code C is s-error detecting if whenever at least one but at most s errors are made in any codeword, then the resulting word is not a codeword.*

From this definition and that of the distance of a code, it is clear that a code of distance d can detect up to $d-1$ errors. A more important definition for us is that of error correcting.

Definition 1.6. *A code C is said to be t-error correcting if a word is a distance at most t from some codeword, then its distance from every other codeword is greater than t.*

To help to understand this definition, suppose that a codeword v is transmitted and at most t errors are made, meaning at most t components of the codeword are changed, resulting in a word w. If the code is t-error correcting, then since w is a distance of at most t from v, then the distance from w to any other codeword is greater than t. Therefore, v is the closest codeword to w. Using MLD, we would correct w to v, and thus recover the correct codeword.

Example 1.7. The code $\{0, 1\}$ cannot correct any errors, nor can $\{00, 11\}$. However, $\{000, 111\}$ can correct 1 error. If 111 is transmitted but one error is made, the result has distance 2 from 000. The same is true starting with 000. The code $\{00000, 11111\}$ can correct two errors.

Example 1.8. The Hamming code can correct one error. Moreover, there is a nice decoding algorithm, if $F = \mathbb{F}_2$. Suppose that v is a codeword, and one error is made in transmitting v, resulting in a word w. Then $w = v + e_i$, where e_i is the i-th standard basis vector of F^7. Multiplying by the Hamming matrix H, we get $Hw = H(v + e_i) = Hv + He_i = He_i$. Now, an easy calculation shows that He_i is the i-th column of H. Therefore, if we identify which column of H is Hw, then we will determine i. Once we know i, we can recover v as $v = w + e_i$. This example illustrates an aspect of coding theory; we would like codes that can correct errors and for which there is an efficient decoding algorithm.

We know show the relation between the distance and the error correction capability of a code.

Theorem 1.9. *Suppose that a code C has distance d. Then C is t-error correcting for $t = \lfloor (d-1)/2 \rfloor$ but is not $(t+1)$-error correcting.*

Proof. We give a metric-theoretic proof of this fact. Let $t = \lfloor (d-1)/2 \rfloor$. Suppose that w is a word a distance at most t from v. We need to prove that w is a distance greater than t from any other codeword. Suppose that u is another codeword. If $d(w,u) \leq t$, then consider the closed disks of radius t centered at v and u, respectively. Then w is in both disks. By the triangle inequality, we have

$$d(v,u) \leq d(v,u) + d(u,w) \leq 2t < d,$$

a contradiction to the definition of d. Therefore, $d(w,u) > t$, so the code is indeed t-error correcting. To see that C is not $(t+1)$-error correcting, let u, v be codewords with $d(u,v) = d$. Note that $d = 2t+1$ or $d = 2t+2$, depending on whether d is odd or even. If we change $t+1$ of the components of u that differ from those of v to make them equal to the corresponding components of v, then we obtain a word w with $d(u,w) = t+1$ but $d(w,v) = d(u,v) - (t+1) = d - (t+1) \leq t+1$. This shows that C is not $(t+1)$-error correcting. □

From this theorem it is clear that the Hamming code is 1-error correcting since its distance is 3.

If we have a linear code C with parity check matrix H, we can use cosets to help decode. Given a word w, the *syndrome* of w is Hw. Therefore, the syndrome is 0 exactly when the word is a codeword. The possible syndromes are in 1-1 correspondence with the cosets of C, since $C + u = C + w$ if and only if $Hu = Hw$. If a codeword v is incorrectly transmitted as w, then $e = w - v$ is the error word. We have $He = Hw$, so e and w have the same syndrome. Since $C + w = C + e$, we look in the coset $C + w$ for the smallest weight vector; this is our error word e by MLD. To use this coset decoding, we need a list of syndromes and corresponding smallest weight error vectors. These vectors are called *coset leaders*. For example, the following table would be the coset decoding table for the Hamming code.

Syndrome	Coset Leader
000	0000000
001	1000000
010	0100000
011	0010000
100	0001000
101	0000100
110	0000010
111	0000001

The Maple worksheet *cosets.mws*, available on the class website, will produce this table.

1.6 Bounds on Codes

Suppose you have to transmit two pieces of information. You could use the code $\{0, 1\}$, but that would give you no error correction. You could use $\{000, 111\}$ to be able to correct one error, or up to $\frac{1}{3}$ of the digits. You could use $\{00000, 11111\}$ and correct up to two errors, or $\frac{2}{5}$ of the digits. In general, by making the strings longer, we can have better error correction. However, longer codewords mean more effort in sending, receiving, and decoding. So, we would like to have codes as short as possible with as good error correction as possible. What restrictions are there? Are there any relationships between the parameters n, k, and d? There are several relationships, all that give upper bounds or lower bounds for d. To state some of these bounds, we set $A_q(n, d)$ to be the largest size of a code of length n and distance d. Much of coding theory has been to determine the values of this function and in determining the asymptotic behavior of $A_q(n, d)$ as a function of $\delta = d/n$ as $n \to \infty$.

To help in some of the proofs, we first give a counting argument. Consider the sphere of radius r centered at a word v. To pick a word a distance i from v, we need to change i of the components of i; there are $\binom{n}{i}$ choices for these components. Each component can be changed in $q - 1$ ways since $|F| = q$. Therefore, there are $\binom{n}{i}(q-1)^i$ total words a distance of i from v. Therefore, the sphere contains a total of $\sum_{i=0}^{r} \binom{n}{i}(q-1)^i$ words.

There are many bounds on codes, although we restrict our attention to just three. The first bound gives a lower bound of the size of a code with given parameters n and d. Most bounds give upper bounds.

Proposition 1.10 (Gilbert-Varshamov Bound).

$$A_q(n, d) \geq \frac{q^n}{\sum_{i=0}^{d-1} \binom{n}{i}(q-1)^i}.$$

Proof. Let C be a code of maximal size of length n and distance d. The spheres of radius $d - 1$ centered at codewords then cover the entire space of words, since if there is a word a distance of at least d from every codeword, then we could add it to C without changing n or d. If M is the size of this code, then from our calculation of the size of spheres, we have $M \cdot \sum_{i=0}^{d-1} \binom{n}{i}(q-1)^i \geq q^n$, which yields the bound. □

This bound says that there exists a code of length n and distance d, and with M elements, such that $M \geq q^n / \left(\sum_{i=0}^{d-1} \binom{n}{i}(q-1)^i\right)$. It does not say anything about arbitrary codes.

Proposition 1.11 (Singleton Bound). *Let C be a code with parameters n, k, and d. Then $d \leq n - k + 1$. Therefore, $A_q(n, d) \leq q^{n-d+1}$.*

Proof. Consider the function $\varphi : C \to F^{n-d+1}$ given by $\varphi(x_1, \ldots, x_n) = (x_1, \ldots, x_{n-d+1})$. In other words, φ removes the last $d - 1$ components of a codeword. Since every codeword of

C has weight at least d, the function φ is 1-1. Consequently, $|C| \leq \left|F^{n-d+1}\right|$. Since this is true for any code of length n and distance d, we have $A_q(n,d) \leq \left|F^{n-d+1}\right| = q^{n-d+1}$. Taking logarithms to the base q, we get $k \leq n-d+1$, or $d \leq n-k+1$. □

If a code C satisfies $d = n-k+1$, then the code is said to be an *MDS code* (maximum distance separable).

The following bound is also called the sphere packing bound.

Proposition 1.12 (Hamming Bound). *If $t = \lfloor (d-1)/2 \rfloor$, then*

$$A_q(n,d) \leq \frac{q^n}{\sum_{i=0}^{t} \binom{n}{i}(q-1)^i}.$$

Proof. Suppose that we have a code with length n and distance d, and for which it has M codewords. Let v be a codeword, and consider the sphere of radius t centered at v. As we have seen this sphere contains $\sum_{i=0}^{t} \binom{n}{i}(q-1)^i$ words. Since the code can correct t errors, the spheres of radius t centered at codewords are disjoint. Therefore, since there are q^n total words,

$$q^n \leq M \sum_{i=0}^{t} \binom{n}{i}(q-1)^i,$$

or

$$M \leq \frac{q^n}{\sum_{i=0}^{t} \binom{n}{i}(q-1)^i}.$$

Since this is true for any such code, we get the Hamming bound.

These last two codes give an upper bound, in terms of n and d, of the number of elements of any code with length n and distance d. □

The codes Goppa constructed from algebraic curves yield a better asymptotic lower bound than the Gilbert-Varshamov bound. By asymptotic bounds, we mean the investigation of

$$\limsup_{n\to\infty} \frac{\log_q(A_q(n,d))}{n}.$$

The reason for considering asymptotic bounds comes from Shannon's channel coding theorem. If we define the *information rate* of a code to be k/n, then the theorem says that given any $\varepsilon > 0$, there is a code with information rate at least R for which the probability of incorrect decoding of a received word is less than ε. However, in order to make the probability of incorrect decoding low with a fixed k/n, then n must be large. This theorem is not stated quite correctly; there is a limit on how large the information rate can be; this limit is called the capacity of the channel. A description of this can be found in Roman's book.

1.7 Cyclic Codes

To define some of the important classes of codes, we first define cyclic codes. First of all, let $\sigma : F^n \to F^n$ be the *shift map*; that is, $\sigma(x_1, \ldots, x_n) = (x_n, x_1, x_2, \ldots, x_{n-1})$. A code C is said to be *cyclic* provided that $\sigma(C) = C$. To give an alternate description of cyclic codes, first consider the F-algebra of polynomials $F[x]$. We may view F^n as a subspace of $F[x]$ via the map $\varphi : (a_0, \ldots, a_{n-1}) \mapsto a_0 + a_1 x + \cdots + a_{n-1}x^{n-1}$. Then F^n is mapped isomorphically onto the subspace of polynomials of degree less than n. We can interpret σ with respect to this embedding. Since $\sigma(a_0, \ldots, a_{n-1}) = (a_{n-1}, a_0, \ldots, a_{n-2})$, we have

$$\begin{aligned}\varphi\left(\sigma(a_0, \ldots, a_{n-2})\right) &= a_{n-1} + a_0 x + \cdots + a_{n-2}x^{n-1} = a_{n-1} + x(a_0 + \cdots + a_{n-2}x^{n-2}) \\ &\equiv x(a_0 + \cdots + a_{n-2}x^{n-2} + a_{n-1}x^{n-1}) \bmod (x^n - 1).\end{aligned}$$

Because of this, if we view $F^n \cong F[x]/(x^n - 1)$, then σ corresponds to multiplication by x. Therefore, a code $C \subseteq F[x]/(x^n - 1)$ is cyclic if and only if $xC = C$. Since C is already an F-vector space, this additional condition is equivalent to that C be an ideal of $F[x]/(x^n - 1)$. Now, since $F[x]$ is a PID, every ideal of $F[x]/(x^n - 1)$ is principal and is generated by a divisor of $x^n - 1$. Thus, cyclic codes of length n are in 1-1 correspondence with divisors of $x^n - 1$.

Suppose that $x^n - 1$ factors as $x^n - 1 = g(x)h(x)$. Consider the code $C = (g(x))/(x^n - 1)$; that is, C is the code consisting of all polynomial cosets that are multiples of $g(x)$. If $g(x) = g_0 + g_1 x + \cdots + x^{n-k}$, then $\left\{\overline{g(x)}, \overline{xg(x)}, \ldots, \overline{x^{k-1}g(x)}\right\}$ is a basis for C, where $\overline{f(x)} = f(x) + (x^n - 1)$. Therefore, $k = n - \deg(g(x))$. We call $g(x)$ the *generator polynomial* of the code.

Let K be a finite extension field of F. If $\alpha_1, \ldots, \alpha_t \in K$, let $f_i(x)$ be the minimal polynomial over F of α_i, and let $g(x)$ be the least common multiple of the $f_i(x)$. Then $g(x)$ is the monic polynomial of least degree for which each α_i is a root. We then can use $g(x)$ to build a cyclic code of length n, where n is any integer for which $g(x)$ divides $x^n - 1$. This code then consists of all polynomial cosets $\overline{p(x)}$ for which $p(\alpha_i) = 0$ for all i. Note that for $g(x)$ to divide $x^n - 1$, we need each α_i to be a root of $x^n - 1$; that is, we need $\alpha_i^n = 1$.

Recall from the theory of finite fields, the multiplicative group of any finite field is cyclic. A generator of the multiplicative group F^* is called a *primitive element* of F. If α is a primitive element of $\mathbb{F}_{n+1}$, then all powers of α are roots of $x^n - 1$, and we can use powers of alpha to build a code of length n. Note that $n + 1$ must be a power of a prime for $\mathbb{F}_{n+1}$ to exist.

Example 1.13. With a rearrangement of its digits, the Hamming code is an example of a cyclic code. To see this, let α be a primitive root of $\mathbb{F}_8$. Since we may view $\mathbb{F}_8$ as $\mathbb{F}_8 = \mathbb{F}_2[x]/(x^3 + x + 1)$, we can view its elements as 3-tuples over $\mathbb{F}_2$. The nonzero elements $1, \alpha, \ldots, \alpha_6$ of $\mathbb{F}_8$ then correspond to the columns of the Hamming matrix H. By reordering the columns if necessary, we view $H = (1, \alpha, \ldots, \alpha^6)$, where we think of these elements as

column vectors. For example, if $\alpha = \overline{x}$, then

$$\begin{aligned}
\alpha^0 &= 100, \\
\alpha^1 &= 010, \\
\alpha^2 &= 001, \\
\alpha^3 &= 110, \\
\alpha^4 &= 011, \\
\alpha^5 &= 111, \\
\alpha^6 &= 101.
\end{aligned}$$

Therefore, we have

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

The condition for a vector $v = (a_0, \ldots, a_6)$ to be in the (rearranged) Hamming code is

$$0 = Hv = (1, \alpha, \ldots, \alpha_6) \cdot \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_6 \end{pmatrix} = a_0 + a_1\alpha + \cdots + a_6\alpha^6.$$

Therefore, by viewing v as a polynomial $v(x) \in \mathbb{F}_2[x]$ of degree at most 6, we see that v is a codeword if and only if $v(\alpha) = 0$. In fact, the generator polynomial is then the minimal polynomial of α, which is $x^3 + x + 1$.

Example 1.14 (BCH Codes). Let $n = q^m - 1$, and let α be a primitive element of the field $\mathbb{F}_{q^m}$. If e is a positive integer with $e \leq n$, we can use the elements $\alpha, \alpha^2, \ldots, \alpha^{e-1}$ to build a code. This code is then the set of polynomial cosets $\overline{p(x)}$ with $p(\alpha) = p(\alpha^2) = \cdots = p(\alpha^{e-1}) = 0$. While we do not show it here, this code has distance at least e, and so can correct $\lfloor (e-1)/2 \rfloor$ errors. The integer e is called the *designated distance* of the code, and is a lower bound of the actual distance.

Example 1.15 (Reed-Solomon Codes). Let $n = q - 1$ and let α be a primitive element of $\mathbb{F}_q$. Then the code of all polynomial cosets $\overline{p(x)}$ with $p(\alpha) = p(\alpha^2) = \cdots = p(\alpha^{d-1}) = 0$ is called a Reed-Solomon code. By the claim of the previous example, the distance of this code is at least d. The generator polynomial is $g(x) = (x - \alpha) \cdots (x - \alpha^{d-1})$, so $k = n - \deg(g(x)) = n - d + 1$. Therefore, by the Singleton bound, the distance of the code is exactly d, and the code is an MDS code.

Example 1.16. Suppose $n = 7$ and $d = 4$. The Reed-Solomon Code we get from this data has $k = 4$. The resulting code is

We now consider some specific examples of BCH codes.

Example 1.17. Consider $\mathbb{F}_{16} = \mathbb{F}_2[x]/(x^4+x+1)$, the splitting field over $\mathbb{F}_2$ of x^4+x+1. A short calculation shows that any root α of x^4+x+1 is a primitive element of $\mathbb{F}_{16}$. To see this, recall that since $\mathbb{F}_{16}^*$ has order 15, Lagrange's theorem implies that α is a primitive element provided that $\alpha^3 \neq 1$ and $\alpha^5 \neq 1$. Since the minimal polynomial of α is $x^4 + x + 1$, which does not divide x^3-1 or x^5-1, neither of these polynomials has α as a root. We consider the BCH code of all polynomials cosets having $\alpha, \alpha^2, \ldots, \alpha^6$ as roots. Then $n = 15$. To calculate k, we calculate the generator polynomial $g(x)$. This is the least common multiple of the minimal polynomials of the six roots. The minimal polynomial of α and α^2 is $x^4 + x + 1$. The minimal polynomial of α^3 and α^6 is $x^4 + x^3 + x^2 + x + 1$ and the minimal polynomial of α^5 is $x^2 + x + 1$. One way to find these is to note that every element of $\mathbb{F}_{16}$ is a root of $x^{16} - x$. If you factor this into irreducible polynomials, you need to check which irreducible has a given α^i as a root. Alternatively, you can find the generator polynomial by using the Maple worksheet *generator.mws*, which is available on the course website. In any case, we get $g(x) = x^{10} + x^8 + x^5 + x^4 + x^2 + x + 1$. This has degree 10, so $k = n - \deg(g(x)) = 5$. This code can correct three errors since the designated distance is 7.

Example 1.18. Consider the code built from the same field $\mathbb{F}_{16}$ but with designated distance 5. Then the code comes from the roots $\alpha, \alpha^2, \alpha^3, \alpha^4$. In this case, the generator polynomial is $x^8 + x^7 + x^6 + x^4 + 1$, and so $n = 15$ and $k = 15 - 8 = 7$. By lowering the designated distance, we have increased the dimension of the code.

2 Introduction to Algebraic Geometry

While the error correcting codes defined by Goppa can be described purely in terms of fields, they are better understood by also viewing them geometrically. We will discuss the basic concepts of algebraic geometry and see how the study of (nonsingular projective) algebraic curves is equivalent to the study of algebraic function fields in one variable. Furthermore, by thinking of fields, we will help to justify why we need to discuss projective curves instead of restricting only to affine curves. A brief treatment of the concepts of algebraic geometry needed for coding theory can be found in the book *Codes and Curves*, by Walker.

To do algebraic geometry we need to use algebraically closed fields. Recall that a field k is algebraically closed if every nonconstant polynomial in $k[x]$ has a root in k. The fundamental theorem of algebra states that $\mathbb{C}$ is algebraically closed. Neither $\mathbb{R}$ nor $\mathbb{Q}$ is algebraically closed as x^2+1 has no root in either field. Nor is any finite field F algebraically closed, for if $F = \{a_1, \ldots, a_n\}$, then $(x-a_1)\cdots(x-a_n)+1$ has no root in F. If F is an arbitrary field, then it has an algebraic extension that is algebraically closed; this field is called an algebraic closure of F. Such a field is unique up to isomorphism.

2.1 Affine Curves

Let k be an algebraically closed field. We denote the polynomial ring over k in 2 variables by $k[x,y]$. Algebraic geometry studies solutions to polynomial equations. We define the *affine place* $\mathbf{A}^2(k)$ to be the set k^2 of all pairs over k. If it is not important to keep track of the field k, we will write $\mathbf{A}^2$ in place of $\mathbf{A}^2(k)$. If $P = (a,b) \in \mathbf{A}^2$ and $f \in k[x,y]$, we will denote by $f(P)$ the evaluation of f at P. If $f \in k[x,y]$, then the *affine curve* $f = 0$ is defined to be

$$Z(f) = \left\{ P \in \mathbf{A}^2 : f(P) = 0 \right\}.$$

This set is sometimes called the zero set of f.

Example 2.1. The curve $y = x^2$ in $\mathbf{A}^2$ is a parabola. In general, a conic section is the zero set of a polynomial of degree 2. For instance, $x^2 + y^2 = 1$ is a circle and $xy = 1$ is a hyperbola. The curve $y^2 = x^3$ is sometimes called a cuspidal cubic curve.

Example 2.2. An elliptic curve is a curve given by an equation $y^2 = f(x)$, where $f(x)$ is a cubic polynomial with no repeated roots. For example, $y^2 = x^3 - x$ is an elliptic curve.

Example 2.3. A hyperelliptic curve is a curve given by an equation $y^2 = f(x)$, where $f(x)$ is a polynomial of degree at least 4 and with no repeated roots. For example, $y^2 = x^5 - 1$ is an elliptic curve over $\mathbb{C}$, or over any algebraically closed field of characteristic not 5.

Example 2.4. Suppose that $k = \mathbb{R}$, a field that is not algebraically closed. There are no solutions to the equation $x^2 + y^2 + 1 = 0$ over $\mathbb{R}$, so $Z(x^2+y^2+1)$ is empty. However, if $k = \mathbb{C}$, then there are solutions, including $(i, 0)$. It is to have solutions to polynomial

equations that the base field is assumed to be algebraically closed. If $f(x, y)$ is a polynomial over an algebraically closed field k, and if $b \in k$, then $f(x, b)$ is a polynomial in the one variable x. If $f(x, b)$ is not constant, then it has roots in k, and so $f(x, y) = 0$ has solutions in $\mathbf{A}^2(k)$.

If X is an algebraic curve over k, then the ideal of X is

$$I(X) = \{f \in k[x, y] : f(P) = 0 \text{ for all } P \in X\}.$$

If $f(x, y) = p(x, y)^{e_1} \cdots p_n(x, y)^{e_n}$ is the factorization of a polynomial f into irreducible factors, then it is easy to see that $I(Z(f)) = (p_1 \cdots p_n)$. Moreover, the *coordinate ring* of X is the quotient ring $\Gamma(X) = k[x, y]/I(X)$. One way of thinking about the coordinate ring is to consider it the ring of polynomial functions on X. For, two polynomials f and g induce the same function $X \to k$ precisely when $f - g \in I(X)$, which is equivalent to the cosets $\overline{f}$ and $\overline{g}$ being equal.

A topological space is said to be *irreducible* if it cannot be written as the union of two proper subcurves. If $f = gh$, then $Z(f) = Z(g) \cup Z(h)$. From this, if f is a squarefree polynomial, it follows that $Z(f)$ is irreducible if and only if f is irreducible. Alternatively, X is irreducible if and only if $\Gamma(X)$ is an integral domain.

2.2 Projective Varieties

For students who have studied topology, the construction of projective varieties parallels that of constructing the real projective plane. We define an equivalence relation $\sim$ on $\mathbf{A}^3 \setminus \{(0, 0, 0)\}$ by defining that $(a, b, c) \sim (a', b', c')$ if there is a nonzero scalar λ with $a' = \lambda a$, $b' = \lambda b$, and $c' = \lambda c$. Geometrically, the equivalence class of a point is the line through the origin that passes through the point. We will write $(a : b : c)$ for the equivalence class of (a, b, c), and $\mathbf{P}^2$ will denote the set of all equivalence classes. This is the *projective plane*. Note that $(a : b : c)$ represents a point in the projective plane only if at least one of the coordinates is nonzero.

Note that polynomial functions are not well defined on points of projective space. However, we can get around this problem. A polynomial $f \in k[x, y, z]$ is said to be *homogeneous* if every monomial of f has the same degree. Alternatively, f is homogeneous if there is an integer m with $f(\lambda x, \lambda y, \lambda z) = \lambda^m f(x, y, z)$ for all $\lambda \in k$. If f is homogeneous and $P \in \mathbf{P}^2$, then the equation $f(P) = 0$ is well defined; in other words, if $P \sim Q$, then $f(P) = 0$ if and only if $f(Q) = 0$. Therefore, we can define zero sets of collections of homogeneous polynomials. If f is a homogeneous polynomial, then the *projective curve* $f = 0$ is the zero set

$$Z(f) = \{P \in \mathbf{P}^n : f(P) = 0\}.$$

We define a projective curve to be irreducible in exactly the same way as for affine curves; if f is a squarefree homogeneous polynomial, then it follows that $Z(f)$ is irreducible if and

only if f is irreducible.

We can define a coordinate ring for projective curves. If X is a projective curve, then $I(X)$ is defined to be

$$I(X) = \langle \{ f \in k[x,y,z] : f \text{ is homogeneous and } f(P) = 0 \text{ for all } P \in X \} \rangle .$$

The homogeneous coordinate ring $S(X)$ is the quotient ring $k[x,y,z]/I(X)$. As with affine curves, a projective curve is irreducible if and only if its ideal is a prime ideal, and this happens exactly when $S(X)$ is an integral domain.

Example 2.5. Consider the projective parabola X given by $yz = x^2$. If $(a : b : c) \in X$, then $bc = a^2$. If $c \neq 0$, then by dividing by c, we may assume that $c = 1$. Then, (a, b) is on the affine parabola $y = x^2$. On the other hand, if $c = 0$, then $a = 0$, and $b \neq 0$ for $(a : b : c)$ to be a valid point. Then, by dividing by b, we have the point $(0 : 1 : 0)$. Thus, we can think of X as the union of the affine parabola $y = x^2$ and the extra point $(0 : 1 : 0)$. Note that the affine parabola is obtained by setting $z = 1$ in the equation $yz = x^2$. Moreover, there is another affine parabola inside X, and this other curve contains $(0 : 1 : 0)$. If we set $y = 1$, then we have the equation $z = x^2$, which again represents a parabola. The point $(0 : 1 : 0)$ is a point of this affine parabola. The purpose of considering this second affine parabola is that any point of X is a point on some affine curve in X.

Example 2.6. Let X be the projective elliptic curve $y^2z = x^3 - xz^2$. The affine elliptic curve $y^2 = x^3 - x$ is an affine curve in X; we can map it into X via $(a, b) \mapsto (a : b : 1)$. The image is $X - \{(0 : 1 : 0)\}$; if $(a : b : c) \in X$ and $c = 0$, then the equation $y^2z = x^3 - xz^2$ forces $a = 0$. Then $b \neq 0$ since $(a : b : c) \in \mathbf{P}^2$. Thus, we may divide by b to assume $b = 1$, and so the only point on X for which $c = 0$ is $(0 : 1 : 0)$. As with the previous example, the affine equation is obtained from the projective equation by setting $z = 1$.

Example 2.7. In the previous examples we started with a projective curve and found an affine curve inside it. In this example we start with an affine curve and produce a projective curve. Let Y be the curve $y = x^3$. By adding a third variable z, we can use $y - x^3$ to get the homogeneous polynomial $yz^2 - x^3$. The affine curve Y sits inside the projective curve $yz^2 = x^3$ as in the previous examples. Similarly, $y^2 + 1 = x^4$ yields the homogeneous polynomial $y^2z^2 + z^4 = x^4$. What we are doing is adding enough copies of z to each monomial so that each has degree equal to the highest degree of a monomial of the original polynomial. To have a formula for doing this, if $f(x, y)$ has degree d, then $z^d f(x/z, y/z)$ is the *homogenization* of $f(x, y)$. The resulting projective curve is called the projective closure of the affine curve $f = 0$. In general, if $Z(f)$ is a projective curve, then any point of the form $(a : b : 0)$ of $Z(f)$ is called a point at infinity. If $g(x, y) = f(x, y, 1)$, then $Z(f)$ is the union of the affine curve $g = 0$ and the points of infinity. The polynomial $g(x, y)$ is called a dehomogenization of $f(x, y, z)$ (at the variable z).

Note that if $f^h(x,y,z)$ is the homogenization of $f(x,y)$, then $f(x,y) = f^h(x,y,1)$. Conversely, if $g(x,y,z)$ is homogeneous of degree d, and if $f(x,y) = g(x,y,1)$, then

$$g(x,y,z) = z^{\deg(g)-\deg(f)} f^h(x,y).$$

We do not, in general, recover $g(x,y,z)$ by homogenizing $g(x,y,1)$ since this polynomial may have smaller degree than the degree of $g(x,y,z)$. For example, if $g(x,y,z) = z^2 + xz + yz$, then $g(x,y,1) = 1 + x + y$ has degree 1 while $g(x,y,z)$ has degree 2.

2.3 The Function Field of a Curve

Suppose that an affine curve X is irreducible. Then its coordinate ring $\Gamma(X)$ is an integral domain, and so has a quotient field, which we denote by $k(X)$. This is the field

$$k(X) = \{g/h : g, h \in \Gamma(X), h \neq 0\}.$$

Let $\overline{x}$ and $\overline{y}$ be the images of x and y in $\Gamma(X)$. The definition of $\Gamma(X)$ shows that $\Gamma(X) = k[\overline{x}, \overline{y}]$. That is, $\Gamma(X)$ is generated as a k-algebra by $\overline{x}$ and $\overline{y}$. The field $k(X)$ then contains the field $k(\overline{x}, \overline{y})$ generated by $\overline{x}$ and $\overline{y}$. Since the quotient field of an integral domain is the smallest field containing the domain, we see that $k(X) = k(\overline{x}, \overline{y})$. Therefore, $k(X)$ is the field extension of k generated by two elements u, v, and subject to the relation $f(u,v) = 0$.

To give other another view of the function field, suppose that X is the zero set of the irreducible polynomial $f(x,y)$. If we define an equivalence relation $\sim$ on the ring

$$S = \{g(x,y)/h(x,y) : g, h \in k[x,y], h \notin (f)\}$$

by $g/h \sim g'/h'$ if $gh' - g'h \in (f) \subseteq k[x,y]$, then $k(X) \cong S/\sim$. To verify this interpretation, note that an arbitrary element of $k(X)$ is of the form $\overline{g}/\overline{h}$ with $g, h \in k[x,y]$, where we continue to write bars to represent the image of a polynomial in $\Gamma(X)$. When are $\overline{g}/\overline{h}$ and $\overline{g'}/\overline{h'}$ equal in $k(X)$? Since $k(X)$ is the quotient field of $\Gamma(X)$, this occurs exactly when $\overline{g}\overline{h'} = \overline{g'}\overline{h}$. Finally, this occurs when $\overline{gh' - g'h} = \overline{0}$, or $gh' - g'h \in (f)$.

We may define a topology on a curve by defining a set to be open if it is empty, or if its complement is finite. Thus, proper closed sets are finite. This is a special example of the Zariski topology on an algebraic variety; curves are special examples of varieties. We will view $k(X)$ as the field of rational functions defined an open subset of the curve X; a rational function is simply a function that can be represented as a quotient of polynomials.

Example 2.8. The function field of the affine parabola $y = x^2$ is the quotient field of $k[x,y]/(y - x^2)$. This ring is isomorphic to $k[t]$, the rational function field in one variable t; they are isomorphic via the map that sends t to $x + (y - x^2)$. Therefore, the function field is isomorphic to $k(t)$.

Example 2.9. The coordinate ring of the curve $X = Z(y^2 - x^3)$ is $k[x,y]/(y^2 - x^3)$. This ring

is isomorphic to $k[t^2, t^3]$, a subring of $k[t]$. These rings are isomorphic via the map $\overline{x} \mapsto t^2$ and $\overline{y} \mapsto t^3$. The function field of this curve is then the quotient field of $k[t^2, t^3]$, which is $k(t)$. Note that y/x represents a rational function on X as does x^2/y. Furthermore, these functions agree since $y/x = x^2/y$ in $k(X)$. Therefore, a rational function can be represented in more than one way as a quotient of polynomials.

Example 2.10. The coordinate ring of the elliptic curve $y^2 = x^3 - x$ is $k[x, y]/(y^2 - x^3 - x)$. We claim that its function field is isomorphic to $k(t)\left(\sqrt{t^3 - t}\right)$. To see this, note that the coordinate ring is $k[\overline{x}, \overline{y}]$, and so its function field is $k(\overline{x}, \overline{y})$. That is, the function field is generated as an extension field of k by $\overline{x}$ and $\overline{y}$. Now, we have $\overline{y}^2 = \overline{x}^3 - \overline{x}$. Therefore $\overline{y}$ is algebraic over $k(\overline{x})$, and so $k(\overline{x}, \overline{y}) = k(\overline{x})(\overline{y}) \cong k(\overline{x})\left(\sqrt{\overline{x}^3 - \overline{x}}\right)$. The element $\overline{x}$ cannot be algebraic over k, so $k(\overline{x}) \cong k(t)$, the rational function field in t.

Let $f(x, y, z)$ be an irreducible homogeneous polynomial, and let $X = Z(f)$. We can also define a function field of X. Let $\sim$ be the equivalence relation on the set

$$\{g(x, y, z)/h(x, y, z) : g, h \in k[x, y, z] \text{ are homogeneous of the same degree, } h \notin (f)\} \cup \{0\}$$

given by $g/h \sim g'/h'$ if and only if $gh' - g'h \in (f)$. The function field $k(X)$ is then the set of equivalence classes, under the natural operations.

Note that any $f/g \in k(X)$ determines a well defined function on an open subset of X: if g, h are both homogeneous of degree d, then if λ is nonzero, then

$$\frac{g(\lambda a, \lambda b, \lambda c)}{h(\lambda a, \lambda b, \lambda c)} = \frac{\lambda^d g(a, b, c)}{\lambda^d h(a, b, c)} = \frac{g(a, b, c)}{h(a, b, c)}.$$

The function g/h is defined for all points P except when $h(P) = 0$. We may then think of $k(X)$ in much the same way as we think of the function field of an affine curve.

We now show that the function field of a projective curve can be calculated by finding the function field of an affine curve.

Proposition 2.11. *Let $f(x, y, z)$ be an irreducible homogeneous polynomial. Let $X = Z(f)$ be a projective curve and if U is the affine curve $Z(f(x, y, 1))$, then $k(X) \cong k(U)$.*

Proof. We will define a ring homomorphism $\varphi : \Gamma(U) \to k(X)$ and show that φ is injective. Thus, φ induces a homomorphism $k(U) \to k(X)$, which necessarily is injective. We will then be done by showing that this map is surjective. To define φ, write $g(x, y) = f(x, y, 1)$. For $h(x, y) \in k[x, y]$, define $\varphi(\overline{h}) = h(x/z, y/z)$. Note that if $\deg(h) = d$, then $h(x/z, y/z) = h^h(x, y, z)/z^d$, and this represents an element of $k(X)$. This map is well defined, since if $h' - h \in (g)$, then $h' = h + gl$ for some $l \in k[x, y]$. Then $h'(x/z, y/z) = h(x/z, y/z) + g(x/z, y/z)l(x/z, y/z)$. By our hypothesis on f, we have $g(x/z, y/z) = f(x, y, z)/z^{\deg(f)}$. Therefore, $g(x/z, y/z)l(x/z, y/z) = 0$ in $k(X)$ by the definition of $k(X)$. Thus, $\varphi(\overline{h}) = \varphi(\overline{h'})$; this proves that φ is well defined. It is a simple exercise to see that φ is a ring

homomorphism. To prove injectivity, suppose that $\varphi(h) = 0$. Then $h(x/z, y/z) = 0$ in $k(X)$. Note that $h(x/z, y/z) = h^h(x, y, z)/z^{\deg(h)}$. For this to be zero in $k(X)$, we have $h^h(x, y, z) = f(x, y, z)m(x, y, z)$ for some m. Dehomogenizing, we get

$$\begin{aligned} h(x, y, z) &= h^h(x, y, 1) = f(x, y, 1)m(x, y, 1) \\ &= g(x, y)m(x, y, 1). \end{aligned}$$

Therefore, $\overline{h} = 0$. Thus, φ is injective. For surjectivity, every element of $k(X)$ is represented by a quotient $h(x, y, z)/l(x, y, z)$ of homogeneous polynomials of the same degree. Let their common degree be d. Then $h(x, y, z)/l(x, y, z) = z^d h(x, y, 1)/z^d l(x, y, 1) = h(x, y, 1)/l(x, y, 1)$. This represents an element of $k(U)$, and this element maps via φ to $h(x, y, z)/l(x, y, z)$. Therefore, φ is surjective, and so $k(U) \cong k(Z)$. □

Example 2.12. Let X be the projective parabola $yz = x^2$. The affine parabola $Y = Z(y - x^2)$ is obtained from dehomogenizing $yz - x^2$. Therefore, $k(X) = k(Y) \cong k(t)$. Similarly, the function field of $yz^2 = x^3$ is $k(t)$, and the function field of the projective elliptic curve $y^2z = x^3 - xz^2$ is $k(t)\left(\sqrt{t^3 - t}\right)$.

There are important subrings of the function field of a curve. First, if X is a projective or affine curve, then

$$k[X] = \{\varphi \in k(X) : \varphi \text{ is defined at every point of } X\}.$$

This is called the *ring of regular functions* on X, and is the ring of globally defined rational functions on X. If X is affine, then one can show that $k[X] \cong \Gamma(X)$, but that if X is projective, then $k[X] = k$. Next, let $P \in X$. Then the *local ring of X at P* is the set

$$\mathcal{O}_P(X) = \{\varphi \in k(X) : \varphi \text{ is defined at } P\}.$$

This is the ring of regular functions defined locally at P. For $\varphi(P)$ to be defined, φ must be defined in an open neighborhood of P, since if $\varphi = g/h$, then $P \notin Z(h)$, and so φ is defined on the open neighborhood $h \neq 0$ of P. For convenience, we will typically write $\mathcal{O}_P$ in place of $\mathcal{O}_P(X)$. It is a local ring; its unique maximal ideal is

$$M_P = \{\varphi \in \mathcal{O}_P : \varphi(P) = 0\}.$$

A short exercise shows that M_P is an ideal of $\mathcal{O}_P$. Furthermore, if $\varphi \in \mathcal{O}_P \setminus M_P$, then we may write $\varphi = f/g$ with $f(P) \neq 0$. Then $g/f \in \mathcal{O}_P$ is an inverse for φ. Therefore, since every element outside M_P is a unit, M_P is the unique maximal ideal of $\mathcal{O}_P$. If $Y = Z(f)$ is an affine curve, then $\mathcal{O}_P(Y) = \left\{\overline{g}/\overline{h} \in k(Y) : h(P) \neq 0\right\}$.

Proposition 2.13. *Let P be a point of an irreducible affine curve $X = Z(f)$, and let $\mathfrak{m}_P = \{g \in k[X] : g(P) = 0\}$. Then $\mathcal{O}_P = k[X]_{\mathfrak{m}_P}$.*

Proof. By definition, $k[X] \subseteq \mathcal{O}_P$. Moreover, everything in $k[X] \setminus \mathfrak{m}_P$ is invertible in $\mathcal{O}_P$, so $k[X]_{\mathfrak{m}_P} \subseteq \mathcal{O}_P$. Conversely, let $\varphi \in \mathcal{O}_P$. Then we may write $\varphi = g/h$ with $g, h \in k[X]$, by the definition of $k(X)$. Since φ is defined at P, we have $h(P) \neq 0$. Therefore, $h \notin \mathfrak{m}_P$, so $g/h \in k[X]_{\mathfrak{m}_P}$. Thus, $\mathcal{O}_P = k[X]_{\mathfrak{m}_P}$. □

It is not hard to show that if $P = (a, b)$, then $\mathfrak{m}_P$ is generated by the images in $k[X]$ of the polynomials $x - a$ and $y - b$.

Let $X = Z(f)$ be an irreducible projective curve, and let $Y = Z(g)$ be the affine curve obtained by dehomogenizing f. We have seen that $k(X) = k(Y)$. If $P \in Y$, the definition of local ring then shows that $\mathcal{O}_P(X) = \mathcal{O}_P(Y)$. Therefore, the local ring at a point can be determined from the previous proposition.

Example 2.14. The line $X = Z(y)$ has function field $k(t)$. We determine the local rings of points on X. Let $P = (a, 0) \in X$. Note that $\Gamma(X) = k[x, y]/(y) \cong k[t]$, and $k(X)$ is the quotient field of this ring. The isomorphism $k(X) \cong k(t)$ sends the image of a polynomial $f(x, y)$ to $f(t, 0)$. We claim that $\mathcal{O}_P(X) = k[t]_{(t-a)}$, the localization of $k[t]$ at the maximal ideal $(t - a)$. For, $g(x, y)/h(x, y) \in \mathcal{O}_P(X)$ if $h(a, 0) \neq 0$. Therefore, $h(t, 0)$ is not divisible by $t - a$, and so $g(t, 0)/h(t, 0) \in k[t]_{(t-a)}$. The reverse inclusion is easy since $g(t, 0)/h(t, 0) \in k[t]_{(t-a)}$ is defined at P since $t - a$ does not divide $h(t, 0)$.

2.4 Nonsingular Curves

To have a complete correspondence between function fields and projective curves, we must restrict to nonsingular curves. Suppose that $f(x, y, z)$ is a homogeneous polynomial. It defines a projective curve X in $\mathbf{P}^2$, and we refer to it as a plane curve. By recalling formulas for tangent lines from calculus, we say that the curve X is nonsingular at a point $P \in X$ if at least one of the three partial derivatives $\partial f/\partial x$, $\partial f/\partial y$, or $\partial f/\partial z$ do not vanish at P. If X is nonsingular at every point P, then we say that X is a *nonsingular curve*. Similarly, an affine curve $g = 0$ is nonsingular at a point P if $\partial g/\partial x$ and $\partial g/\partial y$ do not both vanish at P, and the curve itself is nonsingular if it is nonsingular at all its points.

Example 2.15. The affine parabola $y = x^2$ is nonsingular, for the partial derivatives of $y - x^2$ are $-2x$ and 1; neither vanish simultaneously. Similarly, $Z(y - x^3)$ and $Z(x^2 + y^2 - 1)$ are nonsingular. However, $Z(y^2 - x^3)$ is singular at the origin.

Example 2.16. Consider the projective parabola $yz = x^2$. The partial derivatives are $-2x$, z, y; they only simultaneously vanish at $(0, 0, 0)$, which does not represent a point in $\mathbf{P}^2$. Therefore, the parabola is nonsingular. However, the projective closure of the cubic curve $y = x^3$ is $X = Z(yz^2 - x^3)$. Here, the partial derivatives are $-3x^2$, z^2, $2yz$, which vanishes at $(0 : 1 : 0) \in X$. Therefore, this curve has a singularity. Note that the singularity occurs at the point at infinity.

Example 2.17. The elliptic curve $X = Z(y^2z - x^3 + xz^2)$ is nonsingular if the characteristic of k is not 2. To see this, the three partial derivatives of $y^2z - x^3 + xz^2$ are $-3x^2 + z^2$, $2yz$, and $y^2 + 2xz$. It is easy to see that all three partials vanish only at $x = y = z = 0$, and so there is no point on X for which this happens. In fact, if $y^2 = f(x)$ is any elliptic curve, where $f(x)$ is a cubic polynomial with no repeated roots, then this curve is nonsingular. For, a point $P = (a, b)$ on the curve satisfies $b^2 = f(a)$. The partials of $y^2 - f(x)$ are $-f'(x)$ and $2y$; for these to vanish at P, we must have $b = 0$ and $f'(a) = 0$. Then $f(a) = f'(a) = 0$, and this cannot happen since a would then be a multiple root of $f(x)$.

Example 2.18. Any hyperelliptic curve is nonsingular, as long as the characteristic of the base field is not 2. The argument of the previous example works word for word to prove that a hyperelliptic curve is nonsingular.

By making use of some theorems from Math 582, we can now begin to give the connection between function fields and curves. We first note that a commutative ring is said to have dimension 1 if all nonzero prime ideals are maximal. It is not an obvious result, but the coordinate ring of an affine curve has dimension 1. Also, the local ring at a point on any curve also has dimension 1. The general result is that the geometric dimension of an affine algebraic variety is equal to the ring theoretic dimension of its coordinate ring, and that these are also equal to the transcendence degree over k of the function field $k(X)$. Next, if X is a curve, then $\mathcal{O}_P(X)$ is a local Noetherian domain of dimension 1 for any $P \in X$. We have already remarked that its dimension is 1. It is Noetherian because the coordinate ring of an affine curve is a quotient ring of $k[x, y]$. The coordinate ring is then Noetherian, and since $\mathcal{O}_P(X)$ is a localization of the coordinate ring, it is also Noetherian. We have seen above that $\mathcal{O}_P(X)$ is a local ring. Let A be a local domain with maximal ideal M, and set $F = A/M$, a field. Then A is a discrete valuation ring if and only if it is Noetherian, dimension 1, and $\dim_F(M/M^2) = 1$. Note that $F = k$ for $A = \mathcal{O}_P(X)$; this can be seen indirectly in the proof below.

Theorem 2.19. *Let $X = Z(f)$ be an irreducible curve. Then $P \in X$ is nonsingular if and only if $\mathcal{O}_P(X)$ is a discrete valuation ring of $k(X)$.*

Proof. By the remarks above, it is enough to prove that P is nonsingular if and only if $\dim_k(M_P/M_P^2) = 1$, where M_P is the maximal ideal of $\mathcal{O}_P$. Let $M = (x - a, y - b)$, a maximal ideal of $k[x, y]$. Define a map $\theta : M \to k^2$ by

$$\theta(g) = (\frac{\partial g}{\partial x}(P), \frac{\partial g}{\partial y}(P)).$$

We note that θ yields a vector space isomorphism $M/M^2 \cong k^2$. The polynomial f yields the 1-dimensional subspace $V = ((f) + M^2)/M^2$ of M/M^2, and $\theta(V)$ has dimension 1 if and only if P is nonsingular, and it has dimension 0 otherwise. Recall that $k[X] \cong k[x, y]/(f)$

and $\mathcal{O}_P = k[X]_{\mathfrak{m}_p}$, where $\mathfrak{m}_P = \{g \in k[X] : g(P) = 0\} = M/(f)$. Then

$$\mathfrak{m}_P/\mathfrak{m}_P^2 = M/(f)/\left(M^2 + (f)/\right)(f) \cong M/\left(M^2 + (f)\right)$$

The maximal ideal M_P of $\mathcal{O}_P$ is then $\mathfrak{m}_P\mathcal{O}_P$, and so $M_P/M_P^2 \cong \mathfrak{m}_P/\mathfrak{m}_P^2$ by a Math 582 exercise. Thus,

$$M_P/M_P^2 \cong \mathfrak{m}_P/\mathfrak{m}_P^2 \cong M/((f) + M^2) \cong \left(M/M^2\right)/V$$

by one of the homomorphism theorems of vector space theory. Therefore, M_P/M_P^2 has dimension 1 if and only if P is nonsingular. □

Example 2.20. The curve $X = Z(y^2 - x^3)$ is singular at the origin. The coordinate ring of X is $k[t^2, t^3]$, and so the local ring at the origin is $k[t^2, t^3]_{(t^2,t^3)}$. This is a proper subring of the discrete valuation ring $k[t]_{(t)}$. By another characterization of discrete valuation rings, a local Noetherian domain of dimension 1 is a discrete valuation ring if and only if it is integrally closed. The ring $k[t^2, t^3]_{(t^2,t^3)}$ is not integrally closed since t is integral over it; t satisfies the polynomial $T^2 - t^2$. However, $t \notin k[t^2, t^3]_{(t^2,t^3)}$, which shows that this ring is not integrally closed.

2.5 Curves over non-Algebraically Closed Fields

Classical algebraic geometry uses algebraically closed fields in order to have solutions to polynomial equations. However, in several situations, including coding theory, we need to work with arbitrary fields.

Let k be an arbitrary field, and let K be an algebraically closed extension field of k. An affine curve over k is a curve $Z(f) \subseteq \mathbf{A}^2(K)$ such that $f(x, y) \in k[x, y]$. Similarly, a projective curve over k is a curve $Z(f) \subseteq \mathbf{P}^2(K)$ such that $f(x, y, z) \in k[x, y, z]$. An affine curve $Z(f)$ over k is said to be *absolutely irreducible* if $f(x, y)$ is irreducible in $K[x, y]$. An analogous definition holds for projective curves.

If $X = Z(f)$ is an affine curve over k, then its ideal over k is

$$I(X/k) = I(X) \cap k[x, y] = \{g(x, y) \in k[x, y] : g(P) = 0 \text{ for all } P \in X\}$$

and its k-coordinate ring is

$$\Gamma(X/k) = k[x, y]/I(X/k).$$

This is a subring of the coordinate ring $\Gamma(X)$. If X is absolutely irreducible, then the function field $K(X)$ exists, we define the k-function field $k(X)$ to be the quotient field of $\Gamma(X/k)$. Therefore, we see that $k(X)$ is the set of all quotients of polynomials $g(x, y)/h(x, y)$, where $g/h = g'/h'$ if $gh' - g'h \in fk[x, y]$. We similarly can define the k-function field of an irreducible projective curve. We also can define a k-version of the local ring of a point,

which we also denote by $\mathcal{O}_P(X/k)$. This is the ring

$$\begin{aligned}\mathcal{O}_P(X/k) &= \{\varphi \in k(X) : \varphi(P) \text{ is defined}\} \\ &= \mathcal{O}_P(X) \cap k(X).\end{aligned}$$

This notation does not show the dependence on k, but this will not give us a problem since we will not consider this concept for two fields at a time.

The characterization of nonsingularity in terms of $\mathcal{O}_P$ can be made relative to k as follows. If X is a curve over k, then $P \in X$ is nonsingular if and only if $\mathcal{O}_P(X/k)$ is a discrete valuation ring. The proof of this is a moderately standard use of facts about discrete valuation rings, and we won't give it here.

A point $P = (a, b)$ on a curve X is said to be a *k-rational point* if $a, b \in k$. We write $X(k)$ for the set of all k-rational points of X. Thus, $X = X(K)$. To simplify terminology, we will often refer to a K-rational point as simply a point, and use the terminology rational point to note a point whose coordinates are in k (or some other subfield of K). A curve over k is nonsingular if it is nonsingular as a curve over K. That is, X is nonsingular if every point $P \in \mathbf{A}^2(K)$ lying on X is nonsingular.

As we will see, a major problem in working with codes coming from algebraic geometry is to find curves over $\mathbb{F}_q$ with lots of rational points.

Example 2.21. Let $n \geq 3$ be an integer, and consider the projective curve $x^n + y^n = z^n$. Fermat's last theorem says that there are no $\mathbb{Q}$-rational points of this curve having all nonzero coordinates.

Example 2.22. The projective curve $X = Z(x^2 + y^2 + z^2)$ over $\mathbb{R}$ has no $\mathbb{R}$-rational points. However, it has lots of $\mathbb{C}$-rational points, such as $(1 : 0 : i)$.

Example 2.23. The projective curve $X = Z(x^2 + y^2)$ over $\mathbb{R}$ is not absolutely irreducible even though $x^2 + y^2$ is irreducible in $\mathbb{R}[x, y]$, since over $\mathbb{C}$ the polynomial $x^2 + y^2$ factors as $(x + iy)(x - iy)$. Therefore, the k-curve $Z(f)$ need not be absolutely irreducible even if f is irreducible over k.

Example 2.24. We will see near the end of the course that Weil's proof of the Riemann hypothesis for curves over a finite field $\mathbb{F}_q$ yields a bound on the number of rational points. This bound, which is in terms of the *genus* g of the curve and of the size q of the base field, says that the number N of rational points satisfies the inequality

$$|N - (q + 1)| \leq 2g\sqrt{q}.$$

We will define the genus in the first chapter of Stichtenoth; the proof of this bound is done in Chapter 5. Let X be the curve over $\mathbb{F}_{q^2}$ given by $x^{q+1} + y^{q+1} = 1$. This is called the Hermitian curve over $\mathbb{F}_{q^2}$. It turns out that the genus of this curve is $g = q(q - 1)/2$, and that $N = 1 + q^3$, the largest possible given the bound above.

3 Algebraic Function Fields and Discrete Valuation Rings

Let F/k be a field extension. Then F is said to be an *algebraic function field in one variable* over k if there is an element $x \in F$ that is transcendental over k and for which $[F : k(x)] < \infty$. Recall that if x is transcendental over k, then the field $k(x)$ is isomorphic to a rational function field in one variable over k. We will not distinguish between transcendental elements and variables from now on. These fields are precisely the function fields of algebraic curves defined over k. We have indicated, without full proof, why the function field of a curve is such a field. For the converse, if F is generated over $k(x)$ by a single element y, then since y is algebraic over $k(x)$, there is an irreducible polynomial $p(t) \in k(x)[t]$ for which $p(y) = 0$. By clearing denominators, we may view $p(t) \in k[x][t]$, and the equation $p(y) = 0$ yields an equation $f(x, y) = 0$ with f a polynomial in two variables over k. Then F is the function field of $Z(f) \subseteq \mathbf{A}^2$.

We note one simple but important property of algebraic function fields in one variable, which we state as a proposition.

Proposition 3.1. *Let F/k be an algebraic function field in one variable. If $t \in F$ is transcendental over k, then $[F : k(t)] < \infty$.*

Proof. There is an $x \in F$ transcendental over k and with $[F : k(x)] < \infty$. We will be done by proving that x is algebraic over $k(t)$, since then $[k(t)(x) : k(t)] < \infty$, and then

$$\begin{aligned}[F : k(t)] &= [F : k(t)(x)][k(t)(x) : k(t)] \\ &\leq [F : k(x)][k(t)(x) : k(t)]\end{aligned}$$

is finite since both terms are finite. To show that x is algebraic over $k(t)$, we know that t is algebraic over $k(x)$. Therefore, there are $f_i(x) \in k(x)$ with $t^n + f_{n-1}(x)t^{n-1} + \cdots + f_0(x) = 0$. By writing each $f_i(x)$ as a quotient of polynomials and then clearing denominators, we have an equation of the form $a_n(x)t^n + a_{n-1}(x)t^{n-1} + \cdots + a_0(x) = 0$. Viewing this equation as a polynomial in x with coefficients in $k(t)$, we conclude that x is algebraic over $k(t)$. □

This proposition could also be proved by using the notion of a transcendence basis. A (finite) transcendence basis of a field extension K/k is a set $\{t_1, \ldots, t_n\}$ of elements of K such that t_1 is transcendental over k, and t_{i+1} is transcendental over $k(t_1, \ldots, t_i)$ for each $i \geq 1$, and for which K is algebraic over $k(t_1, \ldots, t_n)$. The number of elements of a transcendence basis is uniquely determined, and is called the transcendence degree of K/k. For an algebraic function field in one variable F/k, the transcendence degree is 1. If t is transcendental over k, then $\{t\}$ is a transcendence basis for F/k by the uniqueness of transcendence degree. Therefore, every element of F is algebraic over $k(t)$, including the element x of the proof of the proposition. Moreover, F is finitely generated over k, so it is also finitely generated over $k(x)$; thus, as F is algebraic over $k(x)$, we see that $[F : k(x)] < \infty$.

Let F/k be an algebraic function field in one variable. If k' is the algebraic closure of k in F, then k' is called the *field of constants* of F/k. Recall that k' is the set of all elements of F that are algebraic over k. We will see the reason for calling elements of k' constants once we interpret elements of F as functions. Now we prove that k' is a finite dimensional extension of k. This is a special case of a theorem from field theory says that if F/k is a finitely generated field extension, then the algebraic closure of k in F is a finite extension of k.

Proposition 3.2. *Let F/k be an algebraic function field in one variable. If k' is the field of constants of F/k, then $[k' : k] < \infty$.*

Proof. Let $x \in F$ be transcendental over k. Then $[F : k(x)] < \infty$ by the previous result. We claim that $[k' : k] \leq [F : k(x)]$. Once we prove this, then we will have proved the proposition. Let $\{t_1, \ldots, t_n\} \subseteq k'$ be a linearly independent set over k. We claim it is also linearly independent over $k(x)$. For, if there is an equation $\sum_i t_i f_i(x) = 0$ with $f_i(x) \in k(x)$, by clearing denominators, we may assume that $f_i(x) \in k[x]$. Write $f_i(x) = \sum_j a_{ij} x^j$ with $a_{ij} \in k$. Then $\sum_{i,j} t_i a_{ij} x^j = 0$, which we may write as $\sum_j \left(\sum_i a_{ij} t_i\right) x^j = 0$. This would show that x is algebraic over k' unless all coefficients are 0, a contradiction since k'/k is algebraic and x is transcendental over k. So, $\sum_i a_{ij} t_i = 0$ for each j. Since $a_{ij} \in k$ and the t_i are independent over k, each $a_{ij} = 0$. This shows that each $f_i(x) = 0$, and so the claim that $\{t_1, \ldots, t_n\}$ is independent over $k(x)$. Therefore, $n \leq [F : k(x)]$. This forces $[k' : k] \leq [F : k(x)]$, as desired. □

If $f(x, y) \in k[x, y]$ is an absolutely irreducible polynomial and $X = Z(f)$ is the corresponding curve over k, then k is the field of constants of $k(X)/k$. We probably will not use this result, and we will not prove it. A proof can be found in Section 22 of my book *Field and Galois Theory*. That section studies function fields of affine algebraic varieties, not just algebraic curves.

3.1 Discrete Valuation Rings

We note that the rational function field $k(x)$ itself is the most simple example of an algebraic function field. One special property for $k(x)$ is that, since this field is the quotient field of the unique factorization domain $k[x]$, every element $\varphi(x)$ of $k(x)$ can be uniquely represented in the form

$$\varphi(x) = \frac{p_1(x)^{e_1} \cdots p_n(x)^{e_n}}{q_1(x)^{f_1} \cdots q_m(x)^{f_m}},$$

where $p_i, \ldots, p_n, q_1, \ldots q_m$ are distinct irreducible polynomials, n and m are positive integers, and each exponent e_i and f_i is nonnegative. We note that we may need to use 0 exponents to write an element in this form. The element φ defines a function to k whose domain is a subset of k; we may talk of zeros and poles of φ: a zero of φ is obviously an element a with $\varphi(a) = 0$. A pole of φ is an element b for which the denominator of φ is 0. Such an

element is a point at which φ is not defined. If b is a pole, then the denominator of φ can be written in the form $(x-b)^r\sigma(x)$ for some polynomial σ for which b is not a root, and for some integer $r \geq 1$. We then see that, while $\varphi(x)$ is not defined at b, the rational function $(x-b)^r\varphi(x)$ is both defined and nonzero at b. We then define the order of the pole b to be r. Similarly, a zero a of φ has order s if $\varphi(x) = (x-a)^s\tau(x)$, where $\tau(x)$ is a rational function defined at a and for which $\tau(a) \neq 0$.

The problem of Riemann is to determine, given points $P_1, \ldots, P_n, Q_1, \ldots, Q_m$ of a curve X and positive integers $e_1, \ldots, e_n, f_1, \ldots, f_m$, which functions $\varphi \in k(X)$ have a zero at P_i of order at least e_i and a pole at Q_i of order at most f_i. While this problem is well defined for curves whose function field is $k(x)$, it is not well defined for other curves. One point of the first chapter of Stichtenoth is to define what is a zero and a pole of a rational function on a curve. For this we need to discuss in some detail the concept of a discrete valuation ring. Before giving the definitions, we give two examples, one from number theory and the other from algebraic geometry.

Example 3.3. Let p be a prime number. We can write every rational number in the form $p^n\frac{a}{b}$, where $n \in \mathbb{Z}$ and where a and b are integers neither divisible by p. The exponent n is uniquely determined, this is a consequence of unique factorization. We define a function $v_p : \mathbb{Q}^* \to \mathbb{Z}$ by $v_p(x) = n$ if $x = p^n\frac{a}{b}$, as above. We point out two properties of this function. Let $x, y \in \mathbb{Q}^*$. Then $v_p(xy) = v_p(x) + v_p(y)$. This follows from the laws of exponents and from the definition of a prime, that implies that if u, v are not divisible by p, then neither is uv. Second, if $x + y \neq 0$, then $v_p(x+y) \geq \min\{v_p(x), v_p(y)\}$. To see why this is true, write $x = p^n\frac{a}{b}$ and $y = p^m\frac{c}{d}$. For sake of argument, suppose that $n \geq m$. Then

$$\begin{aligned} x + y &= p^n\frac{a}{b} + p^m\frac{c}{d} = p^m(p^{n-m}\frac{a}{b} + \frac{c}{d}) \\ &= p^m\left(\frac{p^{n-m}ad + bc}{bd}\right). \end{aligned}$$

The denominator bc is not divisible by p since neither b nor c is divisible by p. The numerator may or may not be divisible by p. Therefore, $v_p(x+y) \geq m = \min\{v_p(x), v_p(y)\}$. By using the function v_p, we can define a subring of $\mathbb{Q}$ by

$$\begin{aligned} \mathcal{O}_p &= \{x \in \mathbb{Q}^* : v_p(x) \geq 0\} \cup \{0\} \\ &= \left\{\frac{a}{b} : a, b \in \mathbb{Z},\ p \nmid b\right\}. \end{aligned}$$

Moreover, $\mathcal{O}_p$ has an ideal M_p defined by

$$\begin{aligned} M_p &= \{x \in \mathbb{Q}^* : v_p(x) > 0\} \cup \{0\} \\ &= \left\{\frac{a}{b} : a, b \in \mathbb{Z},\ p \mid a,\ p \nmid b\right\}. \end{aligned}$$

Furthermore, every element of $\mathcal{O}_p \setminus M_p$ is a unit in$\mathcal{O}_p$, as such an element has the form $\frac{a}{b}$

with both a, b not divisible by p. Then $\frac{b}{a} \in \mathcal{O}_p$ is the inverse of $\frac{a}{b}$ in $\mathcal{O}_p$. Thus, $\mathcal{O}_p$ is a local ring with unique maximal ideal M_p. This ring is sometimes called the p-adic valuation ring of $\mathbb{Q}$. The residue field $\mathcal{O}_p/M_p$ is isomorphic to $\mathbb{F}_p$ via the map $\frac{a}{b} \mapsto \overline{a} \cdot (\overline{b})^{-1}$.

Example 3.4. Let $p(x)$ be an irreducible polynomial in $k[x]$. Every rational function in $k(x)$ can be written in the form $p(x)^n \frac{a(x)}{b(x)}$, where $a(x)$ and $b(x)$ are polynomials neither divisible by $p(x)$. We define a function $v_p : k(x)^* \to \mathbb{Z}$ by $v_p(\varphi(x)) = n$ if $\varphi(x) = p(x)^n \frac{a(x)}{b(x)}$ as above. With the same arguments as in the previous example, we see that $v_p(f(x)g(x)) = v_p(f(x)) + v_p(g(x))$ and $v_p(f(x)+g(x)) \geq \min\{v_p(f(x)), v_p(g(x))\}$ for all $f(x), g(x) \in k(x)^*$. Also, we have a local ring

$$\begin{aligned}\mathcal{O}_{p(x)} &= \{\varphi(x) \in k(x)^* : v_p(\varphi(x)) \geq 0\} \cup \{0\} \\ &= \left\{ \frac{a(x)}{b(x)} : a(x), b(x) \in k[x],\ p(x) \nmid b(x) \right\}\end{aligned}$$

with unique maximal ideal

$$\begin{aligned}M_p &= \{\varphi(x) \in k(x)^* : v_p(\varphi(x)) > 0\} \cup \{0\} \\ &= \left\{ \frac{a(x)}{b(x)} : a(x), b(x) \in k[x],\ p(x) \mid a(x),\ p(x) \nmid b(x) \right\}.\end{aligned}$$

Finally, we note that $\mathcal{O}_p/M_p \cong k[x]/(p(x))$ via the map $\frac{a(x)}{b(x)} \mapsto \overline{a(x)} \cdot \left(\overline{b(x)}\right)^{-1}$. This field is a finite dimensional field extension of k; recall that, up to isomorphism, $k[x]/(p(x))$ is the field extension of k obtained by adjoining to k a root of the polynomial $p(x)$, so its dimension over k is equal to the degree of $p(x)$.

We now give definitions of a discrete valuation ring and a discrete valuation. These definitions are equivalent, in some sense, as the previous examples indicate, although we make this equivalence more precise below.

Definition 3.5. *A discrete valuation ring of a field F is a principal ideal domain $R \subseteq F$ such that for every $a \in F^*$, either $a \in R$ or $a^{-1} \in R$.*

Definition 3.6. *A discrete valuation on a field F is a nonzero function $v : F^* \to \mathbb{Z}$ satisfying the properties (i) $v(ab) = v(a) + v(b)$ for all $a, b \in F^*$ and (ii) $v(a+b) \geq \min\{v(a), v(b)\}$ for all $a, b \in F^*$ with $a + b \neq 0$.*

A discrete valuation is then a group homomorphism from the multiplicative group of F^* to the additive group $\mathbb{Z}$. Our assumption that v is nonzero means v is not identically equal to 0. Since the image of v is a nonzero subgroup of $\mathbb{Z}$, this image is $n\mathbb{Z}$ for some $n > 1$. By replacing the function v by $\frac{1}{n}v$, we assume that v is surjective. To connect these definitions, first let v be a discrete valuation on a field F. Set

$$\mathcal{O}_v = \{a \in F^* : v(a) \geq 0\} \cup \{0\}.$$

The definition of a discrete valuation implies that $\mathcal{O}_v$ is a subring of F. We call $\mathcal{O}_v$ the valuation ring of v. Furthermore, if M_v is defined by

$$M_v = \{a \in F^* : v(a) > 0\} \cup \{0\},$$

then M_v is an ideal of $\mathcal{O}_v$. Moreover, every element a of $\mathcal{O}_v \setminus M_v$ satisfies $v(a) = 0$, and since this implies that $v(a^{-1}) = 0$, the element a is a unit of $\mathcal{O}_v$. This shows that M_v is the unique maximal ideal of $\mathcal{O}_v$. We claim that $\mathcal{O}_v$ is a discrete valuation ring. One property is easy to prove. Let $a \in F^*$. Since v is a homomorphism, $v(a^{-1}) = -v(a)$. Therefore, either $v(a) \geq 0$ or $v(a^{-1}) \geq 0$; this implies that $a \in \mathcal{O}_v$ or $a^{-1} \in \mathcal{O}_v$. To finish the claim, we need to show that $\mathcal{O}_v$ is a principal ideal domain. Let I be a nonzero ideal of $\mathcal{O}_v$. Let $n = \min\{v(a) : a \in I, a \neq 0\}$, and pick $a \in I$ with $v(a) = n$. We claim that $I = a\mathcal{O}_v$. The inclusion $a\mathcal{O}_v \subseteq I$ is clear. For the reverse inclusion, let $x \in I$. Then $v(x) \geq n$ by definition of n. Then $v(xa^{-1}) \geq 0$, so $xa^{-1} \in \mathcal{O}_v$. Therefore, $x \in a\mathcal{O}_v$. Therefore, $I \subseteq a\mathcal{O}_v$, and so $I = a\mathcal{O}_v$, as desired. This finishes the proof that $\mathcal{O}_v$ is a discrete valuation ring.

We now show that every discrete valuation ring of a field F is the valuation ring of some discrete valuation on F. Let R be a discrete valuation ring of F. We first note that R is a local ring with unique maximal ideal $M = \{r \in R : r \notin R^*\}$, the set of nonunits of R. To prove this it suffices to show that M is an ideal of R. It is clear that if $x \in M$ and $r \in R$, then $xr \in M$. It is also clear that if $x \in M$, then $-x \in M$. So, we are left to prove that if $x, y \in M$, then $x + y \in M$. It is enough to assume that both x and y are nonzero. Since R is a discrete valuation ring, either $xy^{-1} \in R$ or $yx^{-1} = (xy^{-1})^{-1} \in R$. Suppose that $xy^{-1} \in R$. Then $x = yr$ for some $r \in R$. Then $x + y = yr + y = y(1 + r)$. This cannot be a unit since it is a multiple of the nonunit y. Therefore, $x + y \in M$. Therefore, M is an ideal of R, and so M is the unique maximal ideal of R.

We define a valuation v on F as follows. First, for $x \in R$ with $x \neq 0$, set $v(x) = n$ if $x \in M^n \setminus M^{n+1}$. That is, $v(x) = n$ if n is the maximum integer satisfying $x \in M^n$. We view $R = M^0$ in this definition. To prove that v is well defined on R, we need to prove that there is such a maximum integer for any nonzero x. This amounts to proving that $\bigcap_{n=1}^{\infty} M^n = (0)$. We do this in the following lemma.

Lemma 3.7. *Let R be a local principal ideal domain. If M is the maximal ideal of R, then $\bigcap_{n=1}^{\infty} M^n = (0)$.*

Proof. Let $J = \bigcap_{n=1}^{\infty} M^n$, an ideal of R. Since R is a principal ideal domain, there is an $a \in R$ with $J = aR$. Write $M = tR$ for some $t \in R$. Then $M^n = t^n R$. For each n there is an $r_n \in R$ with $a = t^n r_n$. In particular, $tr_1 = t^n r_n$ for each n. Since R is a domain, $r_1 = t^{n-1} r_n$. This shows that $r_1 \in J$. If we write $r_1 = as$, then $a = tr_1 = tas$. If $a \neq 0$, then cancelling a gives $1 = ts$. This forces $t \in M$ to be a unit, which is false. Therefore, $a = 0$, and so $J = 0$. □

We point out that this lemma is an easy special case of the Krull intersection theorem, which states that if R is a local Noetherian ring with maximal ideal M, then $\bigcap_{n=1}^{\infty} M^n = (0)$.

As we pointed out before the lemma, we now have a well defined function $v : R\setminus\{0\} \to \mathbb{Z}$ defined by $v(x) = n$ if $x \in M^n \setminus M^{n+1}$. Since R is a principal ideal domain, there is a $t \in M$ with $M = tR$. Then $M^n = t^nR$. Therefore, each nonzero $x \in R$ can be written uniquely in the form $x = t^nu$ for some unit $u \in R^*$. From this fact it is easy to prove that v is a discrete valuation. First, if $x, y \in R$, then $x = t^nu$ and $y = t^mw$ for some nonnegative integers n, m and units u, w. Then $xy = t^{n+m}(uw)$. Since u and w are units, the product uw is also a unit. Then $v(xy) = n + m = v(x) + v(y)$. Second, suppose that $n \geq m$. Then $x + y = t^m(t^{n-m}u + w)$. If we write $t^{n-m}u + w = t^rs$ for some unit s, then $x + y = t^{m+r}s$, and so $v(x + y) = m + r \geq m = \min\{v(x), v(y)\}$. We extend v to F^* by defining $v(a/b) = v(a) - v(b)$. This is well defined; if $a/b = c/d$, then $ad = bc$. Therefore, $v(ad) = v(bc)$, or $v(a)+v(d) = v(b)+v(c)$, so $v(a)-v(b) = v(c)-v(d)$. We have prove that $v(xy) = v(x) + v(y)$ and $v(x + y) \geq \min\{v(x), v(y)\}$ for $x, y \in R$. A short argument shows that these properties hold for $x, y \in F^*$, which will prove that v is a discrete valuation ring.

We point out some more properties of a discrete valuation ring of F. First, we will show that a discrete valuation ring $\mathcal{O}$ of F is integrally closed in F. Recall that if $R \subseteq S$ are commutative rings, an element $a \in S$ is said to be *integral* over R if a satisfies an equation of the form $a^n + r_{n-1}a^{n-1} + \cdots + r_0 = 0$ with each $r_i \in R$. That is, a is integral over R if a satisfies a monic polynomial equation with coefficients in R. A ring R with quotient field F is said to be *integrally closed* if whenever $a \in F$ is integral over R, then $a \in R$.

Lemma 3.8. *Let $\mathcal{O}$ be a discrete valuation ring of a field F. Then $\mathcal{O}$ is integrally closed.*

Proof. Let $a \in F$ be integral over $\mathcal{O}$. Then there are $r_i \in \mathcal{O}$ with $a^n + r_{n-1}a^{n-1} + \cdots + r_0 = 0$. Let v be a valuation on F whose valuation ring is $\mathcal{O}$. Then from the equation $-a^n = r_{n-1}a^{n-1} + \cdots + r_0$, we have

$$nv(a) = v(-a^n) \geq \min_{0\leq i\leq n-1}\left\{v(r_ia^i)\right\} = \min_{0\leq i\leq n-1}\left\{v(r_i) + iv(a)\right\}.$$

If this minimum occurs at $i = j$, then $nv(a) \geq v(r_i) + iv(a) \geq iv(a)$, since $r_i \in \mathcal{O}$. Then $(n - i)v(a) \geq 0$, which forces $v(a) \geq 0$. Thus, $a \in \mathcal{O}$, as desired. □

We will use the following lemma to help describe the discrete valuation rings of $k(x)/k$, although it has other uses.

Lemma 3.9. *Let $\mathcal{O}$ be a discrete valuation ring of F. Then $\mathcal{O}$ is a maximal subring of F. That is, if A is a ring with $\mathcal{O} \subseteq A \subseteq F$, then $A = \mathcal{O}$ or $A = F$.*

Proof. Suppose that A is a ring with $\mathcal{O} \subset A \subseteq F$. Let $a \in A \setminus \mathcal{O}$. Then $a^{-1} \in \mathcal{O}$. Suppose that t is a *uniformizer* of $\mathcal{O}$. Then $a^{-1} \in t\mathcal{O}$, so $a^{-1} = t^nu$ for some unit $u \in \mathcal{O}$. So, $a = t^{-n}u^{-1}$. Because $\mathcal{O} \subseteq A$, we see that $t^{n-1}ua = t^{-1} \in A$. We use this to prove that $A = F$. Let $x \in F$ be nonzero, and set $m = v(x)$. Then $t^{-m}x \in \mathcal{O}$. So, $x = t^mc$ for some $c \in \mathcal{O}$. However, since $t \in \mathcal{O}$ and $t^{-1} \in A$, both are in A, so $t^m \in A$. Thus, $x = t^mc \in A$, which proves that $A = F$. □

We finish this section by discussing valuation rings of a field and of a subfield. Let K/F be a finite dimensional extension of fields, and let $\mathcal{O}$ be a discrete valuation ring of K. Suppose also that v is a discrete valuation of K whose valuation ring is $\mathcal{O}$. It is immediate to see that $v|_F$ is a valuation on F once we see that $v|_F$ is not identically 0. If $v|_F$ is the 0 function, then $F \subseteq \mathcal{O}$. Since K is algebraic over F, every element of Kwould be integral over $\mathcal{O}$, which would force $K = \mathcal{O}$ since $\mathcal{O}$ is integrally closed. This is false; thus, $v|_F$ is nontrivial. The valuation ring of $v|_F$ is clearly $F \cap \mathcal{O}$. We will use this in the following way. Let F/k be an algebraic function field in one variable, and let $x \in F$ be transcendental over k. Then $F/k(x)$ is a finite extension. If $\mathcal{O}$ is a discrete valuation ring of F, then $\mathcal{O} \cap k(x)$ is a discrete valuation ring of $k(x)$. We will be able use this idea once we know what are the discrete valuation rings of $k(x)$. We describe these rings in the next section.

3.2 Discrete Valuation Rings of $k(x)/k$

In this section we determine the discrete valuation rings of $k(x)/k$. Recall that we have shown that if k is algebraically closed, then $k(x)$ is the function field of the affine curve $Z(y)$, and that there is a 1-1 correspondence between points $(a, 0)$ on this curve and the discrete valuation rings $k[x]_{(x-a)}$ of $k(x)$. We generalize this example by allowing k to be arbitrary. We will see more valuation rings, which correspond to non-rational points of $Z(y)$.

Let $\mathcal{O}$ be a discrete valuation ring of $k(x)/k$. Then $x \in \mathcal{O}$ or $x^{-1} \in \mathcal{O}$. We first assume that $x \in \mathcal{O}$. Then $k[x] \subseteq \mathcal{O}$. Let P be the maximal ideal of $\mathcal{O}$, and set $M = P \cap k[x]$, a prime ideal of $k[x]$. Then $k[x]_M \subseteq \mathcal{O}$ since $k[x] \setminus M \subseteq \mathcal{O} \setminus P$ and $\mathcal{O} \setminus P = \mathcal{O}^*$ is the set of units of $\mathcal{O}$. Since $\mathcal{O} \neq k(x)$, the ideal M is nonzero. Since nonzero prime ideals of $k[x]$ are maximal, we now know that M is maximal. Therefore, $M = (p(x))$ for some irreducible polynomial $p(x)$. We have seen that $k[x]_{(p(x))}$ is a discrete valuation ring of $k(x)$. Moreover, since $k[x]_{(p(x))} \subseteq \mathcal{O}$ and since discrete valuation rings are maximal subrings of their quotient field, we have $\mathcal{O} = k[x]_{(p(x))}$. This proves that any discrete valuation ring of $k(x)/k$ that contains x is of the form $k[x]_{(p(x))}$. Now suppose that $x \notin \mathcal{O}$. Then $x^{-1} \in \mathcal{O}$; furthermore, $x^{-1} \in P$ else x is a unit in $\mathcal{O}$, and then $x \in \mathcal{O}$. We consider the ring $k[x^{-1}]$, which is isomorphic to the polynomial ring $k[x]$ via the map $f(x^{-1}) \longmapsto f(x)$. Thus, we may think of x^{-1} as the variable of the polynomial ring $k[x^{-1}]$. By the previous argument, $M = k[x^{-1}] \cap P$ is a maximal ideal of $k[x^{-1}]$. However, $x^{-1} \in M$, so $(x^{-1}) \subseteq M$. However, x^{-1} is clearly irreducible, so $M = (x^{-1})$. Continuing to mimic the argument above, this forces $\mathcal{O} = k[x^{-1}]_{(x^{-1})}$. We have thus shown that the discrete valuation rings of $k(x)/k$ are

$$\left\{k[x]_{(p(x))} : p(x) \in k[x] \text{ is irreducible}\right\} \cup \left\{k[x^{-1}]_{(x^{-1})}\right\}.$$

To make a geometric correspondence, if $X = \mathbf{A}^1(K)$ with K an algebraic closure of k, and if $a \in K$ is a point of X, let $p(x)$ be the minimal polynomial of a over k. Then $p(x)$ is an irreducible polynomial, and arguments similar to those we saw earlier show that $\mathcal{O}_P(X/k) = k[x]_{(p(x))}$. Since every irreducible polynomial $p(x)$ has a root in K, every valuation ring of

the form $k[x]_{(p(x))}$ occurs as the local ring of a point of X. These account for all but one of the discrete valuation rings of $k(x)/k$. If we consider the projective line $\mathbf{P}^1(K)$, by viewing $\mathbf{A}^1 \subseteq \mathbf{P}^1$ via $a \longmapsto (a:1)$, then we have accounted for the local rings of all but one point, the point $(1:0)$. If we view $\mathbf{A}^1$ inside $\mathbf{P}^1$ in a different way, via $a \longmapsto (1:a)$, then these two copies have a great deal of overlap. If $a \neq 0$, then $(a:1) = (1:a^{-1})$. Therefore, if we interpret $x \in k(x)$ as the rational function for which $x(a:1) = a$, then $x(1:a^{-1}) = a$. Therefore, $x^{-1}(1:a^{-1}) = a$. Thus, x^{-1} acts as the "variable" on the second affine piece. This indicates that replacing the first affine line by the second should correspond to replacing x by x^{-1}. Since the valuation ring of $(0:1)$ is $k[x]_{(x)}$, by this symmetry argument, the valuation ring of $(1:0)$ is $k[x^{-1}]_{(x^{-1})}$. Therefore, by considering the projective line instead of the affine line, we account for all discrete valuation rings of $k(x)/k$.

3.3 Discrete Valuation Rings of F/k

Let F/k be an algebraic function field in one variable. A discrete valuation ring of F that contains k is said to be a discrete valuation ring of F/k. If k is a finite field, then an exercise shows that any discrete valuation ring of F contains k. Therefore, in this case, any discrete valuation ring of F is also a discrete valuation ring of F/k.

Lemma 3.10. *Let $\mathcal{O}$ be a discrete valuation ring of F/k. If k' is the field of constants of F/k, then $k' \subseteq \mathcal{O}$.*

Proof. Let $a \in F$ be algebraic over k. We wish to show that $a \in \mathcal{O}$. Since a is algebraic over k, there are $\alpha_i \in k$ with $a^n + \alpha_{n-1}a^{n-1} + \cdots + \alpha_0 = 0$. Note that each $\alpha_i \in k$. Thus, a is integral over $\mathcal{O}$. Since $\mathcal{O}$ is integrally closed in F, we see that $a \in \mathcal{O}$, as desired. □

A *place* of F/k is a maximal ideal of a discrete valuation ring of F/k. Let $\mathbf{P}_F$ be the set of all places of F/k. This set takes the place of an algebraic curve; if k is algebraically closed, then there is a 1-1 correspondence between points on a nonsingular projective curve X and places of $k(X)/k$. To mimic the notation of algebraic curves, we typically denote places of F/k by P, and we denote the valuation ring corresponding to the place P by $\mathcal{O}_P$. Since $\mathcal{O}_P$ is a discrete valuation ring, $P = t\mathcal{O}_P$ for some t; such an element is called a uniformizer of $\mathcal{O}_P$. Given a place P, the *residue field* of P is defined as

$$k(P) = \mathcal{O}_P/P.$$

Furthermore, we define the *degree* of P by $\deg(P) = [k(P):k]$. The field k is a subfield of $k(P)$, since k is a subring of $\mathcal{O}_P$, and $k \cap P = (0)$. Thus, $k \cong k/(0)$ embeds in $\mathcal{O}_P/P = k(P)$.

Proposition 3.11. *Let F/k be an algebraic function field in one variable, and let $P \in \mathbf{P}_F$. If $x \in P$, then* $\deg(P) \leq [F:k(x)] < \infty$.

Proof. First note that any element $a \in k'$ is a unit in $\mathcal{O}_P$ since both a and a^{-1} are in $\mathcal{O}_P$. Thus, if $x \in P$, then x is transcendental over k, so $[F:k(x)] < \infty$. We are left to prove

that $\deg(P) \leq [F : k(x)]$. Let $a_1, \ldots, a_n \in \mathcal{O}_P$ such that their images $\overline{a_1}, \ldots, \overline{a_n}$ in $k(P)$ are linearly independent over k. We claim that $\{a_1, \ldots, a_n\}$ is linearly independent over $k(x)$. Note that since $x \in P \subseteq \mathcal{O}_P$, we have $k[x] \subseteq \mathcal{O}_P$. Furthermore, $P \cap k[x] = (x)$ since $x \in P$ and (x) is a maximal ideal of $k[x]$. Thus, $k[x]_{(x)} \subseteq \mathcal{O}_P$. Since $k[x]_{(x)}$ is a discrete valuation ring of $k(x)$, we have $\mathcal{O}_P \cap k(x) = k[x]_{(x)}$. The residue field of the valuation ring $k[x]_{(x)}$ is k, since $k[x]_{(x)}/xk[x]_{(x)} \cong k$. Now, to prove our claim, suppose that there are $c_i \in k(x)$ with $\sum_i c_i a_i = 0$. If v is the valuation on $k(x)$ corresponding to the valuation ring $k[x]_{(x)}$, by dividing through by the coefficient c_i that has smallest value with respect to v, we may assume that the c_i are elements of the valuation ring $k[x]_{(x)}$, and that at least one has value 0. Reducing modulo P, we have $\sum_i \overline{c_i} \cdot \overline{a_i} = 0$. Now, $\overline{c_i}$ lie not just in the residue field $k(P) = \mathcal{O}_P/P$, but in fact the lie in the residue field k of $k[x]_{(x)}$. Therefore, this equation is a linear dependence relation over k of the $\overline{a_i}$. Since $\{\overline{a_1}, \ldots, \overline{a_n}\}$ is independent over k, each $\overline{c_i} = 0$. However, one of the c_i has value 0, so $\overline{c_i} \neq 0$. This contradiction shows that $\{a_1, \ldots, a_n\}$ is linearly independent over $k(x)$, as desired. □

Let $f \in F$. We view f as a function on $\mathbf{P}_F$ by defining $f(P) = \overline{f} \in \mathcal{O}_P/P = k(P)$ if $f \in \mathcal{O}_P$. If $f \notin \mathcal{O}_P$, then we do not define $f(P)$. We call P a *zero* of f if $f(P) = 0$ and P a *pole* of f if $f(P)$ is not defined. Since $f(P) = 0$ if and only if $f \in P$, and $f(P)$ is not defined if and only if $f \notin \mathcal{O}_P$, if v_P is the valuation of F corresponding to $\mathcal{O}_P$, then P is a zero of f if and only if $v_P(f) > 0$ and P is a pole of f if and only if $v_P(f) < 0$. If P is a zero of f, then the positive integer $v_P(f)$ is called the *order of the zero* P. If P is a pole of f, then the positive integer $-v_P(f)$ is called the *order of the pole* P. Note that P is a zero of f if and only if P is a pole of f^{-1}. We will prove that every nonconstant $f \in F$ has only finitely many zeros and only finitely many poles. In fact, we will give much more detailed information about the number of zeros and poles and their orders.

If $F = k(x)$, then these definitions of zero and pole are the same as we gave earlier, as is the definitions of order. For, if $\varphi(x) = g(x)/h(x)$ is a rational function having a zero of order r at a, then $g(x) = (x-a)^r g_1(x)$ for some polynomial $g_1(x)$ with $g_1(a) \neq 0$. Also, we assume that $g(x)/h(x)$ is in reduced form, so $h(a) \neq 0$. If v is the valuation corresponding to the discrete valuation ring $k[x]_{(x-a)}$, then $v(\varphi(x)) = v(g(x)) - v(h(x)) = r$, since $x - a$ generates the maximal ideal P of this valuation ring, and so $\varphi(x) \in P^r \setminus P^{r+1}$. The argument for poles is similar.

We now show that every $f \in F$ has at least one zero and at least one pole, provided that f is not a constant. We use a theorem from the handout on Dedekind domains: if K/F is a finite extension, if A is a Dedekind domain of F, and if B is the integral closure of K, then B is a Dedekind domain. Recall that a discrete valuation ring is a special example of a Dedekind domain, being nothing more than a Dedekind domain that is also a local ring.

Proposition 3.12. *Let $x \in F$ be nonconstant. Then x has at least one zero and at least one pole.*

Proof. Let $x \in F$ be nonconstant, and consider the finite extension $F/k(x)$. Let B be the integral closure of $k[x]_{(x)}$ in F. Then B is a Dedekind domain of F. If M is a maximal ideal

of B, then B_M is a discrete valuation ring. Moreover, $M \cap k[x]_{(x)}$ is a nonzero prime ideal of $k[x]_{(x)}$, so $M \cap k[x]_{(x)} = xk[x]_{(x)}$. Therefore, $x \in M$. Thus, if we set $B_M = \mathcal{O}_P$ with $P \in \mathbf{P}_F$, then $x \in P$, as $M = P \cap B$. Therefore, P is a zero of x. By using the same argument with x replaced by x^{-1}, we see that x^{-1} has a zero Q. Then Q is a pole of x. $\square$

Let F/k be an algebraic function field in one variable, and let k' be the field of constants. For any $P \in \mathbf{P}_F$, we have seen that $k' \subseteq \mathcal{O}_P$. The field k' then embeds in the residue field $\mathcal{O}_P/P = k(P)$ via $f \mapsto f + P$. Therefore, if $f \in k'$ is nonzero, $v_P(f) = 0$ for all P. Therefore, $f(P) = f + P \in k(P)$ is equal to f under the embedding $k' \subseteq \mathcal{O}_P$. Thus, by viewing f as a function on $\mathbf{P}_F$, it is a constant function. This justifies our naming k' to be the field of constants.

4 Divisors and the Riemann-Roch Theorem

In algebraic number theory the main ring of study is the ring of integers in an algebraic number field F. This is the integral closure A of $\mathbb{Z}$ in F. The ring A is a Dedekind domain. Associated to A is the group of fractional ideals, a group under multiplication (of ideals). The principal ideals form a subgroup, and the resulting quotient group is called the ideal class group of A. The study of this group has important consequences. For example, A is a unique factorization domain if and only if the class group is 0. We start by discussing the geometric analogue of this notion, which are divisors, principal divisors, and divisor classes.

4.1 Divisors of a Function Field

Let F/k be an algebraic function field in one variable. By replacing k by the field of constants, we will assume that k is the exact constant field of F/k in this chapter. Recall notation we used in the previous chapter: $\mathbf{P}_F$ is the set of places of F/k. If $P \in \mathbf{P}_F$, then $\mathcal{O}_P$ is the corresponding discrete valuation ring and v_P is the discrete valuation associated to $\mathcal{O}_P$.

Definition 4.1. *Let F/k be an algebraic function field in one variable. A divisor is an element of the free Abelian group on $\mathbf{P}_F$. That is, a divisor is a finite sum $\sum n_P P$, where $n_P \in \mathbb{Z}$ and $P \in \mathbf{P}_F$.*

As part of this definition, two divisors $\sum_P n_P P$ and $\sum_P m_P P$ are equal if and only if $n_P = m_P$ for all P. In particular, $\sum_P n_P P = 0$ if and only if each $n_P = 0$. If $D = \sum_P n_P P$ is a divisor, we define the *support* of D to be $\operatorname{supp}(D) = \{P : n_P \neq 0\}$. This is a finite subset of $\mathbf{P}_F$. If $P \in \mathbf{P}_F$ and D is a divisor, we set $v_P(D)$ to be the coefficient of P in D. Then $v_P(D) = 0$ for all but finitely many P. There is a partial ordering $\leq$ on the set D_F of divisors of F/k, given by $D \leq E$ if $v_P(D) \leq v_P(E)$ for all $P \in \mathbf{P}_F$. A divisor is said to be *positive* if $D \geq 0$. Note that a divisor need not be positive or negative. For example, if $P, Q \in \mathbf{P}_F$, then the divisor $P - Q$ is neither positive or negative. The *degree* of a divisor is defined as $\deg(D) = \sum_P v_P(D) \deg(P)$. This is a finite sum since only finitely many $v_P(D)$ are nonzero and since $\deg(P)$ is finite for any P. If D is a positive divisor, then $\deg(D) \geq 0$. Furthermore, if $D \leq E$, then $\deg(D) \leq \deg(E)$.

Lemma 4.2. *Let D and E be divisors on F/k. Then $\deg(D+E) = \deg(D) + \deg(E)$ and $\deg(-D) = -\deg(D)$. Thus, the degree map is a group homomorphism from D_F to $\mathbb{Z}$.*

Proof. Let $D = \sum_P n_P P$ and $E = \sum_P m_P P$. Then $D + E = \sum_P (n_P + m_P) P$. Thus,

$$\begin{aligned} \deg(D+E) &= \sum_P (n_P + m_P) \deg(P) = \sum_P n_P \deg(P) + \sum_P m_P \deg(P) \\ &= \deg(D) + \deg(E). \end{aligned}$$

Also, $\deg(-D) = \sum_P -n_P \deg(P) = -\sum_P n_P \deg(P)$, so $\deg(-D) = -\deg(D)$. □

Definition 4.3. *Let $f \in F^*$. The divisor (f) of f is defined as $(f) = \sum_P v_P(f)P$. The zero divisor of f is $(f)_0 = \sum_{v_P(f)>0} v_P(f)P$ and the pole divisor is $(f)_\infty = \sum_{v_P(f)<0} -v_P(f)P$.*

A divisor D is said to be *principal* if $D = (f)$ for some $f \in F^*$. If f is a constant, then $v_P(f) = 0$ for all P, so $(f) = 0$. Note that $(f) = (f)_0 - (f)_\infty$. In order to see that the definition above makes sense, we need to prove that any $f \in F^*$ has only finitely many zeros and only finitely many poles. This result is perhaps the most important theorem leading up to the Riemann-Roch theorem. The following lemma will allow us to use one of the results in the Dedekind domain handout in proving this theorem.

Lemma 4.4. *Let F/k be an algebraic function field in one variable, and let $x \in F$ be transcendental over k. Let $\mathcal{O}$ be a discrete valuation ring of F/k with maximal ideal P such that $x \in P$. Then $\mathcal{O} \cap k(x) = k[x]_{(x)}$.*

Proof. Since $x \in P \subseteq \mathcal{O}$, and since $k \subseteq \mathcal{O}$ by hypothesis, $k[x] \subseteq \mathcal{O}$. Let $M = P \cap k[x]$, a prime ideal of $k[x]$. We have $x \in M$, so $(x) \subseteq M$. However, x is irreducible in $k[x]$, so (x) is a maximal ideal. Thus, $M = (x)$. Furthermore, since $k[x] \setminus (x) \subseteq \mathcal{O} \setminus P = \mathcal{O}^*$, the group of units of $\mathcal{O}$, we have $k[x]_{(x)} \subseteq \mathcal{O}$. Finally, as $k[x]_{(x)}$ is a discrete valuation ring, $k[x]_{(x)} = \mathcal{O}$ since discrete valuation rings are maximal subrings of their quotient fields. □

Theorem 4.5. *Let $x \in F \setminus k$. Then $\deg((x)_0) = \deg((x)_\infty) = [F : k(x)]$. Therefore, x has only finitely many zeros and only finitely many poles.*

Proof. Let $x \in F$ be nonconstant. Suppose that $P_1, \ldots, P_t$ are zeros of x. Thus, $x \in P_i$ for each i. By the previous lemma, $\mathcal{O}_{P_i}$ contains $k[x]_{(x)}$. Conversely, if P is a point with $k[x]_{(x)} \subseteq \mathcal{O}_P$, then $x \in P$, since $\mathcal{O}_P \cap k[x]_{(x)} = xk[x]_{(x)}$. The zeros of x are then precisely those P for which $\mathcal{O}_P$ contains $k[x]_{(x)}$. We proved in the Dedekind domain handout that there are only finitely many discrete valuation rings of F that contain the discrete valuation ring $k[x]_{(x)}$. Therefore, x has only finitely many zeros. Moreover, if P_i is a zero of order e_i, then $v_{P_i}(x) = e_i$. Let B be the integral closure of $k[x]_{(x)}$ in F. If $M_i = P_i \cap B$, then $xB = M_1^{e_1} \cdots M_t^{e_t}$, which was proved in the handout. Thus, by the handout, $[F : k(x)] = \sum_i e_i f_i$, where $f_i = [B/M_i : k[x]_{(x)}/xk[x]_{(x)}]$. The field $k[x]_{(x)}/xk[x]_{(x)}$ is equal to k, and $B/M_i = \mathcal{O}_{P_i}/P_i = k(P_i)$. Thus, $f_i = \deg(P_i)$. Therefore, $[F : k(x)] = \sum_i e_i \deg(P_i)$, and since $(x)_0 = \sum_i e_i P_i$, we have $\deg((x)_0) = [F : k(x)]$. Finally, the pole divisor of x is equal to the zero divisor of x^{-1}. Thus, this argument, applied to x^{-1}, shows that $\deg((x)_\infty) = \deg((x^{-1})_0) = [F : k(x)]$. □

Corollary 4.6. *Let $f \in F^*$. Then $\deg((f)) = 0$.*

Proof. In the previous theorem we proved that $\deg((x)_0) = \deg((x)_\infty)$. Since $(x) = (x)_0 - (x)_\infty$ and the degree function preserves addition , $\deg((x)) = \deg((x)_0)) - \deg((x)_\infty)) = 0$. □

Corollary 4.7. *Let D be a divisor with $\deg(D) = 0$. Then the following conditions are equivalent.*

1. *D is principal,*

2. *$\dim(D) \geq 1$,*

3. *$\dim(D) = 1$.*

Proof. We have seen in an example that if $D = (f)$ is principal, then $L(D) = f^{-1}k$. Therefore, $\dim(D) = 1$. This proves (1) $\implies$ (3). The proof of (3) $\implies$ (2) is obvious. Finally, for (2) $\implies$ (1), suppose that $\dim(D) \geq 1$. Take $f \in L(D)$ be nonzero. Then $D + (f) \geq 0$. This is then a positive divisor of degree equal to $\deg(D) = 0$ by Corollary 4.6. This forces $D + (f) = 0$, so $D = -(f) = (f^{-1})$ is principal. □

By using principal divisors, we can define a new group. First, note that $(f) + (g) = (fg)$ and $-(f) = (f^{-1})$. Therefore, the set of principal divisors is a subgroup of D_F. The *divisor class group* C_F is defined to be the quotient group of D_F modulo the subgroup of principal divisors. If D and E are divisors that have the same divisor class in C_F, then we say that D and E are equivalent. In other words, D is *equivalent* to E if there is an $f \in F^*$ with $D = E + (f)$. We will write $D \sim E$ if D and E are equivalent.

Definition 4.8. *Let D be a divisor on F/k. Then $L(D) = \{f \in F^* : D + (f) \geq 0\} \cup \{0\}$.*

Example 4.9. If 0 is the trivial divisor, then $L(0) = k$. To see this, if $f \in L(0)$ is nonzero, then $(f) + 0 \geq 0$, so $(f) \geq 0$. Therefore, $v_P(f) \geq 0$ for all $P \in \mathbf{P}_F$. This forces f to have no poles. Since any nonconstant function has a pole, f must be a constant. Conversely, if f is a constant, then $(f) = 0$, so $(f) + 0 \geq 0$. This proves the equality $L(0) = k$.

Example 4.10. Let $D < 0$ be a negative divisor. Then $L(D) = 0$. For, if $f \in L(D)$ was nonzero, then $D + (f) \geq 0$. However, this would force $\deg(D + (f)) \geq 0$. Since $\deg(D + (f)) = \deg(D) + \deg((f)) = \deg(D) < 0$ by Corollary 4.6, this is impossible. Thus, $f = 0$.

Example 4.11. Let $D = (f)$ be a principal divisor. We claim that $L(D) = f^{-1}k$. First, it is clear that $f^{-1}k \subseteq L(D)$ since if a is any constant, then $(af^{-1}) + (f) = (f^{-1}) + (f) = 0$. Conversely, if $g \in L(D)$, then $(g) + (f) \geq 0$. Then (gf) is a positive divisor. This forces gf to be a constant since any nonconstant function has a pole. So, $gf = a$ for some $a \in k$, and so $g = af^{-1} \in f^{-1}k$.

Example 4.12. Let $F = k(x)$, and let P_∞ correspond to the discrete valuation ring $k[x^{-1}]_{(x^{-1})}$. We first note that if v_∞ is the valuation that corresponds to this ring, then $v_\infty(f(x)) = -\deg(f(x))$ for any $f(x) \in k[x]$. To see this, we have $v_\infty(x^{-1}) = 1$ since x^{-1} is a generator of P_∞. Therefore, $v_\infty(x) = -1 = -\deg(x)$. For $f(x) = a_n x^n + \cdots + a_0$ a polynomial of degree n, we can write $f(x) = x^n(a_n + a_{n-1}x^{-1} + \cdots + a_0 x^{-n})$. The second term has value 0 since $v_\infty(a_n) = 0$ and all other terms have positive value. Thus, $v_\infty(f(x)) = v_\infty(x^n) = -n = -\deg(f(x))$.

Let n be a positive integer, and consider the divisor nP_∞. Then

$$L(nP_\infty) = \{\varphi(x) \in k(x) : (\varphi(x)) + nP_\infty \geq 0\}.$$

This means $\varphi(x) \in L(nP_\infty)$ if and only if $v_\infty(\varphi(x)) \geq -n$ and $v_P(\varphi(x)) \geq 0$ for all $P \neq P_\infty$. If we write $\varphi(x) = f(x)/g(x)$ in reduced form, then for any irreducible factor $p(x)$ of $g(x)$, we have $v_{p(x)}(\varphi(x)) < 0$; since this is impossible, we see that $g(x) = 1$. Therefore, $\varphi(x) \in k[x]$. By our description of v_∞, we then see that $L(nP_\infty) = \{f(x) \in k[x] : \deg(f(x)) \leq n\}$. This is a k-vector space with basis $\{1, x, \ldots, x^n\}$, so its dimension is $n+1$. Note that $\deg(P_\infty) = 1$, so $\deg(nP_\infty) = n$. In other words, we see that $\dim_k(L(nP_\infty)) = 1 + \deg(nP_\infty)$.

Example 4.13. Let $F = k(x)$, and let $a_1, \ldots, a_n \in k$. Consider $P_1, \ldots, P_n$ corresponding, respectively, to the irreducible polynomials $x - a_1, \ldots, x - a_n$. We calculate $L(D)$, where $D = e_1P_1 + \cdots + e_nP_n$, given positive integers e_i. Let $\varphi(x) \in L(D)$. Then $v_{P_i}(\varphi(x)) \geq -e_i$. Write $\varphi(x) = f(x)/g(x)$ in reduced form; thus, $f(x)$ and $g(x)$ have no common term. Then the poles of $\varphi(x)$ are the zeros of $g(x)$, and so a_i can be a zero of $g(x)$ of order at most e_i. Note that a_i need not actually be a pole of $\varphi(x)$. Since $g(x)$ can have no zeros other than the various a_i, we see that $g(x) = (x-a_1)^{d_1} \cdots (x-a_n)^{d_n}$ with $0 \leq d_i \leq e_i$ for each i. If we write $\varphi(x) = f(x)/\left((x-a_1)^{d_1} \cdots (x-a_n)^{d_n}\right)$, then $v_{P_i}(\varphi(x)) \geq -e_i$, as desired. Furthermore, if P is any other place other than P_∞, then $v_P(\varphi(x)) \geq 0$ since the value of the denominator is 0. We only need to consider P_∞, whose valuation v_∞ satisfies $v_\infty(f(x)) = -\deg(f(x))$ for any polynomial $f(x)$. Therefore, $v_\infty(\varphi(x)) = (\sum_i d_i) - \deg(f)$. Thus,

$$L(D) = \left\{ \frac{f(x)}{(x-a_1)^{d_1} \cdots (x-a_n)^{d_n}} : f(x) \in k[x],\ 0 \leq d_i \leq e_i, \deg(f) \leq \sum_{i=1}^{n} d_i \right\}.$$

In fact, one can give a slightly simpler description of $L(D)$ as

$$L(D) = \left\{ \frac{f(x)}{(x-a_1)^{e_1} \cdots (x-a_n)^{e_n}} : f(x) \in k[x], \deg(f(x)) \leq \sum_{i=1}^{n} e_i \right\}.$$

Therefore, $L(D)$ is isomorphic to the space of all polynomials of degree $\leq \sum_{i=1}^n e_i = \deg(D)$. This is a space of dimension $1+\deg(D)$. It is not a coincidence that $\dim_k(L(D)) = 1+\deg(D)$. We will see that this is a general fact for divisors of $k(x)/k$ of positive degree.

We point out some simple properties about the spaces $L(D)$.

Lemma 4.14. *Let F/k be an algebraic function field in one variable, and let D be a divisor on F/k.*

1. *If $f \in F$, then $f \in L(D)$ if and only if $v_P(f) \geq -v_P(D)$ for all $P \in \mathbf{P}_F$.*
2. *The set $L(D)$ is a k-subspace of F.*

3. *The space $L(D)$ is nonzero if and only if D is equivalent to a positive divisor.*

4. *If E is equivalent to D, then $L(E) \cong L(D)$ as k-vector spaces.*

Proof. (1) We have $f \in L(D)$ if and only if $D + (f) \geq 0$, if and only if $v_P(D) + v_P(f) \geq 0$ for all $P \in \mathbf{P}_F$, and this occurs if and only if $v_P(f) \geq -v_P(D)$ for all $P \in \mathbf{P}_F$. This proves (1). For (2), let $f, g \in L(D)$. Then $v_P(f) \geq -v_P(D)$ and $v_P(g) \geq -v_P(D)$ for all P. By the definition of a discrete valuation, $v_P(f+g) \geq \min\{v_P(f), v_P(g)\}$. Thus, $v_P(f+g) \geq -v_P(D)$ for all P, so $f + g \in L(D)$. Also, if $f \in L(D)$ and $a \in k$, then $(af) = (a) + (f) = (f)$. Therefore, $D + (af) = D + (f) \geq 0$, so $af \in L(D)$. Thus, $L(D)$ is a k-vector space. To prove (3), recall that D is equivalent to $D + (f)$ for any $f \in F^*$. If $f \in L(D)$ is nonzero, then $D + (f) \geq 0$. This is a positive divisor equivalent to D. Conversely, if D is equivalent to a positive divisor E, then $E = D + (f)$ for some $f \in F^*$, and then $f \in L(D)$ since $E \geq 0$. Thus, $L(D)$ is nonzero. This proves (3). Finally, for (4), suppose that E is equivalent to D. Then $E = D + (f)$ for some $f \in F^*$. We define a map $\varphi : L(D) \to L(E)$ by $\varphi(x) = xf$. Then

$$E + (\varphi(x)) = E + (xf) = E + (x) + (f) = D + (x).$$

Therefore, $E + (\varphi(x)) \geq 0$ if and only if $D + (x) \geq 0$. Therefore, $x \in L(D)$ if and only if $\varphi(x) \in L(E)$. This shows that φ does map $L(D)$ to $L(E)$ and that φ is surjective. The map φ is k-linear since $\varphi(x + y) = (x + y)f = xf + yf = \varphi(x) + \varphi(y)$ for any $x, y \in L(D)$ and $\varphi(ax) = (ax)f = a(xf) = a\varphi(x)$ for any $a \in k$. Finally, φ is injective since if $\varphi(x) = 0$, then $xf = 0$. Since $f \neq 0$ and F is a field, $x = 0$. Therefore, $L(E) \cong L(D)$ as k-vector spaces. □

If D is a divisor, we set $\dim(D) = \dim_k(L(D))$. We will show later that $L(D)$ is a finite dimensional k-vector space.

Corollary 4.15. *Let F/k be an algebraic function field in one variable.*

1. *Let D be a divisor with $\deg(D) < 0$. Then $\dim(D) = 0$.*

2. *If E and D are equivalent divisors, then $\deg(E) = \deg(D)$ and $\dim(E) = \dim(D)$.*

Proof. To prove (1), if $f \in L(D)$ is nonzero, then $D+(f) \geq 0$. Then $\deg(D+(f)) = \deg(D) < 0$, while $\deg(D + (f)) \geq 0$ since $D + (f) \geq 0$. This is a contradiction. Show, $L(D) = 0$, so $\dim(D) = 0$. For (2), we proved in the lemma that if E and D are equivalent, then $L(E) \cong L(D)$. Therefore, $\dim(E) = \dim(D)$. We have already seen that $\deg(E) = \deg(D)$ since the degree of any principal divisor is 0, by Corollary 4.6. □

We now start to investigate the dimensions of divisors. The next lemma shows that the quotient space $L(E)/L(D)$ is finite dimensional, if $D \leq E$. This is the first step toward showing that $L(E)$ itself is finite dimensional.

Lemma 4.16. *Let D and E be divisors with $D \leq E$. Then $L(D) \subseteq L(E)$, and the quotient vector space $L(E)/L(D)$ satisfies $\dim_k(L(E)/L(D)) \leq \deg(E) - \deg(D)$.*

Proof. We first suppose that $E = D+P$ for some $P \in \mathbf{P}_F$. Then $\deg(E)-\deg(D) = \deg(P)$. Choose $t \in F$ with $v_P(t) = v_P(E)$. Note that $v_P(E) = v_P(D) + 1$. If $f \in L(E)$, then $v_Q(f) \geq -v_Q(E)$ for all Q. In particular, $v_P(f) \geq -v_P(E) = -v_P(t)$. Thus, $v_P(tf) \geq 0$. We define a map $\varphi : L(E) \to k(P)$ by $\varphi(f) = (tf)(P)$. This is well defined since $tf \in \mathcal{O}_P$, and so $(tf)(P) = \overline{tf} \in \mathcal{O}_P/P = k(P)$ is defined. We show that φ is a k-linear map, its kernel is $L(D)$, and that it is surjective. This will prove that $L(E)/L(D)$ is isomorphic to a k-subspace of $k(P)$ as k-vector spaces, and so

$$\dim_k(L(E)/L(D)) \leq \dim_k(k(P)) = \deg(P) = \deg(E) - \deg(D).$$

First, if $f, g \in L(E)$, then $\varphi(f+g) = (t(f+g))(P) = (tf+tg)(P) = (tf)(P) + (tg)(P)$, so $\varphi(f+g) = \varphi(f) + \varphi(g)$. Next, if $a \in k$, then $\varphi(af) = (taf)(P) = a(tf)(P) = a\varphi(f)$. The second to last equality holds because a is a constant function; so, $(taf)(P) = a(P)(tf)(P) = a(tf)(P)$. Alternatively, with the identification of k as a subfield of $k(P)$ under the map $a \mapsto \overline{a}$, we have $(taf)(P) = \overline{taf} = \overline{a} \cdot \overline{tf} = \overline{a} \cdot \varphi(f)$. Thus, φ is k-linear. Finally, to see that $\ker(\varphi) = L(D)$, if $f \in L(D)$, then $v_P(f) \geq -v_P(D) = -v_P(E) + 1$. Thus, $v_P(f) > -v_P(E) = -v_P(t)$, so $v_P(tf) > 0$. Therefore, $(tf)(P) = 0$, so $\varphi(f) = 0$. Conversely, if $f \in L(E)$ with $\varphi(f) = 0$, then $(tf)(P) = 0$, so $v_P(tf) > 0$. This yields $v_P(f) > -v_P(t)$, or $v_P(f) \geq 1 - v_P(t) = 1 - v_P(E) = -v_P(D)$. For any $Q \neq P$, we have $v_Q(D) = v_Q(E)$, so since $v_Q(f) \geq -v_Q(E)$, we have $v_Q(f) \geq -v_Q(D)$. All these inequalities say that $f \in L(D)$. This finishes the proof in the case that $E = D + P$.

In the general case, we may write $E = D + P_1 + \cdots + P_n$, with the P_i not necessarily distinct. Set $D_i = D + P_1 + \cdots + P_i$, so $E = D_n$. By the previous paragraph, $\dim_k(L(D_{i+1})/L(D_i)) \leq \deg(P_i)$ for all i. By induction, if $\dim_k(L(D_{i-1})/L(D)) \leq \deg(P_1) + \cdots + \deg(P_{i-1})$, then

$$\begin{aligned}\dim_k(L(D_i)/L(D)) &= \dim_k(L(D_i)/L(D_{i-1})) + \dim_k(L(D_{i-1})/L(D)) \\ &\leq \deg(P_1) + \cdots + \deg(P_{i-1}) + \deg(P_i)\end{aligned}$$

because

$$L(D_i)/L(D_{i-1}) \cong \frac{L(D_i)/L(D)}{L(D_{i-1})/L(D)}$$

and the dimension of a quotient is the difference of the dimensions of the terms. So, induction shows that $\dim_k(L(E)/L(D)) \leq \sum_{i=1}^{n} \deg(P_i) = \deg(E) - \deg(D)$. □

If D is a divisor, we may write $D = D_+ - D_-$, where D_+ is the sum of the terms in D with positive coefficients and D_- is the negative of the sum of the terms with negative coefficients. Then D_+ and D_- are both positive divisors. The following result shows that the spaces $L(D)$ are actually finite dimensional over k.

Proposition 4.17. *Let E be a divisor. Then $\dim(E) \leq \deg(E_+) + 1$. Therefore, $L(E)$ is a finite dimensional k-vector space.*

Proof. We have $E_+ - E = E_-$, a positive divisor, so $E \leq E_+$. Therefore, $L(E) \subseteq L(E_+)$. It suffices to prove that $\dim(E_+) \leq \deg(E_+) + 1$. Therefore, by replacing E by E_+, we may assume that E is a positive divisor. By the previous lemma applied to E and $D = 0$, we have $\dim_k(L(E)/L(0) \leq \deg(E) - \deg(0)$. However, $L(0) = k$, so $\dim(0) = 1$. Also, $\deg(0) = 0$. Thus, $\dim_k(L(E)) = \dim_k(L(E)/L(0)) + 1 \leq \deg(E) + 1$, as desired. □

This result shows that, for a positive divisor E, we have $\dim(E) - \deg(E) \leq 1$. In fact, if D is any divisor with $\deg(D) > 0$, the same inequality holds. For, if $L(D) = 0$, then $\dim(D) - \deg(D) = -\deg(D) < 0$. However, if $f \in L(D)$ is nonzero, then $E = D + (f)$ is positive. Since $\dim(E) = \dim(D)$ and $\deg(E) = \deg(D)$, the inequality $\dim(E) - \deg(E) \leq 1$ says that $\dim(D) - \deg(D) \leq 1$. Written another way, $0 \leq 1 + \deg(D) - \dim(D)$. This gives a lower bound for this expression. We will now show that there is an upper bound for this expression, as D ranges over all divisors.

Lemma 4.18. *There is a constant γ, depending only on F/k, such that if E is any divisor of F/k, then $\deg(E) - \dim(E) \leq \gamma$. In particular, $1 + \deg(E) - \dim(E)$ is bounded above by $\gamma + 1$ for any divisor E.*

Proof. We give a technical looking argument to prove this lemma. Let $x \in F$ be nonconstant, and set $B = (x)_\infty$. If $n = [F : k(x)]$, then Theorem 4.5 shows that $\deg(B) = n$. Let $u_1, \ldots, u_n$ be a $k(x)$-basis for F. Let C be any positive divisor containing $u_1, \ldots, u_n$. To see that such a divisor exists, let $\{P_1, \ldots, P_r\}$ be a set of points containing all the poles of all the u_i. For each j, let $e_j \geq \max\{-v_{P_j}(u_i) : 1 \leq i \leq n\}$. Then $v_{P_j}(u_i) \geq -e_j$ for all i. Therefore, if $C = \sum_j e_j P_j$, then $u_i \in L(C)$ for each i. We next claim that, for any positive integer l, the divisor $lB + C$ satisfies $\dim(lB + C) \geq (l+1)n$. To see this, we note that $x^i u_j \in L(lB + C)$ for all j and for all i with $0 \leq i \leq l$. For, $(x^i u_j) + lB + C = i(x) + (u_j) + lB + C$. Now, $(u_j) + C \geq 0$. Furthermore, $i(x) + lB = i(x)_0 - i(x)_\infty + l(x)_\infty = i(x)_0 + (l-i)(x)_\infty$. Since $(x)_0$ and $(x)_\infty$ are positive, and $0 \leq l - i$, the divisor $i(x) + lB$ is positive, and so $(x^i u_j) + lB + C$ is positive. There are $(l+1)n$ elements in the set $\{x^i u_j : 0 \leq i \leq l,\ 1 \leq j \leq n\}$, and these elements are k-linearly independent since x is transcendental over k and the u_j are $k(x)$-linearly independent. This proves our claim that $\dim(lB + C) \geq (l+1)n$. On the other hand, Proposition 4.17 and Lemma 4.16 applied to $E = lB + C$ and $D = C$ shows that $\dim(lB + C) - \dim(lB) \leq \deg(C)$. These two inequalities show that

$$(l+1)n \leq \dim(lB + C) \leq \dim(lB) + \deg(C).$$

Written another way, as $\deg(lB) = l \deg(B) = ln$,

$$\deg(lB) - \dim(lB) \leq \gamma$$

if $\gamma = n - \deg(C)$. This holds for all positive integers l.

Let A be any divisor. Our final claim is that there are divisors A_1 and D such that

$A_1 \geq A$, $A_1 \sim D$, and $D \leq lB$ for some l. Given this claim, we have

$$\begin{aligned}\deg(A) - \dim(A) &\leq \deg(A_1) - \dim(A_1) \\ &= \deg(D) - \dim(D) \\ &\leq \deg(lB) - \dim(lB) \\ &\leq \gamma\end{aligned}$$

where the first and third lines follow from Lemma 4.16, the second line holds since A_1 and D are equivalent, and the final line comes from the inequality above. The only thing remaining is then to prove the claim. Choose A_1 any divisor with $A_1 \geq A$. By Lemma 4.16, we have

$$\begin{aligned}\dim(lB - A_1) &\geq \dim(lB) - \deg(A_1) \\ &\geq \deg(lB) - \gamma - \deg(A_1) \\ &= nl - \gamma - \deg(A_1) \geq 0\end{aligned}$$

if l is large enough, since $n \geq 1$. This finishes the proof of the lemma. □

This lemma shows that $\max\{\deg(D) - \dim(D) + 1 : D \in D_F\}$ exists. We define the *genus* of F/k to be the integer $g = \max\{\deg(D) - \dim(D) + 1 : D \in D_F\}$. By definition, we have $g \geq \deg(0) - \dim(0) + 1$. Since $\deg(0) = 0$ and $\dim(0) = 1$, we see that $g \geq 0$. The definition (and existence) of g is usually stated as Riemann's theorem.

Theorem 4.19 (Riemann's Theorem). *Let F/k be an algebraic function field in one variable and let g be its genus.*

1. *If D is any divisor of F/k, then* $\dim(D) \geq \deg(D) + 1 - g$.
2. *There is a constant c, depending only on F/k, such that if* $\deg(D) \geq c$, *then* $\dim(D) = \deg(D) + 1 - g$.

Proof. The first statement is just the definition of g. To see the second, by the definition of g, there is a divisor A with $g = \deg(A) - \dim(A) + 1$. Let $c = \deg(A) + g$. If $\deg(D) \geq c$, then

$$\begin{aligned}\dim(D - A) &\geq \deg(D - A) + 1 - g \geq \deg(D) - \deg(A) + 1 - g \\ &= \deg(D) + 1 - c \geq 1.\end{aligned}$$

Therefore, $L(D - A)$ is nonzero. Let $x \in L(D - A)$ be nonzero. Set $D' = D + (x)$. Then $D' \geq A$ since $(x) + D - A \geq 0$. Then

$$\deg(D) - \dim(D) = \deg(D') - \dim(D') \geq \deg(A) - \dim(A) = g - 1.$$

Therefore, $\dim(D) \leq \deg(D) + 1 - g$. Since the reverse inequality holds by definition of g, we have $\dim(D) = \deg(D) + 1 - g$. □

4.2 The Riemann-Roch Theorem

Riemann's theorem is a very useful fact about divisors, but because it gives an inequality, it does not always allow us to determine the dimension of a divisor. The Riemann-Roch theorem is an improvement of this result.

Theorem 4.20 (Riemann-Roch). *Let F/k be an algebraic function field in one variable. If C is a canonical divisor of F/k, then for any divisor D, we have* $\dim(D) = \deg(D) + 1 - g + \dim(C - D)$.

To save time and tedious details, we aren't going to prove this theorem in class. However, we will define what is a canonical divisor. To do this we discuss derivations and differentials. We point out that Section 3 of Chapter 4 of Stichtenoth gives the relation between differentials in the sense we will introduce with the differentials he defines in Chapter 1. A *differential* of F/k is a F-linear combination of formal symbols df, with $f \in F$. These symbols satisfy the properties $da = 0$ if $a \in k$, $d(f + g) = df + dg$, and $d(fg) = gdf + fdg$. The F-vector space of differentials is denoted $\Omega_{F/k}$. We formally define $\Omega_{F/k}$ as the quotient of the F-vector space with basis all symbols of the form df for $f \in F$, modulo the submodule generated by all elements of the form da for $a \in k$, and all elements of the form $d(f + g) - df - dg$, and all elements of the form $d(fg) - gdf - fdg$. For the rest of this discussion, we assume that F contains an element $x \in F$ transcendental over k for which $F/k(x)$ is a separable extension. We state without proof that if k is a perfect field, then such an element exists. In particular, if k is a finite field, such an element exists. We need to know that $\Omega_{F/k}$ is a 1-dimensional F-vector space. It is fairly easy to prove that its dimension is at most 1 (and we do so below); to see that it is nonzero, we need a little more information about $\Omega_{F/k}$. First, we define a related notion. If M is an F-vector space, then a *k-derivation* is a function $D : F \to M$ is a k-linear function such that $D(ab) = D(a)b + aD(b)$ for all $a, b \in F$. We point out the universal mapping property for $\Omega_{F/k}$: if $D : F \to M$ is a derivation, then there is a unique F-linear map $\varphi : \Omega_{F/k} \to M$ such that $D(x) = \varphi(dx)$ for all $x \in F$. The proof of this fact is not hard, although we do not give it.

Proposition 4.21. *With x as above, the F-vector space $\Omega_{F/k}$ is one-dimensional with basis dx.*

Proof. By assumption, $F/k(x)$ is separable. Let $t \in F$. If $p(T) \in k(x)[T]$ is the minimal polynomial of t over $k(x)$, then $p(T)$ has no repeated roots. Let $p(T) = \sum_{i=0}^{n} a_i(x)T^i$. Then $0 = \sum_{i=0}^{n} a_i(x)t^i$. Applying the defining rules for $\Omega_{F/k}$, we see that $d(a(x)) = a'(x)dx$, where $a'(x)$ is the derivative of $a(x)$. Thus, we have

$$0 = \sum_{i=0}^{n} a_i'(x)t^i dx + \left(\sum_{i=0}^{n} a_i(x) i t^{i-1} \right) dt = \sum_{i=0}^{n} a_i'(x)t^i dx + p'(t)dt.$$

Since $p(T)$ has no repeated roots, $p'(t) \neq 0$. Therefore, we may solve for dt, obtaining

$$dt = \frac{-\sum_{i=0}^{n} a_i'(x)t^i}{p'(t)}dx.$$

Therefore, $dt \in Fdx$. Since $\Omega_{F/k}$ is spanned by dt for $t \in F$, this proves that $\Omega_{F/k} = Fdx$. To finish the proof, we need to show that $\Omega_{F/k}$ is nonzero. To do this we produce a nonzero derivation $D : F \to F$. We have the derivation $\frac{d}{dx}$ on $k(x)$. The argument above actually shows that $\frac{d}{dx}$ extends uniquely to a derivation on F. For, if $t \in F$ and $p(T)$ is its minimal polynomial over $k(x)$, then we define D by the formula

$$D(t) = \frac{-\sum_{i=0}^{n} a_i'(x)t^i}{p'(t)}.$$

□

Let x be as above, and let $\omega = fdx$ be a differential. We then define the divisor (ω) of the differential ω as follows. We first assume that $\omega = dx$. Let $P \in \mathbf{P}_F$, and let t be a generator of the ideal P such that $dt \neq 0$. We point out that such a t exists. For, if t is any generator of P, if $dt \neq 0$, then we are done. If $dt = 0$, write $x = t^n u$ for some $n \in \mathbb{Z}$ and unit u. Then $dx = nt^{n-1}udt + tdu = tdu$ since $dt = 0$. So, $du \neq 0$. Then $d(tu) = udt + tdu = tdu \neq 0$, and so tu is a generator of P whose differential is nonzero. Write $dx = fdt$ for some $f \in F$. We write $f = dx/dt$. Then the coefficient of P in (dx) is $v_P(dx/dt)$. In other words, if we write $(dx) = \sum_P v_P(dx)P$, then $v_P(dx) = v_P(dx/dt)$. We note without proof that if $f \in \mathcal{O}_P$, then $df/dt \in \mathcal{O}_P$. From this fact it follows that the definition of (dx) is well defined. For, if t' is another generator of P, write $t' = tu$ for some unit u. Then $dt'/dt = u + tdu/dt \in \mathcal{O}_P$ by the previous line. Similarly, $dt/dt' \in \mathcal{O}_P$. Since dt'/dt and dt/dt' are inverses, each is a unit. Finally, $dx/dt' = dx/dt \cdot dt/dt'$ differs from dx/dt by the unit dt/dt', so $v_P(dx/dt) = v_P(dx/dt')$. For a general differential $\omega = fdx$, we set $(\omega) = (f) + (dx)$. Therefore, any two differentials have equivalent divisors, so the class of the divisor of a differential is uniquely determined. Any such divisor is called a *canonical divisor*.

Example 4.22. Let $F = k(x)$. We compute a canonical divisor of F/k. We can use x since F is obviously separable over $k(x)$. First, consider a point P that is the maximal ideal of $k[x]_{(p(x))}$ for some irreducible polynomial $p(x)$. Then $p(x)$ is a generator of the maximal ideal. We have $dp(x) = p'(x)dx$. Now, since $p'(x)$ is not divisible by $p(x)$, we see that $v_{p(x)}(p'(x)) = 0$. Thus, $v_{p(x)}(dx) = 0$. This handles all but one point. The remaining point is P_∞, the point at infinity. The element x^{-1} is a generator of this maximal ideal. Since $dx^{-1} = -x^{-2}dx$ and $v_\infty(f) = -\deg(f)$, we see that $v_\infty(dx) = -v_\infty(x^{-2}) = \deg(x^{-2}) = -2$. Therefore, the canonical divisor (dx) is $-2P_\infty$. Since we see that $\deg(-2P_\infty) = -2$ and its degree is $2g - 2$, as we will see in the next corollary, this yields $g = 0$. We will give another proof of this in the next section.

We give some elementary consequences of the Riemann-Roch theorem.

Corollary 4.23. *Let C be a canonical divisor of F/k. Then* $\deg(C) = 2g-2$ *and* $\dim(C) = g$.

Proof. The Riemann-Roch theorem applied to $D = 0$ yields $1 = 1 - g + \dim(C)$. Therefore, $\dim(C) = g$. Next, applying the theorem to $D = C$ yields $\dim(C) = \deg(C)+1-g+\dim(0)$, or $g = \deg(C) + 1 - g + \dim(0)$, which then gives $\deg(C) = 2g - 2$. □

These formulas for the degree and the dimension uniquely determine the divisor class of C, as the next corollary shows.

Corollary 4.24. *A divisor D is a canonical divisor if and only if* $\deg(D) = 2g - 2$ *and* $\dim(D) \geq g$.

Proof. Let D be a divisor with $\deg(D) = 2g-2$ and $\dim(D) \geq g$. If C is a canonical divisor, we need to show that $D \sim C$. By the Riemann-Roch theorem, $\dim(D) = \deg(D) + 1 - g + \dim(C-D)$. Since $\dim(D) \geq g$ and $\deg(D) = 2g-2$, we get $g \leq 2g-2+1-g+\dim(C-D)$, or $1 \leq \dim(C - D)$. Since $\deg(C - D) = \deg(C) - \deg(D) = 0$, Corollary 4.7 shows that $C - D$ is principal. Thus, $D \sim C$. Note that this forces $\dim(D) = g$. □

We can improve on the second statement of Riemann's theorem.

Corollary 4.25. *If D is a divisor with* $\deg(D) > 2g - 2$, *then* $\dim(D) = \deg(D) + 1 - g$.

Proof. Suppose that $\deg(D) > 2g - 2$. Then $\deg(C - D) < 0$, so $\dim(C - D) = 0$. The corollary then follows from the Riemann-Roch theorem. □

We will not say very much about how one can find the genus of a function field; in fact, it is difficult in general to calculate the genus. We do state without proof one fact: If F/k is the function field of the affine curve $Z(f(x, y))$ for some polynomial $f(x, y) \in k[x, y]$ of degree d. Then the genus g of F/k satisfies $g \leq \frac{1}{2}(d - 1)(d - 2)$, and equality holds if the projective version of this affine curve is nonsingular. We will use this below to show that the function field of an elliptic curve has genus 1.

4.3 Fields of Genus 0 and 1

In this section we classify fields of genus 0 and 1. We first consider genus 0. To start, let us see a second argument for why the genus of the rational function field $k(x)/k$ is 0. Let $P = P_\infty$, a point of degree 1. The valuation v_P corresponding to $\mathcal{O}_P$ is given by $v_P(\varphi(x)) = -\deg(\varphi(x))$. If $D = nP$, then by Riemann's theorem, if n is large enough, then $\dim(D) = \deg(D) + 1 - g$. Since $\deg(D) = n$, it is enough to calculate $L(D)$. As we saw in an earlier example, the space $L(D)$ is the set of all $f \in k[x]$ with $\deg(f(x)) \leq n$, which has dimension $n + 1$. Then $g = \deg(D) - \dim(D) + 1 = 0$. Note that P_∞ is a rational point of $k(x)/k$.

For a partial converse, suppose that F/k has genus 0. Recall that we are assuming that k is the exact field of constants. Moreover, suppose that there is a *rational point* $P \in \mathbf{P}_F$. That is, suppose that $\deg(P) = 1$. By Riemann's theorem, $\dim(P) \geq \deg(P) + 1 = 2$. Let $x \in L(P)$ be nonconstant. Then x exists since the set of constants has dimension 1 and $\dim(L(P)) \geq 2$. Because $x \in L(P)$, we have $v_P(x) \geq -v_P(P) = -1$ and $v_Q(x) \geq 0$ for all $Q \neq P$. As the nonconstant x must have a pole, we see that $v_P(x) = -1$. Then $x^{-1} \in P$. By Theorem 4.5, $[F : k(x)] = \deg(x)_\infty$. However, since P is the only pole of x, and since $v_P(x) = -1$, the pole divisor $(x)_\infty = P$. So, $\deg(x)_\infty = \deg(P) = 1$, which shows that $[F : k(x)] = 1$. Therefore, $F = k(x)$, so F is a rational function field in one variable over k.

The assumption that F has a rational point is necessary. Let $k = \mathbb{R}$ and let F be the function field of the curve $x^2 + y^2 + 1 = 0$ over k. Then $F = k(x)(\sqrt{-1 - x^2})$. It can be proven that F/k has genus 0 but that F has no rational point. Thus, F is not isomorphic to a rational function field in one variable over k. However, if $k = \mathbb{C}$, then the genus of F/k is also 0, but any point is a rational point, so $k(x)(\sqrt{-1 - x^2})$ is isomorphic to a rational function field in one variable. In particular, an exercise will show that F is generated over k by $(i - x)^{-1}\sqrt{-1 - x^2}$.

We now consider fields of genus 1. We restrict to the case $\mathrm{char}(k) \neq 2$. We will show that a function field F/k with a rational point has genus 1 if and only if $F = k(x)(y)$ for some $y \in F$ with $y^2 = f(x)$ for some cubic polynomial $f(x)$ with no repeated roots in k. This means that such a field is precisely the function field of an elliptic curve. First, suppose that $F = k(x, y)$, where x is transcendental over k, and with $y^2 = f(x)$ for some cubic $f(x)$ with no repeated roots. We first show that F/k has a rational point P. Consider the extension $F/k(x)$, a quadratic extension. Let P be any point for which $\mathcal{O}_P$ extends the valuation ring $k[x^{-1}]_{(x^{-1})}$ of $k(x)$. Let e be the ramification index and f the residue degree. Since the residue field of $k[x^{-1}]_{(x^{-1})}$ is k, we see that $f = \deg(P)$. Thus, to show that P is a rational point, we need to show that $f = 1$. Recall from the Dedekind domain handout that $ef \leq [F : k(x)] = 2$. Thus, if we prove that $e = 2$, then $f = 1$ is forced upon us. To see this, let v_P be the valuation corresponding to $\mathcal{O}_P$. Then $v|_{k(x)} = ev_\infty$, where v_∞ is the valuation on $k(x)$ corresponding to $k[x^{-1}]_{(x^{-1})}$. Since $v_\infty(\varphi(x)) = -\deg(\varphi(x))$, and $y^2 = f(x)$, we have $v_\infty(f(x)) = -3$. So, $2v(y) = v(y^2) = -3e$. For $v(y) \in \mathbb{Z}$, this forces $e = 2$, as desired. To see that he genus is 1, let $g(x, y) = y^2 - f(x)$. Then g is a polynomial of degree 3. Its homogenization is $y^2z - z^3f(x/z)$. We have stated that the corresponding projective variety is nonsingular since $f(x)$ has no repeated roots. Therefore, by the genus formula given earlier, $g = \frac{1}{2}(3-1)(3-2) = 1$. Alternatively, we could compute the genus by computing the degree of a canonical divisor (dx). If we do so, we would see that $\deg(dx) = 0$. Since this degree is $2g - 2$, we obtain $g = 1$.

Now suppose that F/k has genus 1 and that F/k has a rational point P. We produce $x, y \in F$ such that x is transcendental over k and $y^2 = f(x)$ for some cubic $f(x)$ with no repeated roots in k. By the Riemann-Roch theorem, if $\deg(D) > 2g - 2 = 0$, then $\dim(D) = \deg(D) + 1 - g = \deg(D)$. Therefore, $\dim(nP) = n$ if $n > 0$. Furthermore, $L(nP) \subseteq L((n+1)P)$ since $(n+1)P \geq nP$. There are then elements $x \in L(2P) \setminus L(P)$ and

$y \in L(3P) \setminus L(2P)$. So, for all $Q \neq P$, we have $v_Q(x) \geq 0$ and $v_Q(y) \geq 0$, while at P we have $v_P(x) = -2$ and $v_P(y) = -3$. The element x is transcendental over k since it is not constant (as $v_P(x) \neq 0$). Since P is the only pole of x or y, we see that $(x)_\infty = 2P$ and $(y)_\infty = 3P$. From Theorem 4.5, this implies that $[F : k(x)] = 2$ and $[F : k(y)] = 3$. So, $F/k(x)$ and $F/k(y)$ have no intermediate fields, since these degrees are prime. Therefore, $F = k(x, y)$. We wish to show that y satisfies a cubic over $k(x)$. To see this, note that the seven elements $1, x, x^2, x^3, y, xy, y^2$ all lie in $L(6P)$. Since $\dim(6P) = 6$, these elements are linearly dependent over k. There are constants with $ay^2 + bxy + cy + dx^3 + ex^2 + hx + j = 0$. At least one of a, b, c is nonzero since x is transcendental over k. From this it follows that $a \neq 0$ else $y \in k(x)$, which is not true. By dividing by a, we may assume that $a = 1$. Similarly, $d \neq 0$ else x satisfies a quadratic over $k(y)$, which is impossible since $F = k(y)(x)$ and $[F : k(y)] = 3$. By completing the square, we have $(y + \frac{1}{2}(bx + c))^2 - f(x) = 0$, where $f(x)$ is a cubic in x. Finally, if we replace y by $y + \frac{1}{2}(bx + c)$, then F is still equal to $k(x, y)$, and $y^2 = f(x)$. It remains to show that $f(x)$ has no repeated roots. If, instead, we can write $f(x) = \alpha(x - \beta)^2(x - \gamma)$ for some $\alpha, \beta, \gamma \in k$, then $(y/(x - \beta))^2 = \alpha(x - \gamma)$. Replacing y by $y/(x - \beta)$, we still have that x, y generate F over k. However, since $y^2 = \alpha(x - \gamma)$, we see that $x \in k(y)$. Therefore, $F = k(y)$, and so F is a rational function field in one variable over k. However, this would yield $g = 0$, while we are assuming that $g = 1$. Thus, $f(x)$ has no repeated roots in k.

5 Goppa Codes

We now put the Riemann-Roch theorem to work to define a class of codes. These codes were discovered by Goppa in 1981. Let $F/\mathbb{F}_q$ be an algebraic function field in one variable. We assume that $\mathbb{F}_q$ is the exact constant field. We denote the genus of $F/\mathbb{F}_q$ by g. Let $P_1, \ldots, P_n$ be rational points in $\mathbf{P}_F$. Recall that P is a rational point if $\deg(P) = 1$. Thus, the residue field of a point is $\mathbb{F}_q$ if and only if the point is a rational point. Consider the divisor $D = P_1 + \cdots + P_n$, and let G be a divisor whose support is disjoint from $\{P_1, \ldots, P_n\}$. Define an evaluation function $\mathrm{ev}_D : L(G) \to \mathbb{F}_q^n$ by $\mathrm{ev}_D(f) = (f(P_1), \ldots, f(P_n))$. To see that this is well defined, if $f \in L(G)$, then $v_{P_i}(f) \geq 0$ since P_i is not in the support of G. Therefore, $f(P_i)$ is defined and is an element of $\mathbb{F}_q = \mathbb{F}_q(P_i)$ for each i. It is easy to see that ev_D is $\mathbb{F}_q$-linear. For addition, we have

$$\begin{aligned}\mathrm{ev}_D(f+g) &= ((f+g)(P_1), \ldots, (f+g)(P_n)) \\ &= (f(P_1) + g(P_1), \ldots, f(P_n) + g(P_n) \\ &= (f(P_1), \ldots, f(P_n)) + (g(P_1), \ldots, g(P_n)) \\ &\quad \mathrm{ev}_D(f) + \mathrm{ev}_D(g).\end{aligned}$$

If $a \in \mathbb{F}_q$, then

$$\begin{aligned}\mathrm{ev}_D(af) &= ((af)(P_1), \ldots, (af)(P_n)) \\ &= (a(P_1)f(P_1), \ldots, a(P_n)f(P_n)) \\ &= (af(P_1), \ldots, af(P_n)) = a\,\mathrm{ev}_D\,\varphi(f)\end{aligned}$$

since a is a constant function. The image of ev_D is then a $\mathbb{F}_q$-subspace of $\mathbb{F}_q^n$.

Definition 5.1. *With notation as above, the Goppa code $C_L(D, G)$ is the image of ev_D. That is, $C_L(D, G) = \{(f(P_1), \ldots, f(P_n)) : f \in L(G)\}$.*

From basic facts about divisors, we can determine the parameters of a Goppa code.

Theorem 5.2. *The parameters of the Goppa code $C_L(D, G)$ has length $n = \deg(D)$, dimension $k = \dim(G) - \dim(G - D)$, and distance $d \geq n - \deg(G)$.*

Proof. The length of the code is the number n of rational points $P_1, \ldots, P_n$. Since $D = \sum_{i=1}^n P_i$ and each P_i has degree 1, we see that $n = \deg(D)$. To determine the dimension, since ev_D is linear, we see that $k = \dim(L(G)) - \dim(\ker(\mathrm{ev}_D))$. We show that $\ker(\mathrm{ev}_D) = L(G - D)$. Let $f \in \ker(\mathrm{ev}_D)$. Then $f(P_i) = 0$ for all i. Then $v_{P_i}(f) \geq 1$, so $v_{P_i}(f) \geq -v_{P_i}(G - D) = 1$. If Q is any other point, then $v_Q(f) \geq -v_Q(G) = -v_Q(G - D)$ since $f \in L(G)$. Therefore, we see that $f \in L(G-D)$. Conversely, if $f \in L(G-D)$, then $f \in L(G)$ since $G-D \leq G$, and $f(P_i) = 0$ for all i since $v_{P_i}(f) \geq -v_{P_i}(G-D) = 1$. Thus, $f \in \ker(\mathrm{ev}_D)$. We have then shown that $\ker(\mathrm{ev}_D) = L(G - D)$, and so $k = \dim(L(G) - \dim(L(G - D)) = \dim(G) - \dim(G - D)$. Finally, let $\mathrm{ev}_D(f) \in C_L(D, G)$ have weight d. Then $n - d$ of the

coordinates of $\mathrm{ev}_D(f)$ are zero. If these zeros are at the points $P_{i_1}, \ldots, P_{i_{n-d}}$, then we see that $f \in L(G - P_{i_1} - \cdots - P_{i_{n-d}})$. This space is then nonzero, so the degree of $G - P_{i_1} - \cdots - P_{i_{n-d}}$ is nonnegative. In other words, $\deg(G) - (n-d) \geq 0$, so $d \geq n - \deg(G)$, as desired. □

The Riemann-Roch theorem will give us more information about the parameters of this code.

Corollary 5.3. *Suppose that* $\deg(G) < n$. *Then* ev_D *is injective, and so* $k = \dim(G) \geq \deg(G) + 1 - g$ *and* $d \geq n - \deg(G)$. *Furthermore, if* $n > \deg(G) > 2g - 2$, *then* $k = \deg(G) + 1 - g$.

Proof. We are claiming nothing new about d. If $\deg(G) < n$, then $\deg(G - D) < 0$, so $L(G - D) = 0$. Thus, $\ker(\mathrm{ev}_D) = 0$, so ev_D is injective. This implies that $k = \dim(G)$. From Riemann's theorem, we have $\dim(G) \geq \deg(G) + 1 - g$, which yields $k \geq \deg(G) + 1 - g$. Finally, if $n > \deg(G) > 2g - 2$, then the Riemann-Roch theorem yields $\dim(G) = \deg(G) + 1 - g$, so $= \deg(G) + 1 - g$. □

As an immediate consequence, if $2g - 2 < \deg(G) < n$, then $k + d \geq n + 1 - g$. Therefore, if $F = \mathbb{F}_q(x)$ is a rational function field over $\mathbb{F}_q$, then $g = 0$, and so $k + d \geq n + 1$. However, the Singleton bound says that $k + d \leq n + 1$. Therefore, $k + d = n + 1$. In other words, Goppa codes associated to $\mathbb{F}_q(x)/\mathbb{F}_q$ are MDS codes; that is, codes that attain the Singleton bound. The only problem with this is that there are only $q + 1$ rational points of $\mathbb{F}_q(x)/\mathbb{F}_q$; these are the points corresponding to the discrete valuation rings $k[x]_{(x-a)}$ for $a \in \mathbb{F}_q$ along with $k[x^{-1}]_{(x^{-1})}$. So, if $q = 2$, we can only build Goppa codes of length at most 3 if we use $F = \mathbb{F}_2(x)$.

Proposition 5.4. *Suppose that* $\deg(G) < n$. *If* $\{x_1, \ldots, x_k\}$ *is a basis for* $L(G)$, *then*

$$\begin{pmatrix} x_1(P_1) & x_1(P_2) & \cdots & x_1(P_n) \\ x_2(P_1) & x_2(P_2) & \cdots & x_2(P_n) \\ \vdots & \vdots & \vdots & \vdots \\ x_k(P_1) & x_k(P_2) & \cdots & x_k(P_n) \end{pmatrix}$$

is a generator matrix for $C_L(D, G)$.

Proof. Recall that a generator matrix for a code is a $k \times n$ matrix whose rows form a basis for the code. Since $\deg(G) < n$, the map $\mathrm{ev}_D : L(G) \to \mathbb{F}_q^n$ is injective. Therefore, if $\{x_1, \ldots, x_k\}$ is a basis for $L(G)$, then $\{\mathrm{ev}_D(x_1), \ldots, \mathrm{ev}_D(x_n)\}$ is a basis for the image of ev_D, which is $C_L(D, G)$. The i-th row of the matrix above is simply the vector $\mathrm{ev}_D(x_i)$, so the matrix is indeed a generator matrix for the Goppa code. □

Definition 5.5. *The integer* $d^* = n - \deg(G)$ *is called the designated distance of the code* $C_L(D, G)$. *It is a lower bound for the actual distance of the code.*

5.1 Goppa Codes coming from $\mathbb{F}_q(x)/\mathbb{F}_q$

In this section we will relate BCH codes and Reed Solomon codes to Goppa codes. We will see that such codes arise from Goppa codes associated to the function field $\mathbb{F}_q(x)$. First note that there are $q+1$ rational points of $\mathbb{F}_q(x)/\mathbb{F}_q$; these correspond to the discrete valuation rings $\mathbb{F}_q[x]_{(x-a)}$ for $a \in \mathbb{F}_q$ and to $\mathbb{F}_q[x^{-1}]_{(x^{-1})}$. Therefore, any Goppa code constructed from this function field is limited to have its length n satisfy $n \leq q+1$. In particular, if we work with $\mathbb{F}_2$, we can have codes of length at most 3. To avoid repeatedly making reference to the function field $\mathbb{F}_q(x)/\mathbb{F}_q$, we call a Goppa code a *rational Goppa code* if it is associated to $\mathbb{F}_q(x)/\mathbb{F}_q$.

Recall our standard notation for the parameters of a code; n is the length of the code, k is the dimension of the code, and d is the distance of the code. We note what are the parameters of a rational Goppa code in terms of the degree of G in the next proposition.

Proposition 5.6. *Let $C_L(D,G)$ be a rational Goppa code with parameters n, k, and d. Then $k=0$ if and only if $\deg(G)<0$, and $k=n$ if and only if $\deg(G)>n-2$. Furthermore, if $0 \leq \deg(G) \leq n-2$, then $k=1+\deg(G)$ and $d=n-\deg(G)$. In particular, $C_L(D,G)$ is an MDS code.*

Proof. If $\deg(G)<0$, then $L(G)=0$, so $C_L(D,G)=0$. If $\deg(G)>n-2$, then $\deg(G-D)>-2=2g-2$, where $g=0$ is the genus of $\mathbb{F}_q(x)/\mathbb{F}_q$. Thus, by the Riemann-Roch theorem,

$$\dim(G-D)=\deg(G-D)+1-g=\deg(G-D)+1=\deg(G)-n+1.$$

Since $k=\dim(G)-\dim(G-D)$, and $\dim(G)=\deg(G)+1-g$, also by Riemann-Roch, we see that $k=\deg(G)+1-(\deg(G)-n+1)=n$. Finally, if $0 \leq \deg(G) \leq n-2$, then $\dim(G-D)=0$ since $\deg(G-D) \leq -2 < 0$. Thus, $k=\dim(G)=\deg(G)+1$ by Riemann-Roch. Also, $d \geq n-\deg(G)$. Since $k+d \leq n+1$ by the Singleton bound, this forces $d=n-\deg(G)$. The code is then an MDS code since the Singleton bound is met. □

Recall the definition of a Reed-Solomon code over $\mathbb{F}_q$. Let $n=q-1$, and let α be a primitive element of $\mathbb{F}_q$. We identify the vector space $\mathbb{F}_q^n$ with the vector space of polynomials of degree less than n under the association $(a_0,\ldots,a_{n-1}) \mapsto a_0+\cdots+a_{n-1}x^{n-1}$. The code consisting of all polynomials $p(x)$ of degree less than n with $p(\alpha)=p(\alpha^2)=\cdots=p(\alpha^{d-1})=0$ is a Reed-Solomon code of dimension $n+1-d$. As we will see, this is (almost) a special example of a Goppa code. In fact, we look at a larger class of codes, which are called generalized Reed-Solomon codes. To motivate the definition, we look at the original definition of Reed and Solomon, which is different than the definition we gave. Set $n=q$, and let α be a primitive element of $\mathbb{F}_q$. Set $\alpha_i=\alpha^i$ for $0 \leq i \leq q-2$, and set $\alpha_{q-1}=0$. Pick an integer $k<n$ and let L be the vector space of polynomials in $\mathbb{F}_q[x]$ of degree less than k. Then the code

$$C=\{(f(\alpha_0),\ldots,f(\alpha_{q-1})) : f \in L\}$$

is the type originally defined by Reed and Solomon. This is a code of length $n = q$ and dimension k; we note that the linear transformation $f \mapsto (f(\alpha_0), \ldots, f(\alpha_{q-1}))$ is injective since $k < n$ and a nonzero polynomial of degree $< k$ cannot have n roots. Thus, $\dim(C) = \dim(L) = k$. Moreover, given any nonzero f, since f has at most $k-1$ roots, at least $n-k+1$ of the components of the vector $(f(\alpha_0), \ldots, f(\alpha_{q-1}))$ are nonzero. This says $d \geq n - k + 1$. The Singleton bound shows that $d = n-k+1$, and so this code is an MDS code. To generalize this definition, let $v = (v_1, \ldots, v_n) \in \mathbb{F}_q^n$ have all nonzero entries, let $\alpha = (\alpha_1, \ldots, \alpha_n)$ have distinct components in $\mathbb{F}_q$, and consider the code

$$\mathrm{GRS}_k(\alpha, v) = \{(v_1 f(\alpha_1), \ldots, v_n f(\alpha_n) : f \in L\}.$$

This is called a generalized Reed-Solomon code. If $\mathbf{1} = (1, \ldots, 1)$ and $\alpha = (\alpha_0, \ldots, \alpha_{q-1})$ as above, then this Reed-Solomon code is the code originally defined by Reed and Solomon. To see the connection between these codes and the Reed-Solomon codes we defined earlier, we first define the *extended code* $\overline{C}$ of a code C. This is the code

$$\overline{C} = \left\{ (c_1, \ldots, c_{n+1}) \in \mathbb{F}_q^{n+1} : (c_1, \ldots, c_n) \in C, \sum_{i=1}^{n+1} c_i = 0 \right\}.$$

If H is a parity check matrix for C, then the matrix whose top left part is H, whose bottom row has every entry 1, and whose last column has all 0 entries except for the bottom is a parity check matrix for $\overline{C}$. The code $\overline{C}$ has length 1 more than the length of C, but whose dimension is the same as the dimension of C. The distance of $\overline{C}$ is either equal to the distance d of C or to $d+1$, depending on whether or not every word in C of weight d has the sum of its coefficients different than 0. We will now show that an extended Reed-Solomon code is a generalized Reed-Solomon code. Let $\alpha = (\alpha_0, \ldots, \alpha_{q-1})$ with the α_i defined as above. First note that since every element of $\mathbb{F}_q$ is a root of $x^q - x$, every nonzero element is a root of $x^{q-1} - 1 = (x-1)(x^{q-2} + \cdots + x + 1)$. Thus, if $c \in \mathbb{F}_q$, then $\sum_{i=0}^{q-2} c^i = 0$ if $c \neq 0, 1$. Now, let $f(x) = \sum_{j=0}^{k-1} a_j x^j$, and set $c_i = f(\alpha^i) = f(\alpha_i)$ for $0 \leq i \leq q-2$. If $1 \leq l \leq q-k-1$, then

$$\sum_{i=0}^{q-2} c_i (\alpha^l)^i = \sum_{i=0}^{q-2} \left(\sum_{j=0}^{k-1} a_j (\alpha^i)^j \right) (\alpha^l)^i = \sum_{j=0}^{k-1} a_j \sum_{i=0}^{q-2} (\alpha^{l+j})^i = \sum_{j=0}^{k-1} a_j \cdot 0 = 0$$

since $\sum_{i=0}^{q-2} (\alpha^{l+j})^i = 0$ by the comment above since $1 \leq l + j \leq q-2$. Therefore, taking the Reed-Solomon code for $d = q - k$, we see that $\sum_{i=0}^{q-2} c_i x^i$ lies in this code. Under the correspondence between polynomials and n-tuples, this polynomial corresponds to $(c_0, \ldots, c_{q-2}) = (f(\alpha_0), \ldots, f(\alpha_{q-2}))$. So, given $(c_0, \ldots, c_{q-1}) \in \mathrm{GRS}_k(\mathbf{1}, \alpha)$, there is a unique $f(x)$ of degree $< n$ such that $f(\alpha_i) = c_i$, and then $(c_0, \ldots, c_{q-2})$ lies in the Reed-Solomon code of dimension k and distance $q - k$. A similar calculation to that above shows that

$\sum_{i=0}^{q-1} c_i = 0$, since

$$\begin{aligned}\sum_{i=0}^{q-1} c_i &= \sum_{i=0}^{q-1} f(\alpha_i) = f(0) + \sum_{i=0}^{q-2} f(\alpha^i) = \sum_{i=0}^{q-2}\sum_{j=0}^{k-1} a_j(\alpha^i)^j \\ &= a_0 + \sum_{j=0}^{k-1} a_j \sum_{i=0}^{q-2} (\alpha^j)^i = a_0 + a_0\left(\sum_{i=0}^{q-2} 1\right) \\ &= a_0 + (q-1)a_0 = qa_0 = 0.\end{aligned}$$

Thus, $\mathrm{GRS}_k(\mathbf{1}, \alpha)$ is the extended code corresponding to the Reed-Solomon code for $d = q-k$.

We now show that any generalized Reed-Solomon code is a rational Goppa code.

Proposition 5.7. *Every generalized Reed-Solomon code is a rational Goppa code.*

Proof. Let $G = nP_\infty$ and let P_i be the point corresponding to $\mathbb{F}_q[x]_{(x-\alpha_i)}$. Set $D = P_1 + \cdots + P_n$. Then $L(G) = \{f \in \mathbb{F}_q[x] : \deg(f) < k\}$, and so

$$\begin{aligned}C_L(G, D) &= \{(f(P_1), \ldots, f(P_n)) : f \in L(G)\} \\ &= \{(f(\alpha_1), \ldots, f(\alpha_n)) : f \in L\}\end{aligned}$$

is the generalized Reed-Solomon code $\mathrm{GRS}_k(\mathbf{1}, \alpha)$. To deal with a general $v = (v_1, \ldots, v_n)$, find $g \in \mathbb{F}[x]$ with $g(P_i) = v_i$ for each i. (This is possible to do with a polynomial of degree $n-1$; finding such a polynomial amounts to solving an appropriate system of n equation and n unknowns.) Consider now $G = nP_\infty - (g)$. The support of G is disjoint from that of D, since each v_i is nonzero, so no P_i is a zero (or pole) of g. Moreover, $L(G) = gL(nP_\infty)$. Thus, if $f \in L(nP_\infty)$, then $gf \in L(G)$, and so

$$\begin{aligned}\mathrm{ev}_D(gf) &= ((gf)(P_1), \ldots, (gf)(P_n)) = (g(P_1)f(P_1), \ldots, g(P_n)f(P_n)) \\ &= (v_1 f(P_1), \ldots, v_n f(P_n)),\end{aligned}$$

which shows that $C_L(G, D) = \mathrm{GRS}_k(v, \alpha)$. □

We now look at the connection between BCH codes and rational Goppa codes. Recall the definition of a BCH code. We set $n = q^m - 1$ and let α be a primitive element of $\mathbb{F}_{q^m}$. We choose a positive integer $e \le n$, and we consider the code of all polynomials $p(x) \in \mathbb{F}_q[x]$ of degree $< n$ with $p(\alpha) = \cdots = p(\alpha^{e-1}) = 0$. The resulting code has designated distance e, which is a lower bound for the actual distance d. A Reed-Solomon code is nothing but a BCH code when $m = 1$.

To connect these codes to Goppa codes, we need to relate codes over $\mathbb{F}_q$ to codes over $\mathbb{F}_{q^m}$.

Definition 5.8. *Let C be a code of length n over $\mathbb{F}_{q^m}$. Then $C|_{\mathbb{F}_q} := C \cap \mathbb{F}_q^n$ is called a subfield subcode of C, or the restriction of C to $\mathbb{F}_q$.*

In other words, $C|_{\mathbb{F}_q}$ consists of all codewords $(c_1, \ldots, c_n) \in C$ such that each $c_i \in \mathbb{F}_q$. Assuming that $C|_{\mathbb{F}_q}$ is not the trivial code, its minimum distance is at least that of C, and its length is n. Moreover, $\dim_{\mathbb{F}_q}(C|_{\mathbb{F}_q}) \leq \dim_{\mathbb{F}_{q^m}}(C)$; to see this, suppose that $\alpha_1, \ldots, \alpha_m$ is an $\mathbb{F}_q$-basis for $\mathbb{F}_{q^m}$. Take $v_1, \ldots, v_r \in C|_{\mathbb{F}_q}$ that are linearly independent over $\mathbb{F}_q$. If there are $a_i \in \mathbb{F}_{q^m}$ with $\sum_i a_i v_i = 0$, write $a_i = \sum_j x_{ij}\alpha_j$ for some $x_j \in \mathbb{F}_q$. We then have

$$0 = \sum_i a_i v_i = \sum_i \left(\sum_j x_{ij}\alpha_j \right) v_i = \sum_j \alpha_j \left(\sum_i x_{ij} v_i \right).$$

By looking at each component of v_i, which is an element of $\mathbb{F}_q$, we see that since $\{\alpha_1, \ldots, \alpha_m\}$ is an $\mathbb{F}_q$-basis of $\mathbb{F}_{q^m}$, we have $\sum_i x_{ij} v_i = 0$ for each j. However, the $\mathbb{F}_q$-independence of the v_i implies that each $x_{ij} = 0$, and so each $a_i = 0$. Thus, an $\mathbb{F}_q$-basis of $C|_{\mathbb{F}_q}$ is an $\mathbb{F}_{q^m}$-independent set in C, which yields the result about dimensions. It is possible that $\dim_{\mathbb{F}_q}(C|_{\mathbb{F}_q}) < \dim_{\mathbb{F}_{q^m}}(C)$; for example, if C is a 1-dimensional code generated by a vector $(c_1, \ldots, c_n)$, where $c_1/c_2 \notin \mathbb{F}_q$, then for no $\alpha \in \mathbb{F}_{q^m}$ do we have $\alpha c_1 \in \mathbb{F}_q$ and $\alpha c_2 \in \mathbb{F}_q$. Therefore, $C|_{\mathbb{F}_q} = 0$.

Proposition 5.9. *Every extended BCH code is a subfield subcode for some rational Goppa code.*

Proof. Let $n = q^m - 1$, and let α be a primitive element of $\mathbb{F}_{q^m}$. Let

$$C = \left\{ p(x) \in \mathbb{F}_q[x] : \deg(p(x)) < n,\ p(\alpha) = \cdots = p(\alpha^{e-1}) = 0 \right\}$$

be a BCH code. As usual, we associate the n-tuple $(a_0, \ldots, a_{n-1})$ with the polynomial $a_0 + \cdots + a_{n-1}x^{n-1}$. If

$$C' = \left\{ p(x) \in \mathbb{F}_{q^m}[x] : \deg(p(x) < n,\ p(\alpha) = \cdots = p(\alpha^{e-1}) = 0 \right\},$$

then C' is a Reed-Solomon code, and $C = C'|_{\mathbb{F}_q}$. We claim that $\overline{C} = \overline{C'}|_{\mathbb{F}_q}$. First, if $(c_1, \ldots, c_{n+1}) \in \overline{C}$, then $(c_1, \ldots, c_n) \in C = C'|_{\mathbb{F}_q}$, and $c_{n+1} = -\sum_{i=1}^n c_i$. Thus, $(c_1, \ldots, c_{n+1}) \in \overline{C'}$, and so it lies in $\overline{C'}|_{\mathbb{F}_q}$ since $c_{n+1} \in \mathbb{F}_q$. Conversely, if $(c_1, \ldots, c_{n+1}) \in \overline{C'}|_{\mathbb{F}_q}$, then $(c_1, \ldots, c_n) \in C'|_{\mathbb{F}_q} = C$ since all $c_i \in \mathbb{F}_q$, and as $c_{n+1} = -\sum_{i=1}^n c_i$, we see that $c_{n+1} \in \mathbb{F}_q$ and $(c_1, \ldots, c_n) \in \overline{C}$. Thus, $\overline{C} = \overline{C'}|_{\mathbb{F}_q}$, as claimed. By the previous result, $\overline{C'}$ is a rational Goppa code. Therefore, $\overline{C}$ is a subfield subcode of a rational Goppa code. □

In order to get Goppa codes of arbitrary large length over a fixed finite field $\mathbb{F}_q$, we need to work with function fields other than $\mathbb{F}_q(x)/\mathbb{F}_q$. We will look in detail at some other examples of function fields in the next chapter. Later, we will consider the problem of determining how many rational points there is on a curve. We will prove the Hasse-Weil theorem, which gives a bound, in terms of q and of the genus of the curve, for the number of rational points.

6 Examples of Function Fields

In this chapter we work in detail with two examples of curves defined over a finite field. To help us make computations, we describe, without proof, the connection between points on the curve and discrete valuation rings of its function field. This connection is a bit more complicated than the situation for algebraically closed fields, where there is a 1-1 correspondence between the points on a nonsingular curve and the discrete valuation rings of its function field.

6.1 The Connection Between Points and Places

To describe the correspondence, let $k = \mathbb{F}_q$ be a finite field, let C be a nonsingular projective curve defined over k, and let $F = k(C)$ be its function field. Recall that for C to be defined over k means $C = Z(f)$ for some homogeneous polynomial $f \in k[x, y, z]$. Let K be an algebraic closure of k. An exercise, using a small amount of Galois theory, will show that we may take K to be the union of all finite fields containing k. We view our curve in projective space $\mathbf{P}^2(K)$ over K. Let P be a point on the curve. The local ring of P is

$$\mathcal{O}_P(C/k) = \{\varphi \in F : \varphi(P) \text{ is defined}\}.$$

Its unique maximal ideal is

$$M_P(C/k) = \{\varphi \in F : \varphi(P) = 0\}.$$

We will write $\mathcal{O}_P$ for $\mathcal{O}_P(C/k)$ and M_P for $M_P(C/k)$. Let $k(P)$ be the residue field of this valuation ring; that is, $k(P) = \mathcal{O}_P/M_P$. Then $k(P)$ is a finite extension of k. In fact, if ev_P is the evaluation map $\mathcal{O}_P \to K$ given by $\mathrm{ev}_P(f) = f(P)$, then ev_P is a ring homomorphism from $\mathcal{O}_P$ to K with kernel M_P. Thus, ev_P induces an injective homomorphism $k(P) \to K$, and this sends the coset $\varphi + M_P$ of a function φ to the evaluation $f(P)$. If $P = (a : b : c) \in C$, then $\varphi(P) = \varphi(a, b, c)$. Recall that we may represent elements of $F = k(C)$ as quotients of homogeneous polynomials over k of the same degree. Thus, we see that $\varphi(a, b, c) \in k(a, b, c)$, the field generated over k by the coordinates of P. From this we see that $k(P) = k(a, b, c)$. In particular, $k(P) = k$ if and only if $a, b, c \in k$. That is, $k(P) = k$ if and only if P is a k-rational point. For any point P, we set $\deg(P) = [k(P) : k]$. This is the same number as $\deg(M_P)$, since $\deg(M_P)$ was defined to be $[\mathcal{O}_P/M_P : k] = [k(P) : k]$. So, if $P \in C(k)$ is a k-rational point of C, then $\deg(M_P) = 1$, so M_P is a rational point of $\mathbf{P}_F$. Furthermore, it can be shown that given any place M of F/k with $\deg(M) = 1$, then there is a unique k-rational point $P \in C$ with $M = M_P$. Thus, there is a 1-1 correspondence between rational points of C and rational points of $\mathbf{P}_F$.

For points of degree greater than 1, the correspondence is somewhat more complicated. Suppose that $P = (a : b : c) \in C$. Any automorphism σ of K/k yields a new point $\sigma(P) = (\sigma(a), \sigma(b), \sigma(c)) \in \mathbf{P}^2(K)$. If C is the curve $Z(f)$ with $f \in k[x, y, z]$, then the

coefficients of f are fixed by σ. Thus, $f(\sigma(P)) = f(P) = 0$, so $\sigma(P) \in C$. We state without proof that the set of points $\{\tau(P) : \tau \in \mathrm{Gal}(K/k)\}$ satisfies

$$|\{\tau(P) : \tau \in \mathrm{Gal}(K/k)\}| = \deg(P),$$

and that this set is the set of all points whose local ring is equal to $\mathcal{O}_P$. Moreover, we have a 1-1 correspondence between places of $\mathbf{P}_F$ and sets of points on C, where the correspondence sends M_P to $\{\tau(P) : \tau \in \mathrm{Gal}(K/k)\}$. Moreover, if $P = (a : b : c)$ has its coordinates in the finite field $\mathbb{F}_{q^r}$, then $\mathrm{Gal}(\mathbb{F}_{q^r}/\mathbb{F}_q)$ is generated by the automorphism σ, where $\sigma(t) = t^q$. Then $\{\sigma^i(P) : 1 \leq i \leq r\}$ is the set of points who share the same local ring. If $\tau \in \mathrm{Gal}(K/k)$ and $\varphi \in k(C)$, by representing φ as a quotient of homogeneous polynomials with coefficients in k, we see that $\varphi(\tau(P)) = \tau(\varphi(P)) = \varphi(P)$. Therefore, φ is defined at P if and only if φ is defined at $\tau(P)$. This is why P and $\tau(P)$ share the same local ring.

Finally, we point out that $f(P) = f + M_P$ for any $P \in C$, although the proof is not too difficult. Since we have defined evaluation of $f \in F$ at a place $M_P \in \mathbf{P}_F$ by $f(M_P) = f + M_P$, we see that our two definition of evaluation agree; that is, $f(P) = f(M_P)$.

To summarize this connection, we state it as a theorem. Recall that an extension $\mathbb{F}_{q^r}/\mathbb{F}_q$ of finite fields is a cyclic Galois extension with Galois group generated by the Frobenius automorphism σ, where σ is given by the formula $\sigma(a) = a^q$ for all $a \in \mathbb{F}_{q^r}$.

Theorem 6.1. *Let F be the function field of a nonsingular projective curve C defined over the base field $k = \mathbb{F}_q$.*

1. *There is a 1-1 correspondence between rational points on the curve and places of degree 1 of F/k, where a rational point P corresponds to the place M_P.*

2. *There is a 1-1 correspondence between places of F/k of degree n and sets of points on the curve with coefficients in $\mathbb{F}_{q^n}$, where a place corresponds to the set of points all having the same local ring. Alternatively, if K is the algebraic closure of k, and if $P \in C$, the place M_P corresponds to the set $\{\tau(P) : \tau \in \mathrm{Gal}(K/k)\}$. If $k(P) = \mathbb{F}_{q^n}$, then $P \in C(\mathbb{F}_{q^n})$, and this set of points is equal to $\{\sigma(P) : \sigma \in \mathrm{Gal}(\mathbb{F}_{q^n}/\mathbb{F}_q)\}$.*

3. *If $P \in C$, and if M_P is the corresponding place, then $k(P) = k(M_P)$. Moreover, if $f \in F$, then $f(P)$ is defined if and only if $f(M_P)$ is defined, and $f(P) = f(M_P)$.*

6.2 Connections Between Divisors

To help to compute divisors of functions, we define the intersection divisor of a curve. Recall that we have defined divisors of F/k; we will refer to the group of such divisors as $\mathrm{Div}(F/k)$. These divisors are integral linear combinations of points in $\mathbf{P}_F$. The divisors we are about to define will be combinations of points on a curve C. We will refer to the group of these divisors as $\mathrm{Div}(C)$. Let $C = Z(g)$ be a curve. If f is a homogeneous polynomial, then the *intersection divisor* $Z(g) \cap Z(f) \in \mathrm{Div}(C)$ is the divisor $\sum_P n_P P$, where P is summed over

all points in both $Z(g)$ and $Z(f)$, and where n_P is the *multiplicity* of P in this intersection. If $D = \sum_i n_i P_i \in \mathrm{Div}(C)$, then the *degree* of D is defined to be $\sum_i n_i$. The degree of the intersection divisor $Z(f) \cap Z(g)$ is $\deg(f)\deg(g)$, by the following theorem of Bézout.

Theorem 6.2 (Bézout). *Let $Z(f)$ and $Z(g)$ be projective curves in $\mathbf{P}^2(K)$, where K is an algebraically closed field. Then the number of points, counted with multiplicity, on both $Z(f)$ and $Z(g)$ is equal to* $\deg(f) \cdot \deg(g)$.

We will not give the definition of multiplicity, but we indicate its meaning in examples below. If $f, g \in k[x, y, z]$, and if $P \in Z(f) \cap Z(g)$, then any point of the form $\tau(P)$ for $\tau \in \mathrm{Gal}(K/k)$ is also in $Z(f) \cap Z(g)$. The difference (and connections) between a divisor of F/k and an intersection divisor of C will be addressed in examples below.

As means of using intersection divisors in $\mathrm{Div}(C)$ to compute divisors in $\mathrm{Div}(F/k)$, we describe the connection between $\mathrm{Div}(F/k)$ and $\mathrm{Div}(C)$. If $P \in \mathbf{P}_F$, set $\Psi(P) = \sum_i P_i$, where $P_i \in C$ are the points whose local ring is $\mathcal{O}_P$. That is, $\Psi(P)$ is the sum of the points on C whose local ring is the discrete valuation ring corresponding to P. The map $\Psi : \mathrm{Div}(F/k) \to \mathrm{Div}(C)$ is a group homomorphism, and it is degree preserving; that is, $\deg(\Psi(D)) = \deg(D)$. This follows since if M is a place, then there are $\deg(M)$ points on C whose corresponding place is M. The map Ψ is injective since distinct places correspond to distinct sets of points. The connection between the divisor of a function and an intersection divisor is the following. If $f \in k(C)$, then we may write $f = g/h$ for some homogeneous polynomials g, h. Then $(f) = \Psi^{-1}(D - E)$, where D is the intersection divisor $C \cap Z(g)$ and E is the intersection divisor $C \cap Z(h)$. In examples below, we will compute principal divisors using this fact.

6.3 Elliptic Curves

An elliptic curve is a curve given by an affine equation of the form $y^2 = f(x)$, where $f(x)$ is a cubic with no repeated roots. We look at a specific example. Let $k = \mathbb{F}_3$, and let C be the projective curve over k given by the equation $y^2 z = x^3 - xz^2 - z^3$. This is an elliptic curve since the dehomogenized version of this equation is $y^2 = x^3 - x - 1$, and $x^3 - x - 1$ can be seen to have no repeated roots over $\mathbb{F}_3$. Let $s = x/z$ and $t = y/z$, elements of the function field $F = k(C)$. Note that $t^2 = s^3 - s - 1$ and that $F = k(s, t)$. Moreover, we claim that k is the exact constant field of F/k. To see this, we could give a direct argument, although instead we quote a theorem that says k is the exact constant field of F/k provided that the defining polynomial (or the dehomogenized polynomial) is *absolutely irreducible*. This latter condition means that the polynomial is irreducible over $L(s, t)$ for any field extension L of k. To see this is true in our example, let $f = y^2 - x^3 + x + 1$. If f factors in $L[x, y]$ as $f = gh$, then either the degrees in y of both g and h are 1, or one has degree 0 and the other degree 2. This latter case cannot happen since one of g and h is then a polynomial in x which would divide every coefficient of f, which is clearly false. If g and h are both linear in y, then we may write $g = ay + b$ and $h = cy + d$ with $a, b, c, d \in k[x]$. Then $f = (ac)y^2 + (ad + bc)y + bd$.

Then $ac = 1$, so a, c are units in $k[x]$. They are then constants, and by multiplying and dividing accordingly, we may assume that $a = c = 1$. Then $b + d = 0$ and $bd = -x^3 + x + 1$. This yields $b^2 = x^3 - x - 1$, which is false since $x^3 - x - 1$ has no repeated roots (and is square free). This contradiction shows that f is irreducible in $L[x, y]$ for any field $L \supseteq k$. Since $C = Z(f)$ is nonsingular, the genus of F/k is given by the formula

$$g = \frac{1}{2}(\deg(f) - 1)(\deg(f) - 2) = 1.$$

It is easy to check that C is a nonsingular curve; essentially this follows from the fact that $x^3 - x - 1$ has no repeated roots over $\mathbb{F}_3$. Therefore, each local ring $\mathcal{O}_P$ for $P \in C$ is a discrete valuation ring of F.

This curve has a unique rational point; to see this, note that if $(a : b : c) \in C(\mathbb{F}_3)$, and if $c \neq 0$, then we may assume that $c = 1$. Then $b^2 = a^3 - a - 1$. Checking the three possibilities of $a \in \mathbb{F}_3$, we see that there are no solutions to this equation. However, if $c = 0$, then $0 = a^3$, so $a = 0$. Finally, since $b \neq 0$ else $(a : b : c)$ is not a valid point in $\mathbf{P}^2$, we see that, by dividing by b, we may assume that $b = 1$. Therefore, $P_\infty = (0 : 1 : 0)$ is the unique rational point of this curve, and it is the unique *point at infinity*; recall that we can view the affine curve $y^2 = x^3 - x - 1$ as sitting inside C by sending a point (a, b) to $(a : b : 1)$. Thus, points at infinity are points whose third component is 0. Let C_0 be the affine curve $y^2 = x^3 - x - 1$, viewed inside C. So, $C = C_0 \cup \{P_\infty\}$.

We now consider points of larger degree. To do this we need to consider field extensions of $\mathbb{F}_3$. The polynomial $T^2 + 1$ is irreducible over $\mathbb{F}_3$. Therefore, $\mathbb{F}_3[T]/(T^2 + 1)$ is a field, and since it has dimension 2 over $\mathbb{F}_3$, it is isomorphic to $\mathbb{F}_9$. If α is a root of $T^2 + 1$, then $\mathbb{F}_9 = \mathbb{F}_3(\alpha)$. A computation will show that

$$C(\mathbb{F}_9) = \{(0 : \alpha : 1), (0 : -\alpha : 1), (1 : \alpha : 1), (1 : -\alpha : 1), (-1 : \alpha : 1), (-1 : -\alpha : 1), P_\infty\}.$$

In fact, from the equation $y^2 = x(x-1)(x+1) - 1$, we see that if $x = 0, 1, -1$, then $y^2 = -1$, so $y = \pm\alpha$. All of these points have degree 2 except for P_∞, since their coefficients generate $\mathbb{F}_9$. By a Maple calculation, one can show that $C(\mathbb{F}_{27})$ has 28 points. All points in $C(\mathbb{F}_{27})$ but P_∞ is a point of degree 3 since each has at least one coordinate outside $\mathbb{F}_3$, so the coordinates generate $\mathbb{F}_{27}$. These calculations can be done with the Maple worksheet POINTS.MWS, which defines procedures for computing points on an affine or a projective plane curve.

We now look at examples of computing the intersection divisors. To give some notation to connect divisors of F/k and divisors of C, if $(a : b : 1) \in C$, we write the corresponding place in $\mathbf{P}_F$ as P_{ab}, or as $P_{a,b}$. We use the notation P_∞ both for the point $(0 : 1 : 0)$ and the corresponding place of F/k.

Example 6.3. Consider the intersection divisor $C \cap Z(z)$. If $(a : b : c) \in C$, then $(a : b : c) \in Z(z)$ if and only if $c = 0$. Thus, the only point on both curves is P_∞. Since C is the zero set of a cubic and z is linear, their degrees multiply to 3. Therefore, this intersection divisor is $3P_\infty$. Similarly, if we consider $C \cap Z(x)$, then $P = (a : b : c)$ is on both curves if $a = 0$

and $b^2c = a^3 - ac^2 - c^3 = -c^3$. Therefore, if $c = 0$, then $b = 1$, and we get P_∞. If $c \neq 0$, we may assume that $c = 1$, and then $b^2 = -1$, so $b = \pm\alpha$. We then have two points $(0 : \alpha : 1)$ and $(0 : -\alpha : 1)$. There are three points in total in this intersection, so the multiplicity of each is 1. Thus, the intersection divisor is $P_\infty + (0 : \alpha : 1) + (0 : -\alpha : 1)$. This is a divisor in $\mathrm{Div}(C)$. Under the map Ψ defined above, this divisor is $\Psi(P_\infty + P_{0\alpha})$.

Example 6.4. Let us consider the divisor of $s = x/z$ in $\mathrm{Div}(F/k)$. As we calculated earlier, the intersection divisors $C \cap Z(x)$ and $C \cap Z(z)$ are $P_\infty + (0 : \alpha : 1) + (0 : -\alpha : 1) = \Psi(P_{0\alpha} + P_\infty)$ and $\Psi(3P_\infty)$, respectively. Thus, by the relation between principal divisors and intersection divisors, $(s) = P_{0\alpha} - 2P_\infty$.

Example 6.5. We calculate the divisor of $t = y/z$. We have already calculated the intersection divisor $C \cap Z(z)$, getting $3P_\infty$. We next must see what is the intersection divisor $C \cap Z(y)$. A point $(a : b : c)$ is on $Z(y)$ if and only if $b = 0$. Since $b^2c = a^3 - ac^2 - c^3$, we have $a^3 = ac^2 + c^3$. If $c = 0$, then $a = 0$; thus, $c \neq 0$. By dividing by c, we may assume that $c = 1$. Then $a^3 - a - 1 = 0$. There are three roots in K to this equation. If we label them a_1, a_2, and a_3, then the points on this intersection divisor are $(a_1 : 0 : 1)$, $(a_2 : 0 : 1)$, and $(a_3 : 0 : 1)$. Therefore, the multiplicity of each point is 1, and so $C \cap Z(y) = (a_1 : 0 : 1) + (a_2 : 0 : 1) + (a_3 : 0 : 1)$. Each a_i generates the field $\mathbb{F}_3(a_i) = \mathbb{F}_{27}$ since $x^3 - x - 1$ is easily seen to be irreducible over $\mathbb{F}_3$ (because it has no roots in $\mathbb{F}_3$). These three points are of the form $\{\tau(a_1 : 0 : 1) : \tau \in \mathrm{Gal}(\mathbb{F}_{27}/\mathbb{F}_3)\}$. Therefore, the three correspond to the one place $P_{a_1 0}$. This is then a place of degree 3, and so $C \cap Z(y) = \Psi(P_{a_1 0})$. Since (g) is the difference of the preimages of these intersection divisors, we have $(t) = P_{a_1 0} - 3P_\infty$.

Example 6.6. We calculate the space $L(nP_\infty)$. If $n > 0 = 2g - 2$, then the Riemann-Roch theorem yields $\dim(nP_\infty) = \deg(nP_\infty) + 1 - g = n$. We claim that

$$\left\{s^i t^j : 0 \leq i, 0 \leq j \leq 1, 2i + 3j \leq n\right\}$$

is a basis for $L(nP_\infty)$. These elements clearly are linearly independent over k since $\{1, t\}$ is a basis for F over $k(s)$ and since s is transcendental over k. Moreover, it is not hard to see that there are n elements in the set, so we only need to show that all are in $L(nP_\infty)$. We see that this is true since

$$\begin{aligned}(s^i y t^j) &= i(s) + j(t) = i\,(P_{0\alpha} - 2P_\infty) + j(P_{a_1 0} - 3P_\infty) \\ &= (iP_{0\alpha} + jP_{a_1 0}) - (2i + 3j)P_\infty.\end{aligned}$$

Therefore, if $2i + 3j \leq n$, then $(s^i t^j) + nP_\infty \geq 0$.

We finish this section by giving a connection between divisors of F/k of degree 0 and rational points on C. While we will not need this for our coding theory purposes, it will help to illustrate one of the first notions in the theory of elliptic curves. First, recall that P_∞ is a

rational point. If Q is another rational point, then $Q - P_\infty$ is a divisor of degree 0. We claim that every divisor of degree 0 is equivalent to a divisor of the form $Q - P_\infty$ for some rational point Q. To prove this, let D be a divisor of degree 0. Then $D + P_\infty$ has degree 1. Since the genus of C is 1, the Riemann-Roch theorem yields $\dim(D + P_\infty) = \deg(D + P_\infty) = 1$. Thus, there is a nonzero $f \in L(D + P_\infty)$. Then $D + P_\infty + (f) \geq 0$ is a positive divisor of degree 1. The only way for this to happen is if $D + P_\infty + (f) = Q$ for some rational point Q, which shows that D is equivalent to $Q - P_\infty$. We have thus proven that every divisor D of degree 1 is equivalent to a divisor of the form $Q - P_\infty$. If D is principal, then D is equivalent to $0 = P_\infty - P_\infty$. On the other hand, if $Q \neq P_\infty$, we claim that $Q - P_\infty$ is not principal. For, if $Q - P_\infty = (f)$ for some f, then $(f)_0 = Q$ and $(f)_\infty = P_\infty$. However, $\deg((f)_\infty) = [F : k(f)]$, which would force $[F : k(f)] = 1$, or $F = k(f)$. Since a rational function field has genus 0, this cannot happen. Thus, $Q - P_\infty$ is not principal. Similarly, if $Q \neq Q'$ are both rational points, then $Q - P_\infty$ is not equivalent to $Q' - P_\infty$. As a consequence, the subgroup of divisor classes of degree 0 is $\{Q - P_\infty : Q \in C(k)\}$.

Example 6.7. To give some motivation for looking at the group of divisor classes of degree 0, a classical result about elliptic curves is that one can define a group structure on the set of rational points such that P_∞ is the identity of the group. Given rational points P and Q, to define $P + Q$, consider the line connecting P and Q. This line will intersect the elliptic curve in three points. Two of these points are P and Q; let R be the third point. Next, the line connecting P_∞ and R hits the curve at a third point. This third point is the point $P + Q$. It is tedious to show that we do get a group structure from this.

(In the picture $O = P_\infty$, $T = P * Q$, and $R = P + Q$.) However, The function that sends Q to the divisor class of $Q - P_\infty$ is a group isomorphism from this group of rational points to the group of divisor classes of degree 0. Thus, a non-geometric way to define $P + Q$ is that $P + Q$ is the unique point R that satisfies $R - P_\infty \sim (Q - P_\infty) + (P - P_\infty)$. To see why this is true, say that the line L passing through P and Q hits C at T, and the line L' passing through P_∞ and T hits C at R. There are linear polynomials f and g with $L = Z(f)$ and $L' = Z(g)$. The divisor (f/g) is the difference between the intersection divisor $C \cap Z(f)$ and the intersection divisor $C \cap Z(g)$. These divisors are $P + Q + T$ and $P_\infty + T + R$, respectively. Thus, in the divisor class group, we have $0 \sim (f/g) = P + Q - P_\infty - R$. Thus, $R \sim P + Q - P_\infty$, and so $R - P_\infty \sim (Q - P_\infty) + (P - P_\infty)$.

6.4 Hermitian Curves

In this section we work with a curve that has many rational points. This curve will allow us to construct Goppa codes of large length. Let q be a power of a prime, and let C be the *Hermitian curve* given by the equation $y^q z + yz^q = x^{q+1}$ over the field $k = \mathbb{F}_{q^2}$. We first note that C is a nonsingular curve since the three partial derivatives of $y^q z + yz^q - x^{q+1}$ simultaneously vanish only at $x = y = z = 0$. Let F be the function field $F = k(C)$ of C and let $s = x/z$ and $t = y/z$. Then $F = k(s, t)$, and $t^q + t = s^{q+1}$. The base field k is the exact field of constants because $x^{q+1} - y^q - y$ is absolutely irreducible. To see that this is true, let L be any field extension of k. The Eisenstein criterion, applied to the principal ideal domain $L[y]$, shows that $x^{q+1} - y^q - y$ is irreducible over $L(y)$ since $y^q + y$ is a square-free polynomial since its derivative is relatively prime to $y^q + y$.

As a consequence of this, we see that $[F : k(s)] = q$ and $[F : k(t)] = q + 1$. For the first equation, we have $F = k(s)(t)$, so $[F : k(s)]$ is equal to the degree of the minimal polynomial of t over $k(s)$. This polynomial clearly is $T^q + T - s^{q+1} \in k(s)[T]$, which has degree q. A similar argument holds for the second equation.

We first consider what are the k-rational points of C. We claim that there are $1+q^3$ such points. First, let $(a : b : c) \in C$. If $c = 0$, then $a = 0$, and so $b \neq 0$. By dividing by b, we see that the only such point is $(0 : 1 : 0)$. This is the only point at infinity; that is, it is the only point not on the affine curve $y^q + y = x^{q+1}$ consisting of all points whose third component is 1. This point is clearly a k-rational point. Now, consider rational points $(a : b : 1)$. If $a \in k$ is any element, then we note that $a^{q+1} \in \mathbb{F}_q$. For, since $a \in \mathbb{F}_{q^2}$, we have $a^{q^2} = a$. Then

$$\left(a^{q+1}\right)^q = a^{q^2+q} = a^{q^2} a^q = a \cdot a^q = a^{q+1},$$

so a^{q+1} is a root of $x^q - x$. Since the elements of $\mathbb{F}_q$ are precisely the roots of this polynomial, we see that $a^{q+1} \in \mathbb{F}_q$, as claimed. Next, if $\alpha \in \mathbb{F}_q$, we claim that there are q elements of $\mathbb{F}_{q^2}$ satisfying $y^q + y = \alpha$. To see this, recall that $\mathbb{F}_{q^2}$ is a Galois extension of $\mathbb{F}_q$, and the Galois group is the cyclic group generated by σ, where $\sigma(t) = t^q$ for all $t \in \mathbb{F}_{q^2}$. Moreover σ has order 2 since $[\mathbb{F}_{q^2} : \mathbb{F}_q] = 2$. An element y satisfying $y^q + y = \alpha$ is then an element y satisfying $\sigma(y) + y = \alpha$. Consider the polynomial $t^2 - \alpha t + \beta \in \mathbb{F}_q[t]$, where $\beta \in \mathbb{F}_q$ is chosen so that this polynomial is irreducible over $\mathbb{F}_q$. This is possible since $\mathbb{F}_{q^2}$ can be generated over $\mathbb{F}_q$ by an element whose minimal polynomial has degree 2 over $\mathbb{F}_q$. This minimal polynomial then has two roots in $\mathbb{F}_{q^2}$, and these roots are not in $\mathbb{F}_q$. Say the roots are y and y'. Since $\sigma(y)$ is also a root, we have $y' = \sigma(y)$. Then $t^2 - \alpha t + \beta = (t - y)(t - \sigma(y))$ shows that $y + \sigma(y) = \alpha$, or $y^q + y = \alpha$. Thus, $\mathbb{F}_{q^2}$ contains a root of $y^q + y = \alpha$. This polynomial then splits over $\mathbb{F}_{q^2}$ since $\mathbb{F}_{q^2}$ is Galois over $\mathbb{F}_q$. Furthermore, since the derivative of $y^q + y - \alpha$ is 1, which is not zero, there are no repeated roots. So, there are exactly q elements of $\mathbb{F}_{q^2}$ satisfying $y^q + y - \alpha$. Since this is true for every $\alpha \in \mathbb{F}_q$, and since $x^{q+1} \in \mathbb{F}_q$ for all $x \in \mathbb{F}_{q^2}$, we see that the number of pairs (a, b) satisfying $b^q + b = a^{q+1}$ is q^3 as there are q^2 choices for a, and for each a there are q solutions to $b^q + b = a^{q+1}$. Finally, adding our point $(0 : 1 : 0)$ at

infinity, we have $1 + q^3$ points over $\mathbb{F}_{q^2}$.

The Hermitian curve $y^q z + yz^q = x^{q+1}$ is an example of a curve that meets the *Hasse-Weil bound*. We will see that this bound is a consequence of the Hasse-Weil theorem, which proves the Riemann Hypothesis for algebraic function fields over finite fields. This bound says that the number of k-rational points N of a curve of genus g satisfies

$$|N - (1 + |k|)| \leq 2g\sqrt{|k|}.$$

If $k = \mathbb{F}_{q^2}$, then $|k| = q^2$, and so this bound is $|N - (1 + q^2)| \leq 2qg$. Because our curve is nonsingular, the genus is given by the formula $g = \frac{1}{2}(\deg f - 1)(\deg f - 2)$, where f is the polynomial defining the curve. This degree is $q + 1$, so $g = \frac{1}{2}q(q - 1)$. The Serre bound then simplifies to

$$\left|N - (1 + q^2)\right| \leq q^2(q - 1),$$

or

$$q^3 - 2q^2 - 1 \leq N \leq 1 + q^3.$$

Therefore, N is as large as possible for a curve of genus g that defined over $\mathbb{F}_{q^2}$.

We now look into calculating some divisors on C. We mimic our notation in the previous section by writing P_{ab} for the place corresponding to the point $(a : b : 1)$, and P_∞ for the place corresponding to $(0 : 1 : 0)$.

Example 6.8. First, let us consider the divisor of x/z. We calculate this as we calculated principal divisors for the elliptic curve example. So, consider first the intersection divisor $C \cap Z(x)$. Since points on the curve satisfy the equation $y^q z + yz^q = x^{q+1}$, if a point $(a : b : c)$ is in $Z(x)$, then $a = 0$ and $b^q c + bc^q = 0$. If $c = 0$, then $b \neq 0$ to have a legitimate point of projective space, so the point is $(0 : 1 : 0)$. If $c \neq 0$, we may divide by c to assume that $c = 1$. Then $b^q + b = 0$. There are exactly q solutions in k to this equation (see the argument above for rational points of C); say they are $b_1, \ldots, b_q$. Then the points in $C \cap Z(x)$ are $(0 : 1 : 0), (0 : b_1 : 1), \ldots, (0 : b_q : 1)$. Thus, the multiplicities of each must be 1, and so $C \cap Z(x) = (0 : 1 : 0) + (0 : b_1 : 1) + \cdots + (0 : b_q : 1)$. As for $C \cap Z(z)$, we see that if $(a : b : c) \in C$ has $c = 0$, then $a = 0$, and so the only point we get is $(0 : 1 : 0)$. This point must then have multiplicity $q + 1$, so $C \cap Z(z) = (q + 1)(0 : 1 : 0)$. Therefore,

$$\Psi((x/z)) = (0 : b_1 : 1) + \cdots + (0 : b_q : 1) - q(0 : 1 : 0).$$

The divisor (x/z) is then equal to $P_{0b_1} + \cdots + P_{0b_q} - qP_\infty$. In particular, this says that P_∞ is a pole of x/z. It is not at first apparent that this is so, because P_∞ is a zero of both x and of z. However, we can manipulate the defining equation of the curve to see why this is so. From $y^q z + yz^q = x^{q+1}$, dividing by $x^q z$ gives

$$\frac{x}{z} = \frac{y^q + yz^{q-1}}{x^q}.$$

The numerator of the right hand quotient is defined at P_∞ and is not 0 at P_∞, while the denominator vanishes at P_∞. Thus, P_∞ is indeed a pole of x/z. Moreover, we see that $v_{P_\infty}(x/z) = -qv_\infty(x)$, which shows that the order of the pole P_∞ is at least q, since $v_\infty(x) \geq 1$. By realizing that P_∞ is the only pole of x/z, we see that $(x/z)_\infty = qP_\infty$ since $[F : k(x/z)] = q$, which we pointed out earlier.

Example 6.9. For another example of calculating divisors, consider $f = y/z$. Since we have already calculated $C \cap Z(z) = (q+1)P_\infty$, we only need to calculate $C \cap Z(y)$. If $(a : b : c) \in C$ lies on $Z(y)$, then $b = 0$. Therefore, we get $a = 0$, since $b^q c + bc^q = a^{q+1}$. Then $c \neq 0$, and so $(a : b : c) = (0 : 0 : 1)$. This point then must have multiplicity $(q + 1)$. Our divisor (y/z) is then equal to

$$(y/z) = (q+1)P_{00} - (q+1)P_\infty = (q+1)\left(P_{00} - P_\infty\right).$$

Next, we calculate the space $L(nP_\infty)$, where n is an arbitrary positive integer.

Proposition 6.10. *The space $L(nP_\infty)$ has basis $\{s^i t^j : 0 \leq i, 0 \leq j \leq q-1, iq + j(q+1) \leq n\}$.*

Proof. We have shown that $(s) = P_{0b_1} + \cdots + P_{0b_q} - qP_\infty$ and $(t) = (q+1)(P_{00} - P_\infty)$. Therefore,

$$(s^i t^j) = i(s) + j(t) = i(P_{0b_1} + \cdots + P_{0b_q} - qP_\infty) + j(q+1)(P_{00} - P_\infty).$$

Thus, $s^i t^j \in L(nP_\infty)$ if and only if $iq + j(q+1) \leq n$. Moreover, since $[F : k(s)] = q$, the elements $1, t, \ldots, t^{q-1}$ span F as a $k(s)$-vector space. We then see that the set of elements given above is k-linearly dependent (since s is transcendental over k). It remains to show that they span $L(nP_\infty)$. To see this, let $f \in L(nP_\infty)$, and write

$$f = \frac{f_0 + f_1 t + \cdots + f_{q-1} t^{q-1}}{g},$$

where $f_i, g \in k[s]$ and g has no factor in common with every f_i. Since $(f) + nP_\infty \geq 0$, the only possible pole of f is P_∞. Thus, the only zero of g can be P_∞. However, since g is a polynomial, P_∞ is not a zero of g, and that g has a zero unless it is constant. Suppose that g is not a constant, and let a be a zero of g. Since there is no factor of g common to each f_i, we see that $f_j(a) \neq 0$ for some j. The polynomial $\sum_i f_i(a)T^i$ is nontrivial, so there is some nonzero value b (in an algebraic closure of $\mathbb{F}_{q^2}$) for which the polynomial evaluated at b is nonzero. Then $\sum_i f_i(a)b^i \neq 0$. Consider a point $(a : b : c)$, where $c \neq 0$ is any element satisfying $(a : b : c) \in C$. Finding a c is possible, since c can be any root of the polynomial $b^q T + bT^q - a^{q+1}$, and this polynomial has a root in an algebraic closure of $\mathbb{F}_{q^2}$ since $b \neq 0$. If P is a place corresponding to the point $(a : b : c)$, then P is a pole of f, which is a contradiction since the only pole of f is P_∞ and $P \neq P_\infty$ as $c \neq 0$. This forces g to be a constant, so f is a linear combination of $\{s^i t^j : 0 \leq j \leq q-1, 0 \leq i\}$, as desired. □

With this example we can nicely illustrate the dependence of $\dim(D)$ on $\deg(D)$. The Maple worksheet HERMITIAN.MWS will calculate the size of the basis of $L(nP_\infty)$; using it, we obtain the following table in the case $q = 8$. Recall that since the genus is $q(q-1)/2 = 28$, if $\deg(D) > 2g - 2 = 54$, then $\dim(D) = \deg(D) + 1 - g$, while $\dim(D) \geq \deg(D) + 1 - g$ in any case.

n	$\deg(nP_\infty) + 1 - g$	$\dim(nP_\infty)$
45	18	21
46	19	21
47	20	21
48	21	22
49	22	23
50	23	24
51	24	25
52	25	26
53	26	27
54	27	28
55	28	28
56	29	29
57	30	30

7 The Hasse-Weil Theorem

In this final chapter we will introduce the Riemann zeta function of a curve defined over a finite field and see how knowledge of its zeros yield bounds on the number of rational points of the curve. We start off by describing the classical Riemann zeta function, including writing it in a way that will allow us to define a zeta function of an arbitrary algebraic number field, and then to define a zeta function for an algebraic function field F in one variable over a finite field.

7.1 The Riemann Zeta Function

The classical Riemann zeta function $\zeta(s)$ is defined by the formula

$$\zeta(s) = \sum_{n=1}^{\infty} n^{-s}.$$

From the integral test, we see that this series converges if $s > 1$. By making use of complex function theory, we can view $\zeta(s)$ as a function of a complex variable s, and by doing so, we get a complex analytic function that converges for $\mathrm{Re}(s) > 1$. This function encodes a lot of information about primes; one indication of this is the product formula for $\zeta(s)$; which says that

$$\zeta(s) = \prod_{p} (1 - p^s)^{-1},$$

where the product is over all prime numbers. The proof of this fact can be found in standard complex analysis texts. To give an intuitive idea of why it is true, first consider $\zeta(s)(1-2^s)$. We have

$$\zeta(s)(1-2^s) = \sum_{n=1}^{\infty} n^{-s} - \sum_{n=1}^{\infty} (2n)^{-s} = \sum_{m} m^{-s},$$

where m runs over all odd numbers. Next,

$$\zeta(s)(1-2^s)(1-3^s) = \sum_{m} m^{-s} - \sum_{m} (3m)^{-s} = \sum_{r} r^{-s},$$

where r runs over all integers not divisible by 2 or 3. Continuing this reasoning, and being careful with the limits, will yield the result.

The Riemann hypothesis, probably the most important unsolved problem in mathematics has to do with the zeros of $\zeta(s)$. The *functional equation* for ζ says that $\zeta(s) = 2^s \pi^{s-1} \sin(\pi s/2) \Gamma(1-s) \zeta(1-s)$, where $\Gamma(s)$ is the Gamma function, a generalization of the factorial function in that $\Gamma(n) = n!$ if n is a nonnegative integer. From this equation we see that there are "trivial" zeros of $\zeta(s)$, occurring when $\sin(\pi s/2) = 0$. Other zeros of $\zeta(s)$ are then called nontrivial zeros. Using the product formula for $\zeta(s)$, it follows that $\zeta(s)$ has no zeros in the half plane $\mathrm{Re}(s) > 1$. From the functional equation, one sees that the only non-

trivial zeros occur in the strip $0 \leq \mathrm{Re}(s) \leq 1$. The Riemann hypothesis conjectures that all of the nontrivial zeros of $\zeta(s)$ are on the line $\mathrm{Re}(s) = 1/2$. While this is still unproven, Weil proved in 1941 that the analogue of the Riemann hypothesis for zeta functions associated to function fields of curves over finite fields is true. It is this fact that we will see in this chapter, and how this fact leads to a bound on the number of rational points of a curve.

7.2 Riemann Zeta Functions of Number Fields

We will define an analogue of the Riemann zeta function which will be associated to a curve defined over a finite field. However, to motivate its definition, we first discuss the Riemann zeta function of an algebraic number field. To motivate the definition of these functions, we view $\zeta(s)$ in another light. To do this we recall some of the theory of Dedekind domains.

Let A be a Dedekind domain with quotient field K. A *fractional ideal* of A is a nonzero A-submodule I of K such that $aI \subseteq A$ for some nonzero $a \in A$. For example, if $r = a/b \in K$, then $(r) = rA$ is a fractional ideal of A since rA is an A-submodule of K, and $b(rA) = aA \subseteq A$. We call (r) a principal fractional ideal. Furthermore, as this example indicates, if I is a fractional ideal and $aI \subseteq A$ for some a, then aI is an ordinary ideal of A. We extend the definition of product of two ideals to the case of fractional ideals, by setting, for fractional ideals I and J,

$$IJ = \left\{ \sum_{i=1}^{n} a_i b_i : a_i \in I, b_i \in J, n \geq 1 \right\},$$

the usual formula for the product of two ideals. It is easy to show that IJ is a fractional ideal. Let G be the set of all fractional ideals of A. Then this multiplication is an operation on A, and the fractional ideal A is the identity for this operation. Multiplication of fractional ideals is an associative operation, so we have a semigroup with identity. However, one very important fact about Dedekind domains is that any fractional ideal has an inverse. If I is a nonzero fractional ideal, then

$$(A : I) = \{x \in K : xI \subseteq A\}$$

is another fractional ideal, and it is the inverse of I in G. Therefore, G is an Abelian group. The set of principal fractional ideals is a subgroup of G, and the resulting quotient group is called the *ideal class group* of A. The ideal class group of A is often denoted by $\mathrm{Cl}(A)$. If $A = \mathbb{Z}$, then $\mathrm{Cl}(A) = 0$. More generally, a Math 582 exercise shows that $\mathrm{Cl}(A) = 0$ if and only if A is a unique factorization domain. Therefore, $\mathrm{Cl}(A)$ measures the factorization properties of A, in some sense. For some terminology, we call a fractional ideal positive if it is an ordinary ideal, and we write $I \geq 0$ when this occurs.

We now rephrase the definition of the Riemann zeta function. Since $\mathbb{Z}$ is a principal ideal domain, every nonzero ideal of $\mathbb{Z}$ can be written uniquely in the form (n) for some $n \geq 1$. If $I = (n)$, then we recover n as $n = |\mathbb{Z}/I|$. We call this number the norm of I, and write

$N(I)$ for this. In other words,

$$N(I) = |\mathbb{Z}/I|.$$

Therefore,

$$\zeta(s) = \sum_{I \geq 0} N(I)^{-s}.$$

In this formulation, we can extend the definition easily to other number fields. Recall that an algebraic number field K is a finite extension of $\mathbb{Q}$. If A is the integral closure of $\mathbb{Z}$ in K; that is, A is the set of all elements of K that satisfy a monic polynomial equation with coefficients in $\mathbb{Z}$, then A is a Dedekind domain. If I is an ideal of A, we set $N(I) = |A/I|$. We define the Riemann zeta function $\zeta_K(s)$ by

$$\zeta_K(s) = \sum_{I \geq 0} N(I)^{-s},$$

where the sum is over all nonzero ideals of A.

To view this in a slightly different way, recall that since A is a Dedekind domain, any nonzero ideal I can be written uniquely in the form $I = P_1^{e_1} \cdots P_r^{e_r}$, where the P_i are prime ideals and $e_i \geq 1$. By the Chinese remainder theorem, it follows that

$$A/I = A/P_1^{e_1} \oplus \cdots \oplus A/P_r^{e_r}.$$

A calculation, which can be found in the handout on Dedekind domains, shows that $|A/P^e| = |A/P|^e$ for any prime ideal P and positive integer e. Therefore,

$$|A/I| = \prod_{i=1}^{r} |A/P_i|^{e_i}.$$

Thus, $N(I) = \prod_i N(P_i)^{e_i}$.

7.3 Riemann Zeta Functions of Curves

To define zeta functions for curves, we point out that the group of fractional ideals of A is the free Abelian group on the set of nonzero prime ideals of A. This is just a fancy way to say that every ideal can be written uniquely as a product of prime ideals. Every fractional ideal is, in multiplicative notation, an integer linear combination of prime ideals. This indicates that the group of fractional ideals of a number ring is the analogue of the group of divisors of a function field $F/\mathbb{F}_q$. Let $\mathbf{P}_F$ be the set of places of $F/\mathbb{F}_q$; this set is the analogue of the set of nonzero prime ideals of a Dedekind domain. For $P \in \mathbf{P}_F$, we define the norm $N(P)$ of P by $N(P) = |\mathbb{F}_q(P)|$, which is equal to $q^{\deg(P)}$ since $\deg(P) = [\mathbb{F}_q(P) : \mathbb{F}_q]$. If $E = \sum_{i=1}^{r} e_i P_i$

is a positive divisor of $F/\mathbb{F}_q$, we define the norm of E by

$$\begin{aligned} N(E) &= \prod_{i=1}^{r} N(P_i)^{e_i} = \prod_{i=1}^{r} q^{\deg(P_i)e_i} = q^{\sum_{i=1}^{e} e_i \deg(P_i)} \\ &= q^{\deg(E)}. \end{aligned}$$

Finally, we define the Riemann zeta function $\zeta_F(s)$ by

$$\zeta_F(s) = \sum_{E \geq 0} N(E)^{-s},$$

where the sum is over all positive divisors of $F/\mathbb{F}_q$. We can rewrite this a little, which will lead us to a related function. We have

$$\zeta_F(s) = \sum_{E \geq 0} N(E)^{-s} = \sum_{E \geq 0} q^{-s \deg(E)} = \sum_{E \geq 0} (q^{-s})^{\deg(E)}.$$

Let us write $t = q^{-s}$ for the moment. Then

$$\zeta_F(s) = \sum_{E \geq 0} t^{\deg(E)} = \sum_{n=1}^{\infty} \left(\sum_{\deg(E)=n} t^n \right) = \sum_{n=1}^{\infty} A_n t^n,$$

where

$$A_n = \left| E \in \mathrm{Div}(F/\mathbb{F}_q) : E \geq 0, \deg(E) = n \right|.$$

Note that A_1 is the number of places of degree 1; this follows since a positive divisor of degree 1 must be a single place, and then the place is necessarily of degree 1.

We need to investigate the sets $\{E : E \geq 0, \deg(E) = n\}$ in order to know that this formulation of $\zeta_F(s)$ is well defined. However, before we do this, we set

$$Z_F(t) = \sum_{n=1}^{\infty} A_n t^n,$$

and call this the Z-function associated to the function field $F/\mathbb{F}_q$.

Lemma 7.1. *If n is a positive integer, then there are only finitely many positive divisors of degree n.*

Proof. If $E = \sum_i e_i P_i$ is a positive divisor of degree n, then $n = \sum_i e_i \deg(P_i)$, so $0 \leq e_i \leq n$ and $\deg(P_i) \leq n$. It is then enough to prove that there are only finitely many places of degree less than or equal to n. To do this, suppose that F is the function field of a curve X defined over $\mathbb{F}_q$. We have seen that the places of degree r correspond to sets of r conjugate points with coefficients in $\mathbb{F}_{q^r}$. Since there are only finitely many such points, since $\mathbb{F}_{q^r}$ is

finite, there can be only finitely many places of degree r. Since this is true for any r, there are only finitely many places of degree $\leq n$. □

Let C_F be the divisor class group of $F/\mathbb{F}_q$. We set C_F^0 to be the subgroup of divisor classes of degree 0.

Lemma 7.2. *The group C_F^0 is a finite group.*

Proof. Let E be a divisor of degree $n \geq g$, where g is the genus of $F/\mathbb{F}_q$. We set $C_F^n = \{[D] : \deg(D) = n\}$. It is elementary to check that C_F^n is the coset $C_F^0 + [E]$. Therefore, $|C_F^0| = |C_F^n|$. It then suffices to show that C_F^n is a finite set. By Riemann's theorem, $\dim(E) \geq \deg(E) + 1 - g \geq 1$ by the choice of E. Therefore, there is a nonzero $f \in L(E)$; the divisor $E + (f)$ is then a positive divisor similar to E. Thus, every divisor of degree n is equivalent to a positive divisor (of degree n). This yields $C_F^n = \{[C] : C \geq 0, \deg(C) = n\}$. By the previous lemma, there are only finitely many positive divisors of a given degree, so $|C_F^n| < \infty$, and so $|C_F^0| < \infty$, as desired. □

Corollary 7.3. *If $A_n = |\{E \in \mathrm{Div}(F/\mathbb{F}_q) : E \geq 0, \deg(E) = n\}|$, then A_n is a nonnegative integer.*

The degree map $\deg : \mathrm{Div}(F/\mathbb{F}_q) \to \mathbb{Z}$ is a group homomorphism. The image is a subgroup of $\mathbb{Z}$; therefore, there is a positive integer ∂ such that the image is $\partial\mathbb{Z}$. The integer ∂ then can be characterized as the smallest degree of a nonzero divisor, or the greatest common divisor of the degrees of places. Furthermore, every divisor's degree is a multiple of ∂. We will see later that $\partial = 1$.

Definition 7.4. *The class number of $F/\mathbb{F}_q$ is the integer $h = |C_F^0|$.*

Note that, from the proof of the lemma above, if C_F^n is the set of divisor classes of degree n, and if C_F^n is nonempty, then C_F^n is a coset of C_F^0, so $|C_F^n| = h$. We now investigate the function $Z(t)$. We need further information about the numbers A_n.

Lemma 7.5. *Let $[C] \in C_F$. Then*

$$|\{A \in [C] : A \geq 0\}| = \frac{1}{q-1}(q^{\dim(C)} - 1).$$

Furthermore, if $n > 2g - 2$ and $\partial \mid n$, then

$$A_n = \frac{h}{q-1}\left(q^{n+1-g} - 1\right).$$

Proof. Let C be a divisor, and set $n = \dim(C)$. Then any divisor $A \in [C]$ is of the form $C + (f)$ for some $f \in F^*$, and if $A \geq 0$, then $f \in L(C)$, by definition of $L(C)$. Now, $C + (f) = C + (g)$ if and only if $(f) - (g) = 0$, or $(fg^{-1}) = 0$. Therefore, as a principal divisor is 0 only when the function is a constant, we see that $C + (f) = C + (g)$ if and

only if $g = \alpha f$ for some $\alpha \in \mathbb{F}_q$. Therefore, for every $A \in [C]$ with $A \geq 0$, there are $q-1$ elements $f \in L(C)$ with $A = C + (f)$. Furthermore, the number of nonzero $f \in L(C)$ is $|L(C)| - 1 = q^{\dim(C)} - 1$. Thus, the formula

$$|\{A \in [C] : A \geq 0\}| = \frac{1}{q-1}(q^{\dim(C)} - 1)$$

is true. For the second statement, suppose that $n > 2g-2$ and $\partial \mid n$. The divisibility condition implies that there are divisors of degree n. Moreover, if $\deg(C) = n$, then $\dim(C) = \deg(C) + 1 - g$ by the Riemann-Roch theorem. By the first part of the lemma and the definitions of A_n and h, we have

$$\begin{aligned} A_n &= |\{[C] : \deg(C) = n\}| \cdot |\{A \in [C] : A \geq 0\}| \\ &= h \cdot \frac{1}{q-1}\left(q^{\dim(C)} - 1\right) = \frac{h}{q-1}\left(q^{n+1-g} - 1\right) \end{aligned}$$

since $h = |C_F^0| = |C_F^n|$, which was shown in the proof of the previous lemma. □

Corollary 7.6. *The power series $Z(t) = \sum_{n=0}^{\infty} A_n t^n$ converges for $|t| < q^{-1}$.*

Proof. We have $Z(t) = \sum_{m=0}^{\infty} A_{m\partial} t^{m\partial}$. A little algebra and a basic fact about power series shows that if

$$\lim_{n\to\infty} \left|\frac{A_{n+1}}{A_n}\right| = c,$$

then the power series converges for $|t| < c^{-1}$. From the previous lemma, we have, for n large enough,

$$\begin{aligned} \lim_{n\to\infty} \left|\frac{A_{n+1}}{A_n}\right| &= \lim_{n\to\infty} \frac{\frac{h}{q-1}\left(q^{n+2-g} - 1\right)}{\frac{h}{q-1}\left(q^{n+1-g} - 1\right)} = \lim_{n\to\infty} \frac{\left(q^{n+2-g} - 1\right)}{\left(q^{n+1-g} - 1\right)} \\ &= \lim_{n\to\infty} \frac{\left(q - q^{-n+g+1}\right)}{\left(1 - q^{-n+g+1}\right)} = q. \end{aligned}$$

Thus, the power series converges for $|t| < q^{-1}$. □

Lemma 7.7. *Let $F/\mathbb{F}_q$ have genus 0. Then $h = 1$.*

Proof. We need to prove that every divisor class of degree 0 is principal. Let C be a divisor with $\deg(C) = 0$. By Riemann's theorem, we have $\dim(C) \geq \deg(C) + 1 - g = 1$. Thus, there is a nonzero $f \in L(C)$. Then $(f) + C \geq 0$, and $(f) + C$ also has degree 0. This forces $(f) + C = 0$, so $C = -(f) = (f^{-1})$ is principal. □

To help with the proof of the following proposition, recall that the power series representation of $1/(1-x)$ in the range $|x| < 1$ is $\sum_{n=0}^{\infty} x^n$. Moreover, $\sum_{n=r}^{\infty} x^n = x^r/(1-x)$ on this range. We now see that $Z(t)$ is a rational function. This indicates how much simpler zeta functions for curves over a finite field are than zeta functions of number fields.

Proposition 7.8. *Let g be the genus of $F/\mathbb{F}_q$. Consider $Z(t)$ on the interval of convergence $|t| < q^{-1}$.*

1. *If $g = 0$, then*
$$Z(t) = \frac{1}{q-1}\left(\frac{q}{1-(qt)^\partial} - \frac{1}{1-t^\partial}\right).$$

2. *If $g \geq 1$, then $Z(t) = F(t) + G(t)$, where*
$$F(t) = \frac{1}{q-1}\sum_{0\leq \deg(C)\leq 2g-2} q^{\dim(C)}t^{\deg(C)},$$
where the sum is over all divisor classes $[C]$ with $0 \leq \deg(C) \leq 2g-2$, and
$$G(t) = \frac{h}{q-1}\left(\frac{q^{1-g}(qt)^{2g-2+\partial}}{1-(qt)^\partial} - \frac{1}{1-t^\partial}\right).$$

Proof. First, suppose that $g = 0$. By the previous lemmas, and since $A_n = 0$ if ∂ does not divide n,
$$\begin{aligned}\sum_{n=0}^{\infty} A_n t^n &= \sum_{m=0}^{\infty} A_{\partial m} t^{\partial m} = \sum_{m=0}^{\infty} \frac{1}{q-1}\left(q^{\partial m+1}-1\right)t^{\partial m} \\ &= \frac{1}{q-1}\left(q\sum_{m=0}^{\infty}((qt)^\partial)^m - \sum_{m=0}^{\infty}(t^\partial)^m\right) \\ &= \frac{1}{q-1}\left(\frac{q}{1-(qt)^\partial} - \frac{1}{1-t^\partial}\right).\end{aligned}$$
The final equality holds for $|t| < q^{-1}$, since both power series converge on this range to the respective rational functions.

For the second part, suppose that $g \geq 1$. We have, by the definition of A_n, and the lemmas,
$$\begin{aligned}\sum_{n=0}^{\infty} A_n t^n &= \sum_{[C],\deg(C)\geq 0} |A \in [C] : A \geq 0| \cdot t^{\deg(C)} = \sum_{[C],\deg(C)\geq 0}\left(\frac{q^{\dim(C)}-1}{q-1}\right)t^{\deg(C)} \\ &= \frac{1}{q-1}\sum_{[C],\deg(C)\geq 0} q^{\dim(C)}t^{\deg(C)} - \frac{1}{q-1}\sum_{[C],\deg(C)\geq 0} t^{\deg(C)} \\ &= \frac{1}{q-1}\sum_{0\leq\deg([C])\leq 2g-2} q^{\dim(C)}t^{\deg(C)} + \frac{1}{q-1}\sum_{\deg([C])>2g-2} q^{\dim(C)}t^{\deg(C)} \\ &\quad -\frac{1}{q-1}\sum_{\deg([C])\geq 0} t^{\deg(C)}\end{aligned}$$

The first term, by definition, is $F(t)$. To analyze the second and third terms, recall that if n is divisible by ∂, then there are exactly h divisor classes of degree n; the set of these divisor classes is the coset $\{E + C : \deg(E) = n, [C] \in C_F^0\}$ of C_F^0, and $h = |C_F^0|$. Therefore, the third term simplifies as

$$\frac{-1}{q-1} \sum_{\deg([C]) \geq 0} t^{\deg(C)} = \frac{-h}{q-1} \sum_{m=0}^{\infty} t^{\partial m} = \frac{-h}{q-1} \left(\frac{1}{1-t^{\partial}} \right).$$

For the second term, we have, by Riemann-Roch,

$$\begin{aligned}
\frac{1}{q-1} \sum_{\deg([C]) > 2g-2} q^{\dim(C)} t^{\deg(C)} &= \frac{1}{q-1} \sum_{\deg([C]) > 2g-2} q^{\deg(C)+1-g} t^{\deg(C)} \\
&= \frac{q^{1-g}}{q-1} \sum_{\deg([C]) > 2g-2} (qt)^{\deg(C)} = \frac{hq^{1-g}}{q-1} \sum_{\partial m > 2g-2} \left((qt)^{\partial} \right)^m \\
&= \frac{hq^{1-g}}{q-1} \sum_{m=\frac{2g-2}{\partial}}^{\infty} \left((qt)^{\partial} \right)^m = \frac{hq^{1-g}(qt)^{2g-2+\partial}}{q-1} \sum_{m=0}^{\infty} \left((qt)^{\partial} \right)^m \\
&= \frac{hq^{1-g}(qt)^{2g-2+\partial}}{q-1} \left(\frac{1}{1-(qt)^{\partial}} \right)
\end{aligned}$$

because ∂ divides $2g-2$, since $2g-2$ is the degree of a divisor, namely the canonical divisor. The sum of the second and third terms is $G(t)$. Since the sums for $G(t)$ converge on $|t| < q^{-1}$, the series for $Z(t)$ converges on the same interval. □

Corollary 7.9. *The function $Z(t)$ is a rational function with a simple pole at $t = 1$.*

Proof. The formula for $F(t)$ shows that $F(t)$ is a polynomial, since there are only finitely many divisor classes $[C]$ with $0 \leq \deg(C) \leq 2g-2$. The formula for $G(t)$ shows that it is a rational function with a simple pole at $t = 1$. Therefore, $Z(t) = F(t) + G(t)$ is a rational function with a simple pole at $t = 1$. □

We now give a product representation for $Z(t)$.

Proposition 7.10. *The function $Z(t)$ can be represented on $|t| < q^{-1}$ as*

$$Z(t) = \prod_{P \in \mathbf{P}_F} (1 - t^{\deg(P)})^{-1}.$$

Therefore, $Z(t) \neq 0$ for $|t| < q^{-1}$.

Proof. We need a fact about convergence of infinite products: the product $\prod_{n=1}^{\infty}(1 + a_n)$ converges absolutely if and only if $\sum_{n=0}^{\infty} a_n$ converges absolutely. This fact can be found in most complex analysis books. Therefore, the product above converges absolutely on $|t| < q^{-1}$

since

$$\sum_{P\in\mathbf{P}_F} t^{\deg(P)} \le \sum_{n=0}^{\infty} A_n t^n = Z(t)$$

converges absolutely on $|t| < q^{-1}$. By writing each term of the product as a geometric series, and doing lots of rearranging of terms, we have

$$\begin{aligned}\prod_{P\in\mathbf{P}_F}\left(\frac{1}{1-t^{\deg(P)}}\right) &= \prod_{P\in\mathbf{P}_F}\sum_{n=0}^{\infty} t^{n\deg(P)} \\ &= \sum_{E\ge 0} t^{\deg(E)} = \sum_{n=0}^{\infty} A_n t^n \\ &= Z(t).\end{aligned}$$

□

To prove that $\partial = 1$, we will need to consider how function fields behave under base extension. Let $X = Z(f)$ be a nonsingular irreducible projective curve defined over $\mathbb{F}_q$, and let X be its function field. Then $F = \mathbb{F}_q(s,t)$, where $s = x/z$ and $t = y/z$, viewed as rational functions on X. We assume that $\mathbb{F}_q$ is the exact constant field of $F/\mathbb{F}_q$. Let $r \ge 1$, and consider the extension $\mathbb{F}_{q^r}/\mathbb{F}_q$. We can view X as a curve over $\mathbb{F}_{q^r}$, and the function field of X over $\mathbb{F}_{q^r}$ is $\mathbb{F}_{q^r}(s,t)$. We write F_r for this function field. What we need to know is information about places of F_r in relation to places of F. Note that, in any case, our curve X is the set of all points $(a:b:c) \in \mathbb{P}^2(\overline{\mathbb{F}_q})$ such that $f(a,b,c) = 0$, where $\overline{\mathbb{F}_q}$ is an algebraic closure of $\mathbb{F}_q$ (or of $\mathbb{F}_{q^r}$). If P is the place of F that corresponds to $(a:b:c)$, then

$$P = \{\varphi \in F : \varphi(a,b,c) = 0\}.$$

Also, if P' is the place of F_r that corresponds to $(a:b:c)$, then

$$P' = \{\varphi \in F_r : \varphi(a,b,c) = 0\}.$$

Thus, $P' \cap F = P$.

In the proof of the following lemma, we need to know some facts about finite fields. First, recall that if q is any prime power, then the fields containing $\mathbb{F}_q$ are exactly the fields of the form $\mathbb{F}_{q^r}$ for some $r \ge 1$. Conversely, if T is a subfield of $\mathbb{F}_{q^r}$ containing $\mathbb{F}_q$, then $T = \mathbb{F}_{q^s}$ for some integer s that divides r, since if $t = [\mathbb{F}_{q^r} : T]$, then $q^r = |\mathbb{F}_{q^r}| = |T|^t = q^{st}$. Moreover, the field $\mathbb{F}_{q^s}$ is characterized as the subfield of $\mathbb{F}_{q^r}$ consisting of the solutions to the equation $x^{q^s} = x$.

Lemma 7.11. *Let P be a place of F, and let P' be a place of F_r for which $P' \cap F = P$. If $m = \deg(P)$, then $\deg(P') = m/\gcd(m,r)$. Moreover, there are $\gcd(m,r)$ many such P'.*

Proof. Let X be a curve whose function field over $\mathbb{F}_q$ is F. Choose a point $(a:b:c) \in X$

whose corresponding place in F is P and whose place in F_r is P'. The residue field of P is $\mathbb{F}_q(a,b,c)$ and the residue field of P' is $\mathbb{F}_{q^r}(a,b,c)$. Then, by definition, $\deg(P') = [\mathbb{F}_{q^r}(a,b,c) : \mathbb{F}_{q^r}]$ and $\deg(P) = [\mathbb{F}_q(a,b,c) : \mathbb{F}_q]$. Set $L = \mathbb{F}_q(a,b,c)$. Then the residue field of P' is the composite field $L\mathbb{F}_{q^r}$. By the theorem of natural irrationalities from Galois theory, $[L\mathbb{F}_{q^r} : \mathbb{F}_{q^r}] = [L : L \cap \mathbb{F}_{q^r}]$. Note that, since $m = \deg(P)$, we have $L = \mathbb{F}_{q^m}$. We claim that $\mathbb{F}_{q^r} \cap \mathbb{F}_{q^m} = \mathbb{F}_{q^d}$, where $d = \gcd(m,r)$. If this is true, then

$$[L : L \cap \mathbb{F}_{q^r}] = \frac{[L : \mathbb{F}_q]}{[L \cap \mathbb{F}_{q^r} : \mathbb{F}_q]} = \frac{[\mathbb{F}_{q^m} : \mathbb{F}_q]}{[\mathbb{F}_{q^d} : \mathbb{F}_q]} = \frac{m}{d},$$

as desired. To prove the claim, note that $\mathbb{F}_{q^d} \subseteq \mathbb{F}_{q^r} \cap \mathbb{F}_{q^m}$ since d divides both r and m. For the reverse inclusion, let $a \in \mathbb{F}_{q^r} \cap \mathbb{F}_{q^m}$. Then $a^{q^r} = a$ and $a^{q^m} = a$. Since $d = \gcd(m,r)$, we may write $d = \alpha m + \beta r$ for some integers α, β. Since d, m, r are all positive, one of α, β is positive and the other is negative. Suppose that $\alpha > 0$. Then $d - \beta r = \alpha m$. A simple induction shows that $a^{q^{\alpha m}} = a$ and $a^{q^{-\beta r}} = a$. Therefore,

$$a = a^{q^{\alpha m}} = a^{q^{d-\beta r}} = a^{q^d q^{-\beta r}} = \left(a^{q^{-\beta r}}\right)^{q^d} = a^{q^d}.$$

This shows that $a \in \mathbb{F}_{q^d}$, as desired. Note that as a consequence of this, we see that the composite $\mathbb{F}_{q^m}\mathbb{F}_{q^r} = \mathbb{F}_{q^l}$, where $l = mr/d = \operatorname{lcm}(m,r)$.

To prove the second statement, let P be a place of F of degree m, and let S be the set of points of X whose corresponding place is P. This set has m elements. The places of F_r lying over P are exactly the places corresponding to points in S. However, each such place of F_r corresponds to l/r points; this is because if p is a point in S, then its residue field $\mathbb{F}_{q^r}(p)$ with respect to F_r has degree $[\mathbb{F}_{q^r}(p) : \mathbb{F}_{q^r}] = [\mathbb{F}_{q^l} : \mathbb{F}_{q^r}] = l/r$, as mentioned above. Now, $l = mr/d$, so $l/r = m/d$. Therefore, if $P'_1, \ldots, P'_t$ are the places of F_r lying over P, each corresponds to m/d points of S, and every point of S corresponds to exactly one of the P'_i. Therefore, there are d such P'_i. This finishes the proof. □

To help prove the following proposition, we need a result about roots of unity. Let r and m be positive integers and set $d = \gcd(r,m)$. Then

$$\left(x^{r/d} - 1\right)^d = \prod_{\delta}(x - \delta^m),$$

where the product runs over the r-th roots of unity in $\mathbb{C}$. That is, δ ranges over the elements $\{\exp(2\pi i/r) : 0 \le i < r\}$. To see why this equation is true, we note that both sides are monic polynomials of degree r. It is enough to see that they have the same roots. The roots of the left hand side are precisely the r/d-th roots of unity. If δ is an r-th root of unity, then $(\delta^m)^{r/d} = \delta^{r(m/d)} = 1$. Conversely, any r/d-th root of unity is of the form δ^m for some r-th root of unity δ, as a group theory exercise shows. If we substitute $x = t^{-m}$ and then multiply

both sides by t^{mr}, a little algebra will show that

$$(1 - t^{mr/d})^d = \prod_{\delta}(1 - (t\delta)^m).$$

We will use this in the following proposition.

Proposition 7.12. *Let $Z_r(t)$ be the Z-function associated to F_r. Then $Z_r(t^r) = \prod_\delta Z(\delta t)$, where δ runs over all r-th roots of unity in $\mathbb{C}$.*

Proof. Since two complex analytic functions are equal if they agree on a nontrivial open disk of the complex plane, and because both sides of the equation above are rational functions in t, it is enough to prove the equality for $|t| < q^{-1}$. We will write $P' \mid P$ if $P' \cap F = P$. In the region $|t| < q^{-1}$, the product representation yields

$$Z_r(t^r) = \prod_{P' \in \mathbf{P}_{F_r}} (1 - t^{r \deg(P')})^{-1} = \prod_{P \in \mathbf{P}_F} \prod_{P' | P} (1 - t^{r \deg(P')})^{-1}.$$

For a fixed $P \in \mathbf{P}_F$, let $m = \deg(P)$ and $d = \gcd(r, m)$. We then have, by the lemma,

$$\begin{aligned}\prod_{P' | P} \left(1 - t^{r \deg(P')}\right) &= (1 - t^{rm/d})^d \\ &= \prod_{\delta}(1 - (\delta t)^m) \\ &= \prod_{\delta} \left(1 - (\delta t)^{\deg(P)}\right).\end{aligned}$$

Therefore,

$$\begin{aligned}Z_r(t^r) &= \prod_{P \in \mathbf{P}_F} \prod_{\delta} \left(1 - (\delta t)^{\deg(P)}\right)^{-1} \\ &= \prod_{\delta} \prod_{P \in \mathbf{P}_F} \left(1 - (\delta t)^{\deg(P)}\right)^{-1} = \prod_{\delta} Z(\delta t).\end{aligned}$$

□

Corollary 7.13. *If $\partial = \gcd\{\deg(E) : E \in D_F\}$, then $\partial = 1$.*

Proof. Set $r = \partial$. Then $\left(1 - (\delta t)^{\deg(P)}\right) = 1 - t^{\deg(P)}$ since $\deg(P)$ is a multiple of ∂, so $\delta^{\deg(P)} = 1$. Therefore,

$$Z(\delta t) = \prod_{P \in \mathbf{P}_F} \left(1 - (\delta t)^{\deg(P)}\right)^{-1} = \prod_{P \in \mathbf{P}_F} \left(1 - t^{\deg(P)}\right)^{-1} = Z(t).$$

Therefore, by the proposition, $Z_r(t^r) = Z(t)^r$. By the description of $Z(t)$ in Proposition 7.10, applied to $Z_r(t)$, the Z-function associated to $F_r/\mathbb{F}_{q^r}$, we see that $Z_r(t^r)$ has a simple

pole at $t = 1$. However, $Z(t)^r$ has a pole of order r at $t = 1$. Thus, as $Z_r(t^r) = Z(t)^r$, we see that $r = 1$. In other words, $\partial = 1$. □

This corollary is not true if the base field is infinite. For example, if F is the function field of the curve $x^2 + y^2 + z^2 = 0$ over $\mathbb{R}$, then the curve has no $\mathbb{R}$-rational point. Since the residue field of any point is a finite extension of $\mathbb{R}$, each residue field must then be $\mathbb{C}$. Therefore, the degree of every place is 2, and so $\partial = 2$ for this function field.

Corollary 7.14. *Any function field $F/\mathbb{F}_q$ of genus 0 is a rational function field, and*

$$Z(t) = \frac{1}{(1-t)(1-qt)}.$$

Proof. Since $\partial = 1$, there is a divisor E of degree 1. By Riemann-Roch, $\dim(E) = 2$, so $L(E)$ is nonzero. Taking a nonzero $f \in L(E)$, we have $(f) + E \geq 0$, a positive divisor of degree 1. Then $(f) + E = P$ for some single place P, which is then necessarily of degree 1. Therefore, $F/\mathbb{F}_q$ has a place of degree 1. We have proven that a function field of genus 0 with a rational point is a rational function field, which says that F is isomorphic to $\mathbb{F}_q(x)$. Finally, from Proposition 7.10, we have

$$\begin{aligned} Z(t) &= \frac{1}{q-1}\left(\frac{q}{1-(qt)^\partial} - \frac{1}{1-t^\partial}\right) = \frac{1}{q-1}\left(\frac{q}{1-qt} - \frac{1}{1-t}\right) \\ &= \frac{1}{(1-qt)(1-t)}, \end{aligned}$$

as a little algebra shows. □

Corollary 7.15. *If $F/\mathbb{F}_q$ has genus $g \geq 1$, then $Z(t) = F(t) + G(t)$, where*

$$F(t) = \frac{1}{q-1} \sum_{0 \leq \deg([C]) \leq 2g-2} q^{\dim(C)} t^{\deg(C)}$$

and

$$G(t) = \frac{h}{q-1}\left(\frac{q^g t^{2g-1}}{1-qt} - \frac{1}{1-t}\right).$$

The following result is an analogue of the functional equation mentioned at the beginning of this chapter. We will use the functional equation to get information about the zeros of $Z(t)$.

Proposition 7.16. *If $Z(t)$ is the Z-function of $F/\mathbb{F}_q$, then*

$$Z(t) = q^{g-1} t^{2g-2} Z(1/qt).$$

First suppose that $g = 0$. Then $Z(t) = 1/(1-t)(1-qt)$. Therefore,

$$\begin{aligned} q^{g-1}t^{2g-2}Z(1/qt) &= \frac{q^{-1}t^{-2}}{(1-(qt)^{-1})(1-q(qt)^{-1})} \\ &= \frac{1}{(qt-1)(t-1)} = \frac{1}{(1-t)(1-qt)}. \end{aligned}$$

Next, suppose that $g \geq 1$. We write $Z(t) = F(t) + G(t)$ as earlier. Let C be a canonical divisor. Recall that $\deg(C) = 2g-2$. Then, by the Riemann-Roch theorem,

$$\begin{aligned} (q-1)F(t) &= \sum_{0 \leq \deg([E]) \leq 2g-2} q^{\dim(E)} t^{\deg(E)} \\ &= \sum_{0 \leq \deg([E]) \leq 2g-2} q^{\deg(E)+1-g+\dim(C-E)} t^{\deg(E)} \\ &= q^{g-1}t^{2g-2} \sum_{0 \leq \deg([E]) \leq 2g-2} q^{\deg(E)-(2g-2)+\dim(C-E)} t^{\deg(E)-(2g-2)} \end{aligned}$$

If E is a divisor with $0 \leq \deg(C) \leq 2g-2$, then $\deg(C-E) = \deg(C) - \deg(E) = 2g-2-\deg(E)$, and so $0 \leq \deg(C-E) \leq 2g-2$. Furthermore, by Riemann-Roch, $\dim(C-E) = \deg(C-E)+1-g+\dim(E)$. Moreover, as $[E]$ ranges over divisors of degree between 0 and $2g-2$, so does $[C-E]$. Therefore, setting $D = C-E$, we have

$$\begin{aligned} (q-1)F(t) &= q^{g-1}t^{2g-2} \sum_{0 \leq \deg([D]) \leq 2g-2} q^{\dim(D)-\deg(D)} t^{-\deg(D)} \\ &\qquad q^{g-1}t^{2g-2} \sum_{0 \leq \deg([D]) \leq 2g-2} q^{\dim(D)} \left(\frac{1}{qt}\right)^{\deg(D)} \\ &= q^{g-1}t^{2g-2}F(1/qt). \end{aligned}$$

Next, for $G(t)$, we have, by the corollary,

$$\begin{aligned} q^{g-1}t^{2g-2}G(1/qt) &= \frac{h}{q-1} q^{g-1}t^{2g-2} \left(q^g \left(\frac{1}{qt}\right)^{2g-1} \frac{1}{(1-q(qt)^{-1})} - \frac{1}{(1-(qt)^{-1})} \right) \\ &= \frac{h}{q-1} \left(\frac{1}{t} \frac{1}{(1-t^{-1})} - \frac{q^{g-1}t^{2g-2}}{(1-(qt)^{-1})} \right) \\ &= \frac{h}{q-1} \left(\frac{1}{t-1} - \frac{q^g t^{2g-1}}{qt-1} \right) = G(t). \end{aligned}$$

Therefore, $q^{g-1}t^{2g-2}Z(1/qt) = Z(t)$.

We can phrase the functional equation in terms of the zeta function $\zeta(s)$. Recall that

$\zeta(s) = Z(q^{-s})$. Substituting $t = q^{-s}$ in the functional equation yields

$$\begin{aligned}\zeta(s) &= q^{g-1}(q^{-s})^{2g-2}Z(q^{s-1}) = q^{g-1}(q^{-s})^{2g-2}\zeta(1-s)\\ &= q^{(g-1)(1-2s)}\zeta\left(1-s\right).\end{aligned}$$

In this case, unlike the classical case, we have no "trivial zeros." Also, if s is a zero of $\zeta(s)$, then so is $1 - s$.

The function $Z(t)$ is a rational function with denominator $(1-t)(1-qt)$. We work with the numerator of $Z(t)$.

Definition 7.17. *The polynomial $L(t) = (1-t)(1-qt)Z(t)$ is called the L-polynomial of $F/\mathbb{F}_q$.*

By the previous two corollaries, we see that $L(t) = 1$ if the genus of $F/\mathbb{F}_q$ is 0 and that in any case, $\deg(L(t)) \leq 2g$; this comes from the fact that $F(t)$ is a polynomial of degree at most $2g-2$, and $G(t)$ is a rational function of degree $2g-2$. Thus, the degree of $Z(t)$ is at most $2g-2$. We will see that much information is encoded in this polynomial. Note that $Z(t)$ is determined by $L(t)$, as

$$Z(t) = \frac{L(t)}{(1-t)(1-qt)}.$$

Therefore, $L(t)$ encodes all the information about the numbers A_n that $Z(t)$ encodes.

Theorem 7.18. *Let $L(t)$ be the L-polynomial of the function field $F/\mathbb{F}_q$.*

1. *$L(t)$ has coefficients in $\mathbb{Z}$, and $\deg(L(t)) = 2g$.*
2. *$L(t) = q^g t^{2g} L(1/qt)$;*
3. *$L(1) = h$, the class number of $F/\mathbb{F}_q$;*
4. *If $L(t) = a_0 + a_1 t + \cdots a_{2g}t^{2g}$, then $a_0 = 1$, $a_{2g} = q^g$, and $a_1 = N - (q+1)$, where N is the number of places of $\mathbf{P}_F$ of degree 1.*

Proof. If the genus $g = 0$, then $L(t) = 1$, and the result is trivial in this case. We thus assume that $g \geq 1$. Since $L(t) = (1-t)(1-qt)\sum_{n=0}^{\infty} A_n t^n$, the polynomial $L(t)$ is a power series in t with integer coefficients. Thus, its coefficients as a polynomial are the same integer coefficients. We have already remarked that $\deg(L(t)) \leq 2g$. When we prove that the coefficient $a_{2g} = q^g$, we will have $\deg(L(t)) = 2g$.

For the functional equation, we have

$$Z(t) = q^{g-1}t^{2g-2}Z(1/qt).$$

Therefore,

$$\begin{aligned}
L(t) &= (1-t)(1-qt)Z(t) = (1-t)(1-qt)q^{g-1}t^{2g-2}Z(1/qt) \\
&= (1-t)(1-qt)q^{g-1}t^{2g-2}\frac{L(1/qt)}{(1-(qt)^{-1})(1-q(qt)^{-1})} \\
&= (1-t)(1-qt)q^{g-1}t^{2g-2}\frac{L(1/qt)}{(1-q^{-1}t^{-1})(1-t^{-1})} \\
&= (1-t)(1-qt)q^{g}t^{2g}\frac{L(1/qt)}{(qt-1)(t-1)} \\
&= q^g t^{2g} L(1/qt).
\end{aligned}$$

This proves the functional equation.

To prove (3), since $Z(t) = F(t) + G(t)$, we may write

$$L(t) = (1-t)(1-qt)F(t) + \frac{h}{q-1}\left(q^g t^{2g-1}(1-t) - (1-qt)\right)$$

by our description of $G(t)$. Since $F(t)$ is a polynomial, $F(1)$ is defined, and this formula then yields

$$L(1) = \frac{h}{q-1}(-(1-q)) = h.$$

Finally, write $L(t) = a_0 + a_1 t + \cdots + a_{2g}t^{2g}$. From $L(t) = (1-t)(1-qt)\sum_{n=0}^{\infty} A_n t^n$, if we multiply this out, we see that $a_0 = A_0$ and $a_1 = A_1 - (q+1)$. We have noted earlier that A_1 is the number of places of degree 1; in other words, $A_1 = N$. Therefore, $a_1 = N - (q+1)$. Also, $A_0 = 1$ since the only positive divisor of degree 0 is 0. Thus, $a_0 = 1$. Finally, the functional equation yields

$$\begin{aligned}
L(t) = q^g t^{2g} L(1/qt) &= q^g t^{2g}\left(a_0 + \frac{a_1}{qt} + \cdots + \frac{a_{2g}}{(qt)^{2g}}\right) \\
&= \frac{a_{2g}}{q^g} + \frac{a_{2g-1}}{q^{g-1}}t + \cdots + a_1 q^{g-1}t^{2g-1} + a_0 q^g t^{2g}.
\end{aligned}$$

Equating the constant terms gives $a_0 = a_{2g}/q^g$, so $a_{2g} = a_0 q^g = q^g$. This completes the proof. □

The polynomial $L(t)$ has degree $2g$. If we work over $\mathbb{C}$, then the fundamental theorem of algebra says that we can factor $L(t)$ into linear factors. It is more convenient to work with the reciprocals $\alpha_1, \ldots, \alpha_{2g}$ of the roots of $L(t)$. Since the leading coefficient of $L(t)$ is q^g, we

write

$$L(t) = q^g \prod_{i=1}^{2g}(t - \alpha_i^{-1}) = q^g \prod_{i=1}^{2g} -\alpha_i^{-1}(1 - \alpha_i t)$$
$$= \frac{q^g}{\alpha_1 \cdots \alpha_{2g}} \prod_{i=1}^{2g}(1 - \alpha_i t).$$

Now, since the constant term of $L(t)$ is 1, we see that the coefficient in front of the product is 1. Therefore,

$$L(t) = \prod_{i=1}^{2g}(1 - \alpha_i t).$$

Corollary 7.19. *If N is the number of places of degree 1 in F, then $N = q + 1 - \sum_{i=1}^{2g} \alpha_i$.*

Proof. This follows immediately from $L(t) = \prod (1 - \alpha_i t) = a_0 + a_1 t + \cdots + a_{2g} t^{2g}$ and $a_1 = N - (q+1)$, once we multiply out the product and gather together the linear terms. □

From this corollary, in order to determine the number of rational points of $F/\mathbb{F}_q$, we need to know about the reciprocals of the roots of $L(t)$. The Riemann hypothesis for function fields of curves over finite fields, proved by André Weil in 1941, gives the needed information about the α_i.

Theorem 7.20 (Hasse-Weil). *The roots of the Riemann zeta function $\zeta_F(s)$ lie on the line $\mathrm{Re}(s) = 1/2$. Therefore, if $\alpha_1, \ldots, \alpha_{2g}$ are the reciprocals of the roots of $L(t)$, then $|\alpha_i| = \sqrt{q}$.*

Because of time constraints, we will not prove this theorem. However, to relate the two statements of the theorem, recall that $\zeta(s)$ and $Z(t)$ are related by the equation $\zeta_F(s) = Z(q^{-s})$. Furthermore, since

$$Z(t) = \frac{L(t)}{(1-t)(1-qt)},$$

the roots of $L(t)$ are the zeros of $Z(t)$. So, if s is a zero of $\zeta_F(s)$, then $Z(q^{-s}) = 0$. If we write $s = 1/2 + yi$, then

$$q^{-s} = e^{-\ln(q)s} = e^{-\ln(q)(1/2+yi)}$$
$$= e^{-1/2 \ln q} e^{-y \ln(q) i}.$$

Therefore, $|q^{-s}| = \left|e^{-1/2 \ln q}\right| = q^{-1/2}$; recall that $e^{i\theta} = \cos\theta + i \sin\theta$, so $\left|e^{i\theta}\right| = 1$ for any θ. Thus, a root of $L(t)$ has absolute value $1/\sqrt{q}$, so any α_i has absolute value $\sqrt{q}$.

Corollary 7.21 (Hasse-Weil Bound). *If N is the number of places of $F/\mathbb{F}_q$ of degree 1, then*

$$|N - (q+1)| \leq 2g\sqrt{q}.$$

Proof. We have $N = q + 1 - \sum_{i=1}^{2g} \alpha_i$. Therefore,

$$\begin{aligned} |N - (q+1)| &= \left| \sum_{i=1}^{2g} \alpha_i \right| \leq \sum_{i=1}^{2g} |\alpha_i| \\ &= 2g\sqrt{q}. \end{aligned}$$

As we have seen, Hermitian curves attain this bound, so this is in general the best one can do. □