18.782 Introduction to Arithmetic Geometry **Fall 2013**
Lecture #2 **09/10/2013**

2.1 Plane conics

A *conic* is a plane projective curve C/k of degree 2. Such a curve is defined by an equation of the form

$$ax^2 + by^2 + cz^2 + dxy + exz + fyz = 0,$$

with $a, b, c, d, e, f \in k$. Assuming the characteristic of k is not 2, we can make $d = e = f = 0$ via an invertible linear transformation. First, if $a = b = c = 0$ we can make one of them nonzero by replacing a variable by its sum with another; in this case one of d, e, f must be nonzero, say d, and then replacing y with $x + y$ yields an equation with $a \neq 0$. So assume without loss of generality that $a \neq 0$. Replacing x with $x - \frac{d}{2a}y$ kills the xy term, and we can similarly kill the xz term by replacing x with $x - \frac{e}{2a}z$ (we are just completing the square). Finally, if $f \neq 0$ we can make b nonzero and then replace y with $y - \frac{f}{2b}z$ to eliminate the yz term. Each of these substitutions corresponds to an invertible linear transformation of the projective plane, as does their composition.

So we now assume $\operatorname{char}(k) \neq 2$, and that C has the diagonal form

$$ax^2 + by^2 + cz^2 = 0. \tag{1}$$

If any of the coefficients a, b, c are zero, then this curve is not irreducible.[1] For example, if the coefficient c is zero, we can factor the LHS of (1) over $\overline{k}$:

$$ax^2 + by^2 = (\sqrt{a}x + \sqrt{-b}y)(\sqrt{a}x - \sqrt{-b}y) = 0.$$

In this case $C(\overline{k})$ is the union of two projective lines that intersect at $(0 : 0 : 1)$ (but $C(k)$ might contain only one point, as when $k = \mathbb{Q}$ and $a, b > 0$, for example).

We now summarize this discussion with the following theorem.

Theorem 2.1. *Over a field whose characteristic is not* 2*, every geometrically irreducible conic is isomorphic to a diagonal curve* $ax^2 + by^2 + cz^2 = 0$ *with* $abc \neq 0$.

Remark 2.2. This does not hold in characteristic 2.

2.2 Parameterization of rational points on a conic

Suppose $(x_0 : y_0 : z_0)$ is a rational point on the diagonal conic $C\colon ax^2 + by^2 + cz^2 = 0$. Without loss of generality, we assume $z_0 \neq 0$ and consider the substitution

$$x = x_0W + U, \qquad y = y_0W + V, \qquad z = z_0W \tag{2}$$

[1] In Lecture 1 we defined a plane projective curve $f(x, y, z) = 0$ to be reducible if $f = gh$ for some $g, h \in \overline{k}[x, y, z]$, where $\overline{k}$ is the algebraic closure of k. Some authors distinguish between irreducibility over k versus $\overline{k}$, referring to the latter as *geometric* (or *absolute*) irreducibility. For us, irreducible will always mean geometrically irreducible.

Our definition of a plane projective curve $f(x, y, z) = 0$ requires f to have no repeated factors in $\overline{k}[x, y, z]$, which precludes the case where two of a, b, c are zero. In more general settings, curves defined by a polynomial with repeated factors are said to be *non-reduced*. In this course all curves are reduced.

Andrew V. Sutherland

where U, V, W denote three new variables. We then have

$$\begin{aligned}
a(x_0W+U)^2 + b(y_0W+V)^2 + c(z_0W)^2 &= 0\\
(ax_0^2+by_0^2+cz_0^2)W + 2(ax_0U+by_0V)W + aU^2 + bV^2 &= 0\\
2(ax_0U+by_0V)W &= -aU^2 - bV^2,
\end{aligned}$$

where we have used $ax_0^2 + by_0^2 + cz_0^2 = 0$ to eliminate the quadratic term in W. After rescaling by $2(ax_0u + by_0v)$ and substituting for W in (2) we obtain the parameterization

$$\begin{aligned}
x &= x_0(-aU^2 - bV^2) + 2(ax_0U + by_0V)U = ax_0U^2 + 2by_0UV - bx_0V^2 = Q_1(U,V)\\
y &= y_0(-aU^2 - bV^2) + 2(ax_0U + by_0V)V = -ay_0U^2 + 2ax_0UV + by_0V^2 = Q_2(U,V)\\
z &= z_0(-aU^2 - bV^2) = -az_0U^2 - bz_0V^2 = Q_3(U,V)
\end{aligned}$$

Thus $(Q_1(U,V) : Q_2(U,V) : Q_3(U,V))$ is a polynomial map defined over k that sends each projective point $(U : V)$ on $\mathbb{P}^1$ to a point on the curve C. Moreover, we can recover the point $(U : V)$ via the inverse map from C to $\mathbb{P}^1$ defined by

$$U = x - \frac{x_0}{z_0}z, \qquad V = y - \frac{y_0}{z_0}z.$$

Thus we have an invertible map from C to $\mathbb{P}^1$ that is given by rational (in fact polynomial) functions that are defined at every point (such a map is said to be *regular*). In this situation we regard C and $\mathbb{P}^1$ as isomorphic curves. This yields the following theorem.

Theorem 2.3. *Let C/k be a geometrically irreducible conic with a k-rational point and assume that* $\mathrm{char}(k) \neq 2$. *Then C is isomorphic over k to the projective line $\mathbb{P}^1$.*

Remark 2.4. This theorem also holds when $\mathrm{char}(k) = 2$, but we will not prove this.

2.3 Conics over $\mathbb{Q}$

We now consider the case $k = \mathbb{Q}$. Given a diagonal conic

$$ax^2 + by^2 + cz^2 = 0$$

with $abc \neq 0$, we wish to either find a rational point (which we can then use to parameterize all the rational points), or prove that there are none. After clearing denominators we can assume a, b, c are nonzero integers, and we note that if they all have the same sign then there are clearly no rational points. So let us assume that this is not the case, and without loss of generality suppose that $a > 0$ and $b, c < 0$. Multiplying both sides by a and setting $d = -ab$ and $n = -ac$, we can put our curve in the form

$$x^2 - dy^2 = nz^2, \tag{3}$$

where d and n are positive integers that we may assume are square-free. Solving this equation is equivalent to expressing $n = (\frac{x}{z} + \frac{y}{z}\sqrt{d})(\frac{x}{z} - \frac{y}{z}\sqrt{d})$ as the norm of an element of the real quadratic field $\mathbb{Q}(\sqrt{d})$.

We now present a recursive procedure for doing this, based on Legendre's method of descent; the algorithm we give here is adapted from [1, Alg. I]. The basic idea is to either determine that there are no integer solutions to (3) (and hence no rational solutions), or to reduce the problem to finding a solution to a similar equation with smaller values of d or n (this is why it is called a *descent*). In order to facilitate the recursion, we allow d and n to also take negative values (but still insist that they be square-free).

Given square-free integers d and n, the procedure $\textsc{Solve}(d, n)$ either returns an integer solution to (3), or determines that no solution exists; we use the notation **`fail`** to indicate that the latter has occured.

```
SOLVE$(d, n)$
  1. If $d, n < 0$ then fail.
  2. If $|d| > |n|$ then let $(x_0, y_0, z_0) =$ SOLVE$(n, d)$ and return $(x_0, z_0, y_0)$.
  3. If $d = 1$ return $(1, 1, 0)$; if $n = 1$ return $(1, 0, 1)$; if $d = -n$ return $(0, 1, 1)$.
  4. If $d = n$ then let $(x_0, y_0, z_0) =$ SOLVE$(-1, d)$ and return $(dz_0, x_0, y_0)$.
  5. If $d$ is not a quadratic residue modulo $n$ then fail.
  6. Let $x_0^2 \equiv d \bmod n$, with $|x_0| \leq |n|/2$, and let $t = t_1 t_2^2 = (x_0^2 - d)/n$ with $t_1$ square-free.
  7. Let $(x_1, y_1, z_1) =$ SOLVE$(d, t_1)$ and return $(x_0 x_1 + d y_1, x_0 y_1 + x_1, t_1 t_2 z_1)$.
```

It is clear that if the algorithm **fail**s in steps 1 or 5 then (3) has no solutions, and that the solutions returned in step 3 are all correct. Assuming the algorithm works correctly when $|d| \leq |n|$, then the solution returned in step 3 is clearly correct, and in step 4 with $d = n$, if SOLVE$(-1, d)$ succeeds then we have

$$\begin{aligned} x_0^2 + y_0^2 &= dz_0^2 \\ dx_0^2 + dy_0^2 &= (dz_0)^2 \\ (dz_0)^2 - dx_0^2 &= dy_0^2 = ny_0^2, \end{aligned}$$

and therefore the solution (dz_0, x_0, y_0) is correct (note that -1 and d are both square-free, assuming the input d is, so our square-free constraint is preserved in the recursive call).

It remains to show that the solution returned in step 7 is correct, and that the algorithm is guaranteed to terminate. If we reach step 6 then we have $|d| < |n|$, and since $x_0^2 - d = nt$, we have

$$|t| \leq \frac{|x_0^2 - d|}{|n|} \leq \frac{|x_0|^2 + |d|}{|n|} \leq \frac{|d|^2}{4|n|} + \frac{|d|}{|n|} < \frac{|n|}{4} + \frac{|d|}{|n|} \leq \frac{|n|}{2},$$

where the last inequality is justified by checking each of the cases $|n| = 2$, $|n| = 3$, and $|n| \geq 4$, remembering that the integer $|d|$ is at least 1 and strictly smaller than $|n|$. It follows that $|t_1| \leq |t| < |n|$, which ensures that the algorithm will terminate, since either $|d|$ or $|n|$ is reduced in every recursive call; indeed, the number of recursive calls is clearly bounded by a logarithmic function of $\max(|d|, |n|)$.

To see that the solution returned in step 7 is correct, we first note that t_1 is square-free as required, and if SOLVE(d, t_1) succeeds then we may inductively assume that $x_1^2 - dy_1^2 = t_1 z_1^2$. Multiplying the LHS by $x_0^2 - d$ and the RHS by $x_0^2 - d = nt$ yields

$$\begin{aligned} (x_0^2 - d)(x_1^2 - dy_1^2) &= ntt_1 z_1^2 \\ x_0^2 x_1^2 - dx_0^2 y_1^2 - dx_1^2 + d^2 y_1^2 &= nt_1 t_2^2 t_1 z_1^2 \\ (x_0 x_1)^2 + (dy_1)^2 - d\left((x_0 y_1)^2 + x_1^2\right) &= n(t_1 t_2 z_1)^2 \\ (x_0 x_1 + dy_1)^2 - d(x_0 y_1 + x_1)^2 &= n(t_1 t_2 z_1)^2, \end{aligned}$$

which shows that $(x_0 x_1 + dy_1, x_0 y_1 + x_1, t_1 t_2 z_1)$ is indeed a solution to (3), as desired.

Computationally, the most expensive step of the algorithm (by far) is the computation of x_0 in step 6. As we will see in the next lecture, it is easy to compute square-roots modulo primes, but in general n may be composite, and the only known algorithm for computing square-roots modulo a square-free composite integer n is to compute square-roots modulo each of its prime factors and use the Chinese remainder theorem to get a square-root modulo n. This requires factoring the integer n, a problem for which no polynomial-time algorithm is known.

As described in [1], the algorithm $\textsc{Solve}(d, n)$ can be modified to avoid factorization in any of its recursive steps so that only one initial factorization is required. This does not yield a polynomial-time algorithm, but it greatly speeds up the process, and in practice it is now feasible to find rational solutions to $ax^2 + by^2 + cz^2 = 0$ even when the coefficients a, b, and c are as large as 10^{100}.

Another deficiency of the algorithm $\textsc{Solve}(d, n)$ is that the solutions it finds are typically much larger than necessary. There is a theorem due to Holzer that gives us an upper bound on the size of the smallest solution to (1), and hence of the smallest solution to (3).

Theorem 2.5 (Holzer)**.** *Let a, b, c be square-free integers that are pairwise coprime and suppose that the equation $ax^2 + by^2 + cz^2 = 0$ has a nonzero rational solution. Then there exists a nonzero integer solution (x_0, y_0, z_0) with*

$$|x_0| \le \sqrt{|bc|}, \qquad |y_0| \le \sqrt{|ac|}, \qquad |z_0| \le \sqrt{|ab|}.$$

Proof. See [2] for a short and elementary proof. □

On Problem Set 1 you will implement a simple improvement to algorithm $\textsc{Solve}(d, n)$ that significantly reduces the size of the solutions it finds (and reduces the number of recursive calls), and generally comes close to achieving the Holzer bounds.

Finally, we note that there is a simple criterion for determining whether or not a diagonal conic has a rational solution that does not require actually looking for one.

Theorem 2.6 (Legendre)**.** *Let a, b, c be square-free integers that are pairwise coprime and whose signs are not all the same. The equation $ax^2 + by^2 + cz^2 = 0$ has a rational solution if and only if the congruences*

$$X^2 \equiv -bc \bmod a, \qquad Y^2 \equiv -ca \bmod b, \qquad Z^2 \equiv -ab \bmod c$$

can be simultaneously satisfied.

The necessity of the condition given in Theorem 2.6 is easy to check; if we look at the equation modulo a, for example, we have $by^2 \equiv -cz^2 \bmod a$, and it follows that $-b/c$ and therefore $-bc$ must be a quadratic residue modulo a. The sufficiency can be proved by showing that if the condition holds than $\textsc{Solve}(d, n)$ will succeed in finding a solution to the corresponding norm equation $x^2 - dy^2 = nz^2$. This is basically how Legendre proved the theorem, but we will prove a more general statement after we have developed the theory of p-adic numbers.

It is worth noting that while the congruences in Legendre's theorem apparently give a very simple criterion for determining whether a conic has a rational point, in order to apply them we need to know the factorization of the integers a, b, c. This means that, in general, the problem of determining the existence of a rational solution is not significantly easier than actually finding one, and we still do not have a polynomial-time algorithm for determining the existence of a rational solution to a conic over $\mathbb{Q}$.

References

[1] J.E. Cremona and D. Rusin, *Efficient solution of rational conics*, Mathematics of Computation **72** (2003), 1417–1441.

[2] T. Cochrane and P. Mitchell, *Small solutions of the Legendre equation*, Journal of Number Theory **70** (1998), 62–66.

18.782 Introduction to Arithmetic Geometry **Fall 2013**
Lecture #3 **09/12/2013**

3.1 Quadratic reciprocity

Recall that for each odd prime p the Legendre symbol $(\frac{a}{p})$ is defined as

$$\left(\frac{a}{p}\right)=\begin{cases}1 & \text{if } a \text{ is a nonzero quadratic residue modulo } p,\\ 0 & \text{if } a \text{ is zero modulo } p,\\ -1 & \text{otherwise.}\end{cases}$$

The Legendre symbol is multiplicative, $(\frac{a}{p})(\frac{b}{p})=(\frac{ab}{p})$, and it can be computed using Euler's criterion:

$$\left(\frac{a}{p}\right)\equiv a^{\frac{p-1}{2}} \bmod p.$$

Both statements follow from the fact that the multiplicative group $(\mathbb{Z}/p\mathbb{Z})^\times$ is cyclic of order $p-1$ with -1 as the unique element of order 2 (the case $a=0$ is clear). We also have the well known law of quadratic reciprocity.

Theorem 3.1 (Gauss)**.** *For all odd primes p and q we have $(\frac{p}{q})(\frac{q}{p})=(-1)^{(\frac{p-1}{2})(\frac{q-1}{2})}$.*

I expect you have all seen proofs of this theorem, but I recently came across the following proof due to Rousseau [4], which Math Overflow overwhelmingly voted as the "best" proof quadratic reciprocity. The proof is quite short, so I thought I would share it with you.

Proof. Let $s=(p-1)/2$, $t=(q-1)/2$ and $u=(pq-1)/2$. Consider the three subsets of $(\mathbb{Z}/pq\mathbb{Z})^\times$ defined by

$$A=\{x : x \bmod p\in[1,s]\},\quad B=\{x : x \bmod q\in[1,t]\},\quad C=\{x : x \bmod pq\in[1,u]\}.$$

These subsets each contain exactly half of the $(p-1)(q-1)=4st$ elements of $(\mathbb{Z}/pq\mathbb{Z})^\times$ and thus have size $2st$. Furthermore, for all $x\in(\mathbb{Z}/pq\mathbb{Z})^\times$ each subset contains exactly one of x or $-x$. It follows that the products a,b,c over the sets A,B,C differ only in sign, so their ratios are all ±1. The intersection of A and B has size st, hence there are $2st-st=st$ sign differences between the elements of A and B, and therefore $a/b\equiv(-1)^{st} \bmod pq$. To complete the proof, we just need to show that $a/b\equiv(\frac{p}{q})(\frac{q}{p}) \bmod pq$, since two numbers that are both equal to ±1 and congruent mod $pq>2$ must be equal over $\mathbb{Z}$.

Considering the product a modulo q, it is clear that $a\equiv(q-1)!^s \bmod q$, since modulo q we are just multiplying s copies of the integers from 1 to $q-1$. To compute c modulo q we first compute the product of the integers in $[1,u]$ that are not divisible by q, which is $(q-1)!^s t!$, and then divide by the product of the integers in $[1,u]$ that are multiples of p, since these do not lie in $(\mathbb{Z}/pq\mathbb{Z})^\times$, which is $p\times2p\times\cdots tp=p^t t!$. Thus $c\equiv(q-1)!^s/p^t \bmod q$, and we have $a/c\equiv p^t\equiv\pm1 \bmod q$. But we know that $a/c\equiv\pm1 \bmod pq$, so this congruence also holds mod pq. By Euler's criterion, we have $a/c\equiv(\frac{p}{q}) \bmod pq$. Similarly, $b/c\equiv(\frac{q}{p}) \bmod pq$, and since $b/c\equiv\pm1 \bmod pq$, we have $c/b\equiv b/c \bmod pq$, and therefore $c/b\equiv(\frac{q}{p}) \bmod pq$. Thus $a/b=(a/c)(c/b)\equiv(\frac{p}{q})(\frac{q}{p}) \bmod pq$, as desired. □

Andrew V. Sutherland

3.2 Finite fields

We recall some standard facts about finite fields. For each prime power q there is, up to isomorphism, a unique field $\mathbb{F}_q$ with q elements (and it is easy to show that the order of every finite field is a prime power). We have the prime fields $\mathbb{F}_p \simeq \mathbb{Z}/p\mathbb{Z}$, and for any positive integer n the field $\mathbb{F}_{p^n}$ can be constructed as the splitting field of the (separable) polynomial $x^{p^n} - x$ over $\mathbb{F}_p$ (thus every finite field is a Galois extension of its prime field). More generally, every degree n extension of $\mathbb{F}_q$ is isomorphic to $\mathbb{F}_{q^n}$, the splitting field of $x^{q^n} - x$ over $\mathbb{F}_q$, and the Galois group $\operatorname{Gal}(\mathbb{F}_{q^n}/\mathbb{F}_q)$ is cyclic over order n, generated by the q-power Frobenius automorphism $x \mapsto x^q$. We have the inclusion $\mathbb{F}_{q^m} \subseteq \mathbb{F}_{q^n}$ if and only if m divides n: if $m|n$ then $x^{q^m} = x$ implies $x^{q^n} = x$, and if $\mathbb{F}_{q^m} \subseteq \mathbb{F}_{q^n}$ then $\mathbb{F}_{q^n}$ has dimension n/m as a vector space over $\mathbb{F}_{q^m}$.

While defining $\mathbb{F}_q = \mathbb{F}_{p^n}$ as a splitting field is conceptually simple, in practice we typically represent $\mathbb{F}_q$ more explicitly by adjoining the root of an irreducible polynomial $f \in \mathbb{F}_p[x]$ of degree n and define $\mathbb{F}_q$ as the ring quotient $\mathbb{F}_p[x]/(f)$. The ring $\mathbb{F}_p[x]$ is a principle ideal domain, so the prime ideal (f) is maximal and the quotient is therefore a field. Such an irreducible polynomial always exists: by the primitive element theorem we know that the separable extension $\mathbb{F}_q/\mathbb{F}_p$ can be constructed as $\mathbb{F}_p(\alpha)$ for some $\alpha \in \mathbb{F}_q$ whose minimal polynomial $f \in \mathbb{F}_p[x]$ is irreducible and of degree n. While no deterministic polynomial time algorithm is known for constructing f (even for $n = 2$ (!)), in practice the problem is readily solved using a randomized algorithm, as discussed below.

Elements s and t of $\mathbb{F}_q \simeq \mathbb{F}_p[x]/(f)$ correspond to polynomials in $\mathbb{F}_p[x]$ of degree at most n. The sum $s+t$ is computed as in $\mathbb{F}_p[x]$, and the product st is computed as a product in $\mathbb{F}_p[x]$ and then reduced modulo f, using Euclidean division and taking the remainder. To compute the inverse of s, one uses the (extended) Euclidean gcd algorithm to compute polynomials $u, v \in \mathbb{F}_p[x]$ that satisfy

$$us + vf = \gcd(s, f) = 1,$$

and u is then the inverse of s modulo f; note that $\gcd(s, f) = 1$ since f is irreducible. Using fast algorithms for polynomial arithmetic, all of the field operations in $\mathbb{F}_q$ can be computed in time that is quasi-linear in $\log q = n \log p$, which is also the amount of space needed to represent an element of $\mathbb{F}_q$ (up to a constant factor).

Example 3.2. $\mathbb{F}_8 \simeq \mathbb{F}_2[t]/(t^3 + t + 1) = \{0, 1, t, t+1, t^2, t^2+1, t^2+t, t^2+t+1\}$ is a finite field of order 8 in which, for example, $(t^2+1)(t^2+t) = t+1$. Note that $\mathbb{F}_2 = \{0, 1\}$ is its only proper subfield (in particular, $\mathbb{F}_4 \not\subseteq \mathbb{F}_8$).

The most thing we need to know about finite fields is that their multiplicative groups are cyclic. This is an immediate consequence of a more general fact.

Theorem 3.3. *Any finite subgroup G of the multiplicative group of a field k is cyclic.*

Proof. The group G must be abelian, so by the structure theorem for finite abelian groups it is isomorphic to a product of cyclic groups

$$G \simeq \mathbb{Z}/n_1\mathbb{Z} \times \mathbb{Z}/n_2\mathbb{Z} \times \cdots \times \mathbb{Z}/n_k\mathbb{Z},$$

where each $n_i > 1$ and we may assume that $n_i|n_{i+1}$. If G is not cyclic, then $k \geq 2$ and G contains a subgroup isomorphic to $\mathbb{Z}/n_1\mathbb{Z} \times \mathbb{Z}/n_1\mathbb{Z}$ and therefore contains at least $n_1^2 > n_1$ elements whose orders divide n_1. But the polynomial $x^{n_1} - 1$ has at most n_1 roots in k, so this is not possible and G must be cyclic. □

3.3 Rational points on conics over finite fields

We now turn to the problem of finding rational points on conics over finite fields. We begin by proving that, unlike the situation over $\mathbb{Q}$, there is always a rational point to find.

Theorem 3.4. *Let $C/\mathbb{F}_q$ be a conic over a finite field of odd characteristic. Then C has a rational point.*

Proof. As shown in Lecture 2, by completing the square we can put C in the form $ax^2 + by^2 + cz^2 = 0$. If any of a, b, c is zero, say c, then $(0 : 0 : 1)$ is a rational point on C, so we now assume otherwise. The group $\mathbb{F}_q^\times$ is cyclic and has even order $q-1$, so it contains exactly $\frac{q-1}{2}$ squares. Therefore the set $S = \{y^2 : y \in \mathbb{F}_q\}$ has cardinality $\frac{q+1}{2}$ (since it also includes 0), as does the set $T = \{-by^2 - c : y \in \mathbb{F}_q\}$, since it is a linear transformation of S. Similarly, the set $U = \{ax^2 : x \in \mathbb{F}_q\}$ has cardinality $\frac{q+1}{2}$. The sets T and U cannot be disjoint, since the sum of their cardinalities is larger than $\mathbb{F}_q$, so we must have some $-by_0^2 - c \in T$ equal to some $ax_0^2 \in U$, and $(x_0 : y_0 : 1)$ is then a rational point on C. □

Corollary 3.5. *Let $C/\mathbb{F}_q$ be a conic over a finite field. Then one of the following holds*

1. *C is geometrically irreducible, isomorphic to $\mathbb{P}^1$, and has $q+1$ rational points.*
2. *C is reducible over $\mathbb{F}_q$, isomorphic to the union of two rational projective lines, and has $2q+1$ rational points.*
3. *C is reducible over $\mathbb{F}_{q^2}$, but not over $\mathbb{F}_q$, isomorphic over $\mathbb{F}_{q^2}$ to the union of two projective lines with a single rational point at their intersection.*

In every case we have $\#C(\mathbb{F}_q) \equiv 1 \bmod q$.

Proof. If C is geometrically irreducible then we are in case 1 and the conclusion follows from Theorem 2.3, since we know by Theorem 3.4 that C has a rational point. Otherwise, C must be the product of two degree 1 curves (projective lines), which must intersect at at a single point. If the lines can be defined over $\mathbb{F}_q$ then we are in case 2 and have $2(q+1)-1 = 2q+1$ projective points and otherwise the lines must be defined over the quadratic extension $\mathbb{F}_{q^2}$. which is case 3. The non-trivial element of the Galois group $\mathrm{Gal}(\mathbb{F}_{q^2}/\mathbb{F}_q)$ swaps the two lines and must fix their intersection, which consequently lies in $\mathbb{F}_q$. □

Remark 3.6. Theorem 3.4 and Corollary 3.5 also hold in characteristic 2.

3.4 Root finding

Let f be a univariate polynomial over a finite field $\mathbb{F}_q$. We now consider the problem of how to find the roots of f that lie in $\mathbb{F}_q$. This will allow us, in particular, to compute the square root of an element $a \in \mathbb{F}_q$ by taking $f(x) = x^2 - a$, which is a necessary ingredient for finding rational points on conics over $\mathbb{F}_q$, and also over $\mathbb{Q}$. Recall that the critical step of the descent algorithm we saw in Lecture 2 for finding a rational point on a conic over $\mathbb{Q}$ required us to compute square roots modulo a square-free integer n; this is achieved by computing square roots modulo each of the prime factors of n and applying the Chinese remainder theorem (of course this requires us to compute the prime factorization of n, which is actually the hard part).

No deterministic polynomial-time algorithm is know for root-finding over finite fields. Indeed, even the special case of computing square roots modulo a prime is not known to have a deterministic polynomial-time solution.[1] But if we are prepared to use randomized algorithms (which we are), we can quite solve this problem quite efficiently. The algorithm we give here was originally proposed by Berlekamp for prime fields [1], and then refined and extended by Rabin [3], whose presentation we follow here. This algorithm is a great example of how randomness can be exploited in a number-theoretic setting. As we will see, it is quite efficient, with an expected running time that is quasi-quadratic in the size of the input.

3.4.1 Randomized algorithms

Randomized algorithms are typically classified as one of two types: *Monte Carlo* or *Las Vegas*. Monte Carlo algorithms are randomized algorithms whose output may be incorrect, depending on random choices made by the algorithm, but whose running time is bounded by a function of its input size, independent of any random choices. The probability of error is required to be less than $1/2 - \epsilon$, for some $\epsilon > 0$, and can be made arbitrarily small be running the algorithm repeatedly and using the output that occurs most often. In contrast, a Las Vegas algorithm always produces a correct output, but its running time may depend on random choices made by the algorithm and need not be bounded as a function of the input size (but we do require its expected running time to be finite). As a trivial example, consider an algorithm to compute $a + b$ that first flips a coin repeatedly until it gets a head and then computes $a + b$ and outputs the result. The running time of this algorithm may be arbitrarily long, even when computing $1 + 1 = 2$, but its *expected* running time is $O(n)$, where n is the size of the inputs (typically measured in bits).

Las Vegas algorithms are generally preferred, particularly in mathematical applications, where we generally require provably correct results. Note that any Monte Carlo algorithm whose output can be verified can always be converted to a Las Vegas algorithm (just run the algorithm repeatedly until you get an answer that is verifiably correct). The root-finding algorithm we present here is of the Las Vegas type.

3.4.2 Factoring with gcds

The roots of our polynomial $f \in \mathbb{F}_q[x]$ all lie in the algebraic closure $\overline{\mathbb{F}}_q$. The roots that actually lie in $\mathbb{F}_q$ are distinguished by the fact that they are fixed by the Frobenius automorphism $x \mapsto x^q$. It follows that the roots of f that lie in $\mathbb{F}_q$ are precisely those that are also roots of the polynomial $x^q - x$. Thus the polynomial

$$g = \gcd(f, x^q - x)$$

has the form $\prod_i (x - \alpha_i)$, where the α_i range over the distinct roots of f that lie in $\mathbb{F}_q$. If f has no roots in $\mathbb{F}_q$ then g will have degree 0, and otherwise we can reduce the problem of finding a root of f to the problem of finding a root of g, a polynomial whose roots are distinct and known to lie in $\mathbb{F}_q$. Note that this already gives us a deterministic algorithm to determine whether or not f actually has any roots in $\mathbb{F}_q$, but in order to actually find one we may need to factor g, and this is where we will use a randomized approach.

In order to compute $\gcd(f, x^q - x)$ efficiently, one does *not* compute $x^q - x$ and then take the gcd with f; this would take time exponential in $\log q$, whereas we want an algorithm whose running time is polynomial in the *size* of f, which is proportional to $\deg f \log q$.

[1] If one assumes the extended Riemann Hypothesis, this and many other special cases of the root-finding problem can be solved in polynomial time.

Instead, one computes $x^q \bmod f$ by exponentiating the polynomial x in the ring $\mathbb{F}_q[x]/(f)$, whose elements are uniquely represented by polynomials of degree less than $d = \deg f$. Each multiplication in this ring involves the computation of a product in $\mathbb{F}_q[x]$ followed by a reduction modulo f. This reduction is achieved using Euclidean division, and can be accomplished within a constant factor of the time required by the multiplication. The computation of x^q is achieved using binary exponentiation (or some other efficient method of exponentiation), where one performs a sequence of squarings and multiplications by x based on the binary representation of q, and requires just $O(\log q)$ multiplications in $\mathbb{F}_q[x](f)$. Once we have computed $x^q \bmod f$, we subtract x and compute $g = \gcd(f, x^q - x)$.

Assuming that q is odd (which we do), we may factor the polynomial $x^q - x$ as

$$x^q - x = x(x^s - 1)(x^s + 1),$$

where $s = (q-1)/2$. Ignoring the root 0 (which we can easily check separately), this factorization splits $\mathbb{F}_q^\times$ precisely in half: the roots of $x^s - 1$ are the elements of $\mathbb{F}_q^\times$ that are quadratic residues, and the roots of $x^s + 1$ are the elements of $\mathbb{F}_q^\times$ that are not. If we compute

$$h = \gcd(g, x^s - 1),$$

we obtain a divisor of g whose roots are precisely the roots of g that are quadratic residues. If we suppose that the roots of g are as likely as not to be quadratic residues, we should expect the degree of h to be approximately half the degree of g, and so long as the degree of h is strictly between 0 and $\deg g$, one of h or g/h is a polynomial of degree at most half the degree of g and whose roots are all roots of our original polynomial f.

To make further progress, and to obtain an algorithm that is guaranteed to work no matter how the roots of g are distributed in $\mathbb{F}_q$, we take a randomized approach. Rather than using the fixed polynomial $x^s - 1$, we consider random polynomials of the form

$$(x + \delta)^s - 1,$$

where δ is uniformly distributed over $\mathbb{F}_q$. We claim that if α and β are any two nonzero roots of g, then with probability 1/2, exactly one of these is a root $(x+\delta)^s - 1$. It follows from this claim that so long as g has at least 2 distinct nonzero roots, the probability that the polynomial $h = \gcd(g, (x+\delta)^s + 1)$ is a proper divisor of g is at least 1/2.

Let us say that two elements $\alpha, \beta \in \mathbb{F}_q$ are of *different type* if they are both nonzero and $\alpha^s \neq \beta^s$. Our claim is an immediate consequence of the following theorem from [3].

Theorem 3.7 (Rabin). *For every pair of distinct $\alpha, \beta \in \mathbb{F}_q$ we have*

$$\#\{\delta \in \mathbb{F}_q : \alpha + \delta \textit{ and } \beta + \delta \textit{ are of different type}\} = \frac{q-1}{2}.$$

Proof. Consider the map $\phi(\delta) = \frac{\alpha+\delta}{\beta+\delta}$, defined for $\delta \neq -\beta$. We claim that ϕ is a bijection form the set $\mathbb{F}_q \backslash \{-\beta\}$ to the set $\mathbb{F}_q \backslash \{1\}$. The sets are the same size, so we just need to show surjectivity. Let $\gamma \in \mathbb{F}_q - \{1\}$, then we wish to find a solution $x \neq -\beta$ to $\gamma = \frac{\alpha+x}{\beta+x}$. We have $\gamma(\beta + x) = \alpha + x$ which means $x - \gamma x = \gamma\beta - \alpha$. This yields $x = \frac{\gamma\beta - \alpha}{1-\gamma}$, which is not equal to $-\beta$, since $\alpha \neq \beta$. Thus ϕ is surjective.

We now note that

$$\phi(\delta)^s = \frac{(\alpha+\delta)^s}{(\beta+\delta)^s}$$

is -1 if and only if $\alpha + \delta$ and $\beta + \delta$ are of different type. The elements $\gamma = \phi(\delta)$ for which $\gamma^s = -1$ are precisely the non-residues in $\mathbb{F}_q \backslash \{1\}$, of which there are exactly $(q-1)/2$. □

We now give the algorithm.

```
Algorithm FindRoot($f$)
Input: A polynomial $f \in \mathbb{F}_q[x]$.
Output: An element $r \in \mathbb{F}_q$ such that $f(r) = 0$, or null if no such $r$ exists.

  1. If $f(0) = 0$ then return 0.
  2. Compute $g = \gcd(f, x^q - x)$.
  3. If $\deg g = 0$ then return null.
  4. While $\deg g > 1$:
       a. Pick a random $\delta \in \mathbb{F}_q$.
       b. Compute $h = \gcd(g, (x + \delta)^s - 1)$.
       c. If $0 < \deg h < \deg g$ then replace $g$ by $h$ or $g/h$, whichever has lower degree.
  5. Return $r = -b/a$, where $g(x) = ax + b$.
```

It is clear that the output of the algorithm is always correct, since every root of the polynomial g computed in step 2 is a root of f, and when g is updated in step 4c it is always replaced by a proper divisor. We now consider its complexity.

It follows from Theorem 3.7 that the polynomial h computed in step 4b is a proper divisor of g with probability at least 1/2, since g has at least two distinct nonzero roots $\alpha, \beta \in \mathbb{F}_q$. Thus the expected number of iterations needed to obtain a proper factor h of g is bounded by 2. The degree of h is at most half the degree of g, and the total cost of computing all the polynomials h during step 4 is actually within a constant factor the cost of computing g in step 2.

Using fast algorithms for multiplications and the gcd computation, the time to compute g can be bounded by

$$O(\mathsf{M}(d \log q)(\log q + \log d)$$

bit operations, where $\mathsf{M}(b)$ denotes the time to multiply to b-bit integers and is asymptotically bounded by $\mathsf{M}(b) = O(b \log b \log\log b)$ (in fact one can do slightly better). The details of this complexity analysis and the efficient implementation of finite field arithmetic will not concern us in this course, we refer the reader to [2] for a comprehensive treatment, or see these notes for a brief overview. The key point is that this time complexity is polynomial in $d \log q$, in fact it is essentially quadratic, and in practice we can quite quickly find roots of polynomials even over very large finite fields. same complexity bound, and the total expected running time is $O(\mathsf{M}(nd)(n + \log d))$.

The algorithm can easily be modified to find all the distinct roots of f, by modifying step 4c to recursively find the roots of both h and g/h, this only increases the running time by a factor of $O(\log d)$. Assuming that d is less than the charcteristic of $\mathbb{F}_q$, one can easily determine the multiplicity of each root of f: a root α of f occurs with multiplicity k if and only if α is a root of $f^{(k)}$ but not a root of $f^{(k+1)}$, where $f^{(k)}$ denotes the kth derivative of f. The time to perform this computation is negligible compared to the time to find the distinct roots.

3.5 Finding rational points on curves over finite fields

Now that we know how to find roots of univariate polynomials in finite fields (and in particular, square roots), we can easily find a rational point on any conic over a finite field (and enumerate all the rational points if we wish). As above, let us assume $\mathbb{F}_q$ has odd characteristic, so we can put our conic C is diagonal form $x^2 + by^2 + cz^2 = 0$. If C is geometrically reducible then, as proved on Problem Set 1, it is singular and one of a, b, c must be 0. So one of $(1 : 0 : 0)$, $(0 : 1 : 0)$, $(0 : 0 : 1)$ is a rational point on the curve, and in the case that C is reducible over $\mathbb{F}_q$ we can determine the equations of the two lines whose union forms C by computing square roots in $\mathbb{F}_q$; for example, if $c = 0$ we can compute $ax^2 + by^2 = (\sqrt{a}x + \sqrt{-b}y)(\sqrt{a}x + \sqrt{-b}y)$. It is then straight-forward to enumerate all the rational points on C.

Now let us suppose that C is geometrically irreducible, in which case we must have $abc \neq 0$. If any of $-a/b, -b/c, -c/a$ is a square in $\mathbb{F}_q$, then we can find a rational point with one coordinate equal to 0 by computing a square-root. Otherwise we know that every rational point $(x_0, y_0, z_0) \in C(\mathbb{F}_q)$ satisfies $x_0y_0z_0 \neq 0$, so we can assume $z_0 = 1$. For each of the $q - 1$ possible nonzero choices for y_0, we get either 0 or 2 rational points on C, depending on whether $-(by_0^2 + c)/a$ is a square or not. By Corollary .refcor:ffconicpts, We know there are a total of $q + 1$ rational points, so for exactly $(q + 1)/2$ values of y_0 we must have $-(by_0^2 + c)/a$ square. Thus if we pick $y_0 \in \mathbb{F}_q$ at random, we have a better than 50/50 chance of finding a rational point on C by computing $\sqrt{-(by_0^2 + c)/a}$. This gives us a Las Vegas algorithm for finding a rational point on C whose expected running time is within a constant factor of the time to compute a square-root in $\mathbb{F}_q$, which is quasi-quadratic in $\log q$. Once we have a rational point on our irreducible conic C, we can enumerate them all using the parameterization we computed in Lecture 2.

Remark 3.8. The argument above applies more generally. Suppose we have a geometrically irreducible plane curve C defined by a homogeneous polynomial $f(x, y, z)$ of some fixed degree d It follows from the Hasse-Weil bounds, which we will see later in course, that $\#C(\mathbb{F}_q) = q + O(\sqrt{q})$. Assuming $q \gg d$, if we pick a random projective pair $(y_0 : z_0)$ and then attempt to find a root x_0 of the univariate polynomial $g(x) = f(x, y_0, z_0)$, we will succeed with a probability that asymptotically approaches $1/d$ as $q \to \infty$. This yields a Las Vegas algorithm for finding a rational point on C in time quasi-quadratic in $\log q$.

References

[1] Elwyn R. Berlekamp, *Factoring polynomials over large finite fields*, Mathematics of Computation **24** (1970), 713–735.

[2] Joachim von zur Gathen and Jürgen Garhard, *Modern Computer Algebra*, third edition, Cambridge University Press, 2013.

[3] Michael O. Rabin, *Probabilistic algorithms in finite fields*, SIAM Journal of Computing **9** (1980), 273–280.

[4] G. Rousseau, *On the quadratic reciprocity law*, Journal of the Australian Mathematical Society (Series A) **51** (1991), 423–425.

18.782 Introduction to Arithmetic Geometry **Fall 2013**
Lecture #4 **09/17/2013**

4.1 Inverse limits

Definition 4.1. An *inverse system* is a sequence of objects (e.g. sets/groups/rings) (A_n) together with a sequence of morphisms (e.g. functions/homomorphisms) (f_n)

$$\cdots \longrightarrow A_{n+1} \xrightarrow{f_n} A_n \longrightarrow \cdots \longrightarrow A_2 \xrightarrow{f_1} A_1.$$

The *inverse limit*

$$A = \varprojlim A_n$$

is the subset of the direct product $\prod_n A_n$ consisting of those sequences $a = (a_n)$ for which $f_n(a_{n+1}) = a_n$ for all $n \ge 1$. For each $n \ge 1$ the *projection map* $\pi_n \colon A \to A_n$ sends a to a_n.

Remark 4.2. For those familiar with category theory, one can define inverse limits for any category. In most cases the result will be another object of the same category (unique up to isomorphism), in which case the projection maps are then morphisms in that category. We will restrict our attention to the familiar categories of sets, groups, and rings. One can also generalize the index set $\{n\}$ from the positive integers to any partially ordered set.

4.2 The ring of p-adic integers

Definition 4.3. For a prime p, the *ring of p-adic integers* $\mathbb{Z}_p$ is the inverse limit

$$\mathbb{Z}_p = \varprojlim \mathbb{Z}/p^n\mathbb{Z}$$

of the inverse system of rings $(\mathbb{Z}/p^n\mathbb{Z})$ with morphisms (f_n) given by reduction modulo p^n (for a residue class $\overline{x} \in \mathbb{Z}/p^{n+1}\mathbb{Z}$, pick an integer $x \in \overline{x}$ and take its residue class in $\mathbb{Z}/p^n\mathbb{Z}$).

The multiplicative identity in $\mathbb{Z}_p$ is $1 = (\bar{1}, \bar{1}, \bar{1}, \ldots)$, where the nth $\bar{1}$ denotes the residue class of 1 in $\mathbb{Z}/p^n\mathbb{Z}$. The map that sends each integer $x \in \mathbb{Z}$ to the sequence $(\bar{x}, \bar{x}, \bar{x}, \ldots)$ is a ring homomorphism, and its kernel is clearly trivial, since 0 is the only integer congruent to $0 \bmod p^n$ for all n. Thus the ring $\mathbb{Z}_p$ has characteristic 0 and contains $\mathbb{Z}$ as a subring. But $\mathbb{Z}_p$ is a much bigger ring than $\mathbb{Z}$.

Example 4.4. If we represent elements of $\mathbb{Z}/p^n\mathbb{Z}$ by integers in $[0, p^n - 1]$, in $\mathbb{Z}_7$ we have

$$\begin{aligned}
2 &= (2, 2, 2, 2, 2, \ldots)\\
2002 &= (0, 42, 287, 2002, 2002, \ldots)\\
-2 &= (5, 47, 341, 2399, 16805, \ldots)\\
2^{-1} &= (4, 25, 172, 1201, 8404, \ldots)\\
\sqrt{2} &= \begin{cases} (3, 10, 108, 2166, 4567\ldots)\\ (4, 39, 235, 235, 12240\ldots) \end{cases}\\
\sqrt[5]{2} &= (4, 46, 95, 1124, 15530, \ldots)
\end{aligned}$$

Note that 2002 is not invertible in $\mathbb{Z}_7$, and that while 2 has two square roots in $\mathbb{Z}_7$, it has only one fifth root, and no cube roots.

Andrew V. Sutherland

While representing elements of $\mathbb{Z}_p$ as sequences (a_n) with $a_n \in \mathbb{Z}/p^n\mathbb{Z}$ follows naturally from the definition of $\mathbb{Z}_p$ as an inverse limit, it is redundant. The value of a_n constrains the value of a_{n+1} to just p of the p^{n+1} elements of $\mathbb{Z}/p^{n+1}\mathbb{Z}$, namely, those that are congruent to a_n modulo p^n. If we uniquely represent each a_n as an integer in the interval $[0, p^n - 1]$ we can always write $a_{n+1} = a_n + p^n b_n$ with $b_n \in [0, p-1]$.

Definition 4.5. Let $a = (a_n)$ be a p-adic integer with each a_n uniquely represented by an integer in $\in [0, p^n - 1]$. The sequence $(b_0, b_1, b_2, \ldots)$ with $b_0 = a_1$ and $b_n = (a_{n+1} - a_n)/p^n$ is called the *p-adic expansion of a.*

Theorem 4.6. *Every element of $\mathbb{Z}_p$ has a unique p-adic expansion and every sequence $(b_0, b_1, b_2, \ldots)$ of integers in $[0, p-1]$ is the p-adic expansion of an element of $\mathbb{Z}_p$.*

Proof. This follows immediately from the definition: we can recover (a_n) from its p-adic expansion $(b_0, b_1, b_2, \ldots)$ via $a_1 = a_0$ and $a_{n+1} = a_n + pb_n$ for all $n \geq 1$. □

Thus we have a bijection between $\mathbb{Z}_p$ and the set of all sequences of integers in $[0, p-1]$ indexed by the nonnegative integers.

Example 4.7. We have the following p-adic expansion in $\mathbb{Z}_7$:

$$\begin{aligned}
2 &= (2,0,0,0,0,0,0,0,0,0,\ldots)\\
2002 &= (0,6,5,5,0,0,0,0,0,0,\ldots)\\
-2 &= (5,6,6,6,6,6,6,6,6,6,\ldots)\\
2^{-1} &= (4,3,3,3,3,3,3,3,3,3,\ldots)\\
5^{-1} &= (3,1,4,5,2,1,4,5,2,1,\ldots)\\
\sqrt{2} &= \begin{cases} (3,1,2,6,1,2,1,2,4,6\ldots) \\ (4,5,4,0,5,4,5,4,2,0\ldots) \end{cases}\\
\sqrt[5]{2} &= (4,6,1,3,6,4,3,5,4,6\ldots)
\end{aligned}$$

You can easily recreate these examples (and many more) in Sage. To create the ring of 7-adic integers, just type `Zp(7)`. By default Sage will use 20 digits of p-adic precision, but you can change this to n digits using `Zp(p,n)`.

Performing arithmetic in $\mathbb{Z}_p$ using p-adic expansions is straight-forward. One computes a sum of p-adic expansions $(b_0, b_1, \ldots) + (c_0, c_1, \ldots)$ by adding digits mod p and carrying to the right (don't forget to carry!). Multiplication corresponds to computing products of formal power series in p, e.g. $\left(\sum b_n p^n\right)\left(\sum c_n p^n\right)$, and can be performed by hand using the standard schoolbook algorithm for multiplying integers represented in base 10, except now one works in base p. But Sage will do happily do all this arithmetic for you; I encourage you to experiment in Sage in order to build your intuition.

4.3 Properties of $\mathbb{Z}_p$

Recall that a sequence of group homomorphisms is *exact* if, for each intermediate group in the sequence, the image of the incoming homomorphism is equal to the kernel of the outgoing homomorphism. In the case of a *short exact sequence*

$$0 \longrightarrow A \xrightarrow{f} B \xrightarrow{g} C \longrightarrow 0,$$

this simply means that f is injective, g is surjective, and $\operatorname{im} f = \ker g$. In this situation the the homomorphism g induces an isomorphism $B/f(A) \simeq C$.

Theorem 4.8. *For each positive integer m, the sequence*

$$0 \longrightarrow \mathbb{Z}_p \xrightarrow{[p^m]} \mathbb{Z}_p \xrightarrow{\pi_m} \mathbb{Z}/p^m\mathbb{Z} \longrightarrow 0,$$

is exact. Here $[p^m]$ is the multiplication-by-p^m map and π_m is the projection to $\mathbb{Z}/p^m\mathbb{Z}$.

Proof. The map $[p^m]$ shifts the p-adic expansion $(b_0, b_1, \ldots)$ of each element in $\mathbb{Z}_p$ to the right by m digits (filling with zeroes) yielding

$$(c_0, c_1, c_2, \ldots) = (0, \ldots, 0, b_0, b_1, b_2, \ldots),$$

with $c_n = 0$ for $n < m$ and $c_n = b_{n-m}$ for all $n \geq m$. This is clearly an injective operation on p-adic expansions, and hence on $\mathbb{Z}_p$, and the image of $[p^m]$ consists of the elements in $\mathbb{Z}_p$ whose p-adic expansion $(c_0, c_1, c_2, \ldots)$ satisfies $c_0 = \cdots = c_{m-1} = 0$.

Conversely, the map π_m sends the p-adic expansion $(b_0, b_1, b_2, \ldots,)$ to the sum

$$b_0 + b_1 p + b_2 p^2 + \cdots b_{m-1} p^{m-1}$$

in $\mathbb{Z}/p^m\mathbb{Z}$. Each element of $\mathbb{Z}/p^m\mathbb{Z}$ is uniquely represented by an integer in $[0, p^m - 1]$, each of which can be (uniquely) represented by a sum as above, with $b_0, \ldots, b_{m-1}$ integers in $[0, p-1]$. It follows that π_m is surjective, and its kernel consists of the elements in $\mathbb{Z}_p$ whose p-adic expansion $(b_0, b_1, b_2, \ldots)$ satisfies $b_0 = \cdots = b_{m-1} = 0$, which is precisely $\operatorname{im}[p^m]$. □

Corollary 4.9. *For all positive integers m we have $\mathbb{Z}_p/p^m\mathbb{Z}_p \simeq \mathbb{Z}/p^m\mathbb{Z}$.*

Definition 4.10. For each nonzero $a \in \mathbb{Z}_p$ the *p-adic valuation* of a, denoted $v_p(a)$, is the greatest integer m for which a lies in the image of $[p^m]$; equivalently, $v_p(a)$ is the index of the first nonzero entry in the p-adic expansion of a. We also define $v_p(0) = \infty$, and adopt the conventions that $n < \infty$ and $n + \infty = \infty$ for any integer n.

Theorem 4.11. *The p-adic valuation v_p satisfies the following properties:*

(1) $v_p(a) = \infty$ *if and only if* $a = 0$.

(2) $v_p(ab) = v_p(a) + v_p(b)$.

(3) $v_p(a + b) \geq \min(v_p(a), v_p(b))$.

Proof. The first property is immediate from the definition. The second two are clear when either a or b is zero, so we assume otherwise and let $m = v_p(a)$ and $n = v_p(b)$.

For (2) we have $a = p^m a'$ and $b = p^n b'$, for some $a', b' \in \mathbb{Z}_p$, and therefore $ab = p^{m+n}a'b'$ lies in $\operatorname{im}[p^{m+n}]$ and $v_p(ab) \geq m + n$. On the other hand, the coefficient of p^m in the p-adic expansion of a and the coefficient of p^n in the p-adic expansion of b are both nonzero, so the coefficient of p^{m+n} in the p-adic expansion of ab is nonzero, thus $v_p(ab) \leq m + n$.

For (3) we assume without loss of generality that $m \leq n$, in which case $\operatorname{im}[p^n] \subseteq \operatorname{im}[p^m]$, so a and b both lie in $\operatorname{im}[p^m]$, as does $a+b$, and we have $v_p(a+b) \geq m = \min(v_p(a), v_p(b))$. □

The p-adic valuation v_p is an example of a *discrete valuation.*

Definition 4.12. Let R be a commutative ring. A *discrete valuation* (on R) is a function $v \colon R \to \mathbb{Z} \bigcup \{\infty\}$ that satisfies the three properties listed in Theorem 4.11.

Corollary 4.13. *$\mathbb{Z}_p$ is an integral domain (a ring with no zero divisors).*

Proof. If a and b are both nonzero then $v_p(ab) = v_p(a) + v_p(b) < \infty$, so $ab \neq 0$. □

Definition 4.14. The group of *p-adic units* $\mathbb{Z}_p^\times$ is the multiplicative group of invertible elements in $\mathbb{Z}_p$.

Theorem 4.15. *The following hold:*

(1) $\mathbb{Z}_p^\times = \mathbb{Z}_p - p\mathbb{Z}_p$; equivalently, $\mathbb{Z}_p^\times = \{a \in \mathbb{Z}_p : v_p(a) = 0\}$.

(2) Every nonzero $a \in \mathbb{Z}_p$ can be uniquely written as $p^n u$ with $n \in \mathbb{Z}_{\geq 0}$ and $u \in \mathbb{Z}_p^\times$.

Proof. We first note $v_p(p^m) = m$ for all $m \geq 0$, and in particular, $v_p(1) = 0$. If $a \in \mathbb{Z}_p^\times$, then a has a multiplicative inverse a^{-1} and we have $v_p(a) + v_p(a^{-1}) = v_p(1) = 0$, which implies that $v_p(a) = v_p(a^{-1}) = 0$, since $v_p(a)$ is nonnegative for all $a \in \mathbb{Z}_p$. Conversely, if $a = (a_n)$ with each $a_n \in \mathbb{Z}/p^n\mathbb{Z}$ and $v_p(a) = 0$, then $a_1 \not\equiv 0 \bmod p$ is invertible in $\mathbb{Z}/p\mathbb{Z}$, and since $a_n \equiv a_1 \not\equiv 0 \bmod p$, each a_n is invertible in $\mathbb{Z}/p^n\mathbb{Z}$. So $a^{-1} = (a_n^{-1}) \in \mathbb{Z}_p$, which proves (1).

For (2), if $a \in \mathbb{Z}_p$ is nonzero, let $v_p(a) = m$. Then $a \in \operatorname{im}[p^m]$ and therefore $a = p^m u$ for some $u \in \mathbb{Z}_p$. We then have

$$m = v_p(a) = v_p(p^m u) = v_p(p^m) + v_p(u) = m + v_p(u),$$

so $v_p(u) = 0$, and therefore $u \in \mathbb{Z}_p^\times$. □

Theorem 4.16. *Every nonzero ideal in $\mathbb{Z}_p$ is of the form (p^m) for some integer $m \geq 0$.*

Proof. Let I be a nonzero ideal in $\mathbb{Z}_p$, and let $m = \inf\{v_p(a) : a \in I\}$. Then $m < \infty$ (since I is nonzero), and every $a \in I$ lies in $\operatorname{im}[p^m] = (p^m)$. On the other hand, $v_p(a) = m$ for some $a \in I$ (since v_p is discrete), and we can write $a = p^m u$ for some unit u. But then $u^{-1}a = p^m \in I$ (since I is closed under multiplication by elements of R), thus $p^m \in I \subseteq (p^m)$ which implies $I = (p^m)$. □

Corollary 4.17. *The ring $\mathbb{Z}_p$ is a principal ideal domain with a unique maximal ideal.*

Definition 4.18. A *discrete valuation ring* is a principal ideal domain which contains a unique maximal ideal and is not a field.

This definition of a discrete valuation ring might seem strange at first glance, since it doesn't mention a valuation. But given a discrete valuation ring R with maximal ideal (p), where p is any irreducible element of R, we can define $v\colon R \to \mathbb{Z} \bigcup\{\infty\}$ by setting $v(0) = \infty$ and for every nonzero $a \in R$ defining $v(a)$ as the greatest positive integer n for which $a \in (p^n)$. It is then easy to check that v is a discrete valuation on R.

Discrete valuation rings are about as close as a commutative ring can get to being a field without actually becoming one. To turn a discrete valuation ring into a field, we only need to invert one element (any generator for its maximal ideal). Another remarkable fact about discrete valuation rings is that (up to units) they are unique factorization domains with exactly one prime!

18.782 Introduction to Arithmetic Geometry **Fall 2013**
Lecture #5 **09/19/2013**

5.1 The field of p-adic numbers

Definition 5.1. The field of *p-adic numbers* $\mathbb{Q}_p$ is the fraction field of $\mathbb{Z}_p$.

As a fraction field, the elements of $\mathbb{Q}_p$ are by definition all pairs $(a, b) \in \mathbb{Z}_p^2$, typically written as a/b, modulo the equivalence relation $a/b \sim c/d$ whenever $ad = bc$. But we can represent elements of $\mathbb{Q}_p$ more explicitly by extending our notion of a p-adic expansion to allow negative indices, with the proviso that only finitely many p-adic digits with negative indices are nonzero. If we view p-adic expansions in $\mathbb{Z}_p$ as formal power series in p, in $\mathbb{Q}_p$ we now have formal Laurent series in p.

Recall that every element of $\mathbb{Z}_p$ can be written in the form up^n, with $n \in \mathbb{Z}_{\geq 0}$ and $u \in \mathbb{Z}_p^\times$, and it follows that the elements of $\mathbb{Q}_p$ can all be written in the form up^n with $n \in \mathbb{Z}$ and $u \in \mathbb{Z}_p^\times$. If $(b_0, b_1, b_2, \ldots)$ is the p-adic expansion of $u \in \mathbb{Z}_p^\times$, then the p-adic expansion of $p^n u$ is $(c_n, c_{n+1}, c_{n+2}, \ldots)$ with $c_{n+i} = b_i$ for all $i \geq 0$ and $c_{n-i} = 0$ for all $i < 0$ (this works for both positive and negative n).

We extend the p-adic valuation v_p to $\mathbb{Q}_p$ by defining $v_p(p^n) = n$ for any integer n; as with p-adic integers, the valuation of any p-adic number is just the index of the first nonzero digit in its p-adic expansion. We can then distinguish $\mathbb{Z}_p$ as the subset of $\mathbb{Q}_p$ with nonnegative valuations, and $\mathbb{Z}_p^\times$ as the subset with zero valuation. We have $\mathbb{Q} \subset \mathbb{Q}_p$, since $\mathbb{Z} \subset \mathbb{Z}_p$, and for any $x \in \mathbb{Q}_p$, either $x \in \mathbb{Z}_p$ or $x^{-1} \in \mathbb{Z}_p$. Note that analogous statement is not even close to being true for $\mathbb{Q}$ and $\mathbb{Z}$.

This construction applies more generally to the field of fractions of any discrete valuation ring, and a converse is true. Suppose we have a field k with a discrete valuation, which we recall is a function $v\colon k \to \mathbb{Z} \bigcup \{\infty\}$ that satisfies:

(1) $v(a) = \infty$ if and only if $a = 0$,

(2) $v(ab) = v(a) + v(b)$,

(3) $v(a + b) \geq \min(v(a), v(b))$.

The subset of k with nonnegative valuations is a discrete valuation ring R, called the *valuation ring of k*, and k is its fraction field. As with p-adic fields, the unit group of the valuation ring of k consists of those elements whose valuation is zero.

5.2 Absolute values

Having defined $\mathbb{Q}_p$ as the fraction field of $\mathbb{Z}_p$ and noting that it contains $\mathbb{Q}$, we now want to consider an alternative (but equivalent) approach that constructs $\mathbb{Q}_p$ directly from $\mathbb{Q}$. We can then obtain $\mathbb{Z}_p$ as the valuation ring of $\mathbb{Q}$.

Definition 5.2. Let k be a field. An *absolute value* on k is a function $\| \,\|\colon k \to \mathbb{R}_{\geq 0}$ with the following properties:

(1) $\|x\| = 0$ if and only if $x = 0$,

(2) $\|xy\| = \|x\| \cdot \|y\|$,

(3) $\|x + y\| \leq \|x\| + \|y\|$.

Andrew V. Sutherland

The last property is known as the *triangle inequality*, and it is equivalent to

(3) $\|x - y\| \geq \|x\| - \|y\|$

(replace x by $x \pm y$ to derive one from the other). The stronger property

(3') $\|x + y\| \leq \max(\|x\|, \|y\|)$

is known as the *nonarchimedean triangle inequality* An absolute value that satisfies (3') is called *nonarchimedean*, and is otherwise called *archimedean*.

Absolute values are sometimes called "norms", but since number theorists use this term with a more specific meaning, we will stick with absolute value. Examples of absolute values are the usual absolute value $|\ |$ on $\mathbb{R}$ or $\mathbb{C}$, which is archimedean and the *trivial absolute value* for which $\|x\| = 1$ for all $x \in k^\times$, which is nonarchimedean. To obtain non-trivial examples of nonarchimedean absolute values, if k is any field with a discrete valuation v and c is any positive real number less than 1, then it is easy to check that $\|x\|_v := c^{v(x)}$ defines a nonarchimedean absolute value on k (where we interpret c^∞ as 0). Applying this to the p-adic valuation v_p on $\mathbb{Q}_p$ with $c = 1/p$ yields the p-adic absolute value $|\ |_p$ on $\mathbb{Q}_p$:

$$|x|_p = p^{-v_p(x)}.$$

We now prove some useful facts about absolute values.

Theorem 5.3. *Let k be a field with absolute value $\|\ \|$ and multiplicative identity 1_k.*

(a) $\|1_k\| = 1$.

(b) $\|-x\| = \|x\|$.

(c) $\|\ \|$ *is nonarchimedean if and only if* $\|n\| \leq 1$ *for all positive integers* $n \in k$.

Proof. For (a), note that $\|1_k\| = \|1_k\| \cdot \|1_k\|$ and $\|1_k\| \neq 0$ since $1_k \neq 0_k$. For (b), the positive real number $\|-1_k\|$ satisfies $\|-1_k\|^2 = \|(-1_k)^2\| = \|1_k\| = 1$, and therefore $\|-1_k\| = 1$. We then have $\|-x\| = \|(-1_k)x\| = \|-1_k\| \cdot \|x\| = 1 \cdot \|x\| = \|x\|$.

To prove (c), we first note that a positive integer $n \in k$ is simply the n-fold sum $1_k + \cdots + 1_k$. If $\|\ \|$ is nonarchimedean, then for any positive integer $n \in k$, repeated application of the nonarchimedean triangle inequality yields

$$\|n\| = \|1_k + \cdots + 1_k\| \leq \max(\|1_k\|, \ldots, \|1_k\|) = 1.$$

If $\|\ \|$ is instead archimedean, then we must have $\|x+y\| > \max(\|x\|, \|y\|)$ for some $x, y \in k^\times$. We can assume without loss of generality that $\|x\| \geq \|y\|$, and if we divide through by $\|y\|$ and replace x/y with x, we can assume $y = 1$. We then have $\|x\| \geq 1$ and

$$\|x + 1\| > \max(\|x\|, 1) = \|x\|.$$

If we divide both sides by $\|x\|$ and let $z = 1/x$ we then have $\|z\| \leq 1$ and $\|z + 1\| > 1$. Now suppose for the sake of contradiction that $\|n\| \leq 1$ for all integers $n \in k$. then

$$\|z + 1\|^n = \|(z + 1)^n\| = \left\| \sum_{i=0}^{n} \binom{n}{i} z^i \right\| \leq \sum_{i=0}^{n} \left\| \binom{n}{i} \right\| \|z\|^i \leq \sum_{i=0}^{n} \left\| \binom{n}{i} \right\| \leq n + 1.$$

But $\|z + 1\| > 1$, so the LHS increases exponentially with n while the RHS is linear in n, so for any sufficiently large n we obtain a contradiction. □

Corollary 5.4. *In a field k of positive characteristic p every absolute value $\| \ \|$ is nonarchimedean and is moreover trivial if k is finite.*

Proof. Every positive integer $n \in k$ lies in the prime field $\mathbb{F}_p \subseteq k$ and therefore satisfies $n^{p-1} = 1$. This means the positive real number $\|n\|$ is a root of unity and therefore equal to 1, so $\|n\| = 1$ for all positive integers $n \in k$ and $\| \ \|$ is therefore nonarchimedean, by part (c) of Theorem 5.3. If $k = \mathbb{F}_q$ is a finite field, then for every nonzero $x \in \mathbb{F}_q$ we have $x^{q-1} = 1$ and the same argument implies $\|x\| = 1$ for all $x \in \mathbb{F}_q^\times$. □

5.3 Absolute values on $\mathbb{Q}$

As with $\mathbb{Q}_p$, we can use the p-adic valuation v_p on $\mathbb{Q}$ to construct an absolute value. Note that we can define v_p without reference to $\mathbb{Z}_p$: for any integer $v_p(a)$, is the largest integer n for which $p^n|a$, and for any rational number a/b in lowest terms we define

$$v_p\left(\frac{a}{b}\right) = v_p(a) - v_p(b).$$

This of course completely consistent with our definition of v_p on $\mathbb{Q}_p$. We then define the *p-adic absolute value* of a rational number x to be

$$|x|_p = p^{-v_p(x)},$$

with $|0|_p = p^{-\infty} = 0$, as above. Notice that rational numbers with *large* p-adic valuations have *small* p-adic absolute values. In p-adic terms, p^{100} is a very small number, and p^{1000} is even smaller. Indeed,

$$\lim_{n\to\infty} |p^n| = \lim_{n\to\infty} p^{-n} = 0.$$

We also have the usual archimedean absolute value on $\mathbb{Q}$, which we will denote by $| \ |_\infty$, for the sake of clarity. One way to remember this notation is to note that archimedean absolute values are unbounded on $\mathbb{Z}$ while nonarchimedean absolute values are not (this follows from the proof of Theorem 5.3).

We now wish to prove Ostrowski's theorem, which states that every nontrivial absolute value on $\mathbb{Q}$ is equivalent either to one of the nonarchimedean absolute values $| \ |_p$, or to $| \ |_\infty$. We first define what it means for two absolute values to be equivalent.

Definition 5.5. Two absolute values $\| \ \|$ and $\| \ \|'$ on a field k are said to be *equivalent* if there is a positive real number α such that

$$\|x\|' = \|x\|^\alpha$$

for all $x \in k$.

Note that two equivalent absolute values are either both archimedean or both nonarchimedean, by Theorem 5.3 part (c), since $c^\alpha \le 1$ if and only if $c \le 1$, for any $c, \alpha \in \mathbb{R}_{>0}$.

Theorem 5.6 (Ostrowski)**.** *Every nontrivial absolute value on $\mathbb{Q}$ is equivalent to some $| \ |_p$, where p is either a prime, or $p = \infty$.*

Proof. Let $\| \ \|$ be a nontrivial absolute value on $\mathbb{Q}$. If $\| \ \|$ is archimedean then $\|b\| > 1$ for some positive integer b. Let b be the smallest such integer and let α be the positive real number for which $\|b\| = b^\alpha$ (such an α exists because we necessarily have $b > 1$). Every other positive integer n can be written in base b as

$$n = n_0 + n_1 b + n_2 b^2 + \cdots + n_t b^t,$$

with integers $n_i \in [0, b-1]$ and $n_t \neq 0$. We then have

$$\begin{aligned}
\|n\| &\le \|n_0\| + \|n_1 b\| + \|n_2 b^2\| + \cdots + \|n_t b^t\| \\
&= \|n_0\| + \|n_1\| b^\alpha + \|n_2\| b^{2\alpha} + \cdots + \|n_t\| b^{t\alpha} \\
&\le 1 + b^\alpha + b^{2\alpha} + \cdots + b^{t\alpha} \\
&= \left(1 + b^{-\alpha} + b^{-2\alpha} + \cdots + b^{-t\alpha}\right) b^{t\alpha} \\
&\le c b^{t\alpha} \\
&\le c n^\alpha
\end{aligned}$$

where c is the sum of the geometric series $\sum_{i=0}^{\infty} (b^{-\alpha})^i$, which converges because $b^{-\alpha} < 1$. This holds for every positive integer n, so for any integer $N \ge 1$ we have

$$\|n\|^N = \|n^N\| \le c(n^N)^\alpha = c(n^{\alpha N})$$

and therefore $\|n\| \le c^{1/N} n^\alpha$. Taking the limit as $N \to \infty$ we obtain

$$\|n\| \le n^\alpha,$$

for every positive integer n. On the other hand, for any positive integer n we can choose an integer t so that $b^t \le n < b^{t+1}$. By the triangle inequality $\|b^{t+1}\| \le \|n\| + \|b^{t+1} - n\|$, so

$$\begin{aligned}
\|n\| &\ge \|b^{t+1}\| - \|b^{t+1} - n\| \\
&= b^{(t+1)\alpha} - \|b^{t+1} - n\| \\
&\ge b^{(t+1)\alpha} - (b^{t+1} - n)^\alpha \\
&\ge b^{(t+1)\alpha} - (b^{t+1} - b^t)^\alpha \\
&= b^{(t+1)\alpha}\left(1 - (1 - b^{-1})^\alpha\right) \\
&\ge d n^\alpha
\end{aligned}$$

for some real number $d > 0$ that does not depend on n. Thus $\|n\| \ge d n^\alpha$ holds for all positive integers n and, as before, by replacing n with n^N, taking Nth roots, and then taking the limit as $N \to \infty$, we deduce that

$$\|n\| \ge n^\alpha,$$

and therefore $\|n\| = n^\alpha = |n|_\infty^\alpha$ for all positive integers n. For any other positive integer m,

$$\begin{aligned}
\|n\| \cdot \|m/n\| &= \|m\| \\
\|m/n\| &= \|m\|/\|n\| = m^\alpha / n^\alpha = (m/n)^\alpha,
\end{aligned}$$

and therefore $\|x\| = x^\alpha = |x|_\infty^\alpha$ for every positive $x \in \mathbb{Q}$, and $\|-x\| = \|x\| = x^\alpha = |-x|_\infty^\alpha$, so $\|x\| = |x|_\infty^\alpha$ for all $x \in \mathbb{Q}$ (including 0).

We now suppose that $\|\ \|$ is nonarchimedean. If $\|b\| = 1$ for all positive integers b then the argument above proves that $\|x\| = 1$ for all nonzero $x \in \mathbb{Q}$, which is a contradiction since $\| \ \|$ is nontrivial. So let b be the least positive integer with $\|b\| < 1$. We must have $b > 1$, so b is divisible by a prime p. If $b \neq p$ then $\|b\| = \|p\|\|b/p\| = 1 \cdot 1 = 1$, which contradicts $\|b\| < 1$, so $b = p$ is prime.

We know prove by contradiction that p is the only prime with $\|p\| < 1$. If not then let $q \neq p$ be a prime with $\|q\| < 1$ and write $up + vq = 1$ for some integers u and v, both of which have absolute value at most 1, since $\| \ \|$ is nonarchimedean.[1] We then have

$$1 = \|1\| = \|up + vq\| \leq \max(\|up\|, \|vq\|) = \max(\|u\| \cdot \|p\|, \|v\| \cdot \|q\|) \leq \max(\|p\|, \|q\|) < 1,$$

which is a contradiction.

Now define the real number $\alpha > 0$ so that $\|p\| = p^{-\alpha} = |p|_p^{\alpha}$. Any positive integer n may be written as $n = p^{v_p(n)}r$ with $v_p(r) = 0$, and we then have

$$\|n\| = \|p^{v_p(n)}r\| = \|p^{v_p(n)}\| \cdot \|r\| = \|p\|^{v_p(n)} = |p|_p^{\alpha v_p(n)} = |n|_p^{\alpha}.$$

This then extends to all rational numbers, as argued above. □

[1] This is a simplification of the argument given in class, as pointed out by Ping Ngai Chung (Brian).

18.782 Introduction to Arithmetic Geometry **Fall 2013**
Lecture #7 **09/26/2013**

In Lecture 6 we proved (most of) Ostrowski's theorem for number fields, and we saw the product formula for absolute values on $\mathbb{Q}$. A similar product formula holds for absolute values on a number field, but in order to state and prove it we need to briefly review/introduce some standard terminology from algebraic number theory.

7.1 Field norms and traces

Let L/K be a finite field extension of degree $n = [L : K]$. Then L is an n-dimensional K-vector space, and each $\alpha \in L$ determines a linear operator $T_\alpha \colon L \to L$ corresponding to multiplication by α (the linearity of T_α is immediate from the field axioms).

Definition 7.1. The *trace* $\operatorname{Tr}_{L/K}(\alpha)$ is the trace of T_α, and the *norm* $N_{L/K}(\alpha)$ is the determinant of T_α.[1]

It follows immediately from this definition that the trace is additive and the norm is multiplicative, and that both take values in K.

The trace and norm can be computed as the trace and determinant of the matrix of T_α with respect to a basis, but their values are intrinsic to α and do not depend on a choice of basis. The Cayley-Hamilton theorem implies that T_α satisfies a characteristic equation

$$f_\alpha(x) = x^n + a_{n-1}x + \cdots + a_1x + a_0 = 0$$

with coefficients $a_i \in K$. We then have

$$\operatorname{Tr}_{L/K}(\alpha) = -a_{n-1} \qquad \text{and} \qquad N_{L/K}(\alpha) = (-1)^n a_0,$$

equivalently, $\operatorname{Tr}_{L/K}(\alpha)$ and $N_{L/K}(\alpha)$ are the sum and product of the roots of f_α, respectively. These roots need not lie in L, but they certainly lie in $\overline{K}$ (in fact in the splitting field of f_α), and in any case their sum and product necessarily lie in K.

Note that α satisfies the same characteristic equation as T_α, since T_α is just multiplication by α, but f_α is not necessarily the minimal polynomial g_α of α over K (which is also the minimal polynomial of the operator T_α). We know that g_α must divide f_α, since the minimal polynomial always divides the characteristic polynomial, but f_α must be a power of g_α. This is easy (and instructive) to prove in the case that L/K is a separable extension, which includes all the cases of interest to us.[2]

Theorem 7.2. *Let L/K be a separable field extension of degree n, let $\alpha \in L$ have minimal polynomial g_α over K and let f_α be the characteristic polynomial of T_α Then*

$$f_\alpha = g_\alpha^{n/d},$$

where $d = [K(\alpha) : K]$.

[1] These are also called the *relative* trace/norm, or the trace/norm *from L down to K* to emphasize that they depend on the fields L and K, not just α.

[2] Recall that *separable* means that minimal polynomials never have repeated roots. In characteristic zero every finite extension is separable, and the same holds for finite fields (such fields are said to be *perfect*).

Andrew V. Sutherland

Proof. There are exactly n distinct embeddings $\sigma_1, \ldots, \sigma_n$ of L into $\overline{K}$ that fix K, and $\sigma_1(\alpha), \ldots, \sigma_n(\alpha)$ are precisely the n (not necessarily distinct) roots of f_α. This list includes the d roots of g_α, since g_α divides f_α, and these d roots are distinct, since L/K is separable. But there are exactly $n/d = [L : K(\alpha)]$ distinct embeddings of L into $\overline{K}$ that fix $K(\alpha)$, and each of these also fixes K and is hence one of the σ_i. It follows that each distinct root of f_α occurs with multiplicity at least n/d, and since f_α has at least d distinct roots, the roots of f_α are precisely the roots of g_α, each occuring with multiplicity n/d. Both f_α and g_α are monic, so $f_\alpha = g_\alpha^{n/d}$. □

7.2 Ideal norms

Now let us fix $K = \mathbb{Q}$, so that L is a number field (a finite extension of $\mathbb{Q}$). Recall that the *ring of integers* of L consists of the elements in L whose minimal polynomials have integer coefficients. This subset forms a ring $\mathcal{O}$ that is a *Dedekind domain*, an integral domain in which every nonzero proper ideal can be uniquely factored into prime ideals (equivalently, a finitely generated Noetherian ring in which every nonzero prime ideal is maximal), and L is its fraction field. The ring of integers is a free $\mathbb{Z}$-module of rank $n = [L : \mathbb{Q}]$, and we can pick a basis for L as an n-dimensional $\mathbb{Q}$-vector space that consists of elements of $\mathcal{O}$ (such a basis is called an *integral basis*). The ring $\mathcal{O}$ then consists of all integer linear combinations of basis elements and can be viewed as an n-dimensional $\mathbb{Z}$-lattice. For proofs of these facts, see any standard text on algebraic number theory, such as [1].

Definition 7.3. Let $\mathfrak{a}$ be a nonzero $\mathcal{O}$-ideal. The (ideal) *norm* $N\mathfrak{a}$ of $\mathfrak{a}$ is the cardinality of the (necessarily finite) ring $\mathcal{O}/\mathfrak{a}$, equivalently, the index $[\mathcal{O} : \mathfrak{a}]$ of $\mathfrak{a}$ as a sublattice of the $\mathbb{Z}$-lattice $\mathcal{O}$.[3] The norm of (0) is zero.

Remark 7.4. In a Dedekind domain every nonzero prime ideal is maximal, so for prime ideals $\mathfrak{p}$ the ring $\mathcal{O}/\mathfrak{p}$ is actually a field of cardinality $N\mathfrak{p} = p^f$, for some prime p and positive integer f called the *inertia degree* (also *residue degree*).

While it may not be immediately obvious from the definition, the ideal norm is multiplicative (for principal ideals this follows from Theorem 7.5 below). For an algebraic integer $\alpha \in L$ we now have two notions of norm: the field norm $N_{L/\mathbb{Q}}(\alpha)$ and the ideal norm $N(\alpha)$ of the prinicipal $\mathcal{O}$-ideal generated by α. These are not unrelated.

Theorem 7.5. *Let α be an algebraic integer in a number field L. Then $N(\alpha) = |N_{L/\mathbb{Q}}(\alpha)|$.*

Proof. Fix an integral basis $\mathcal{B}$ for L. The field norm $N_{L/\mathbb{Q}}(\alpha)$ is the determinant of the matrix of the linear operator T_α with respect to $\mathcal{B}$. The absolute value of this determinant is equal to the volume of a fundamental parallelepiped in the $\mathbb{Z}$-lattice corresponding to the principal ideal (α) as a sublattice of the $\mathbb{Z}$-lattice $\mathcal{O}$ generated by $\mathcal{B}$, relative to the volume of a fundamental parallelepiped in $\mathcal{O}$. But this is precisely the index $[\mathcal{O} : (\alpha)] = N(\alpha)$. □

[3] Like the field norm $N_{L/\mathbb{Q}}$, the ideal norm N depends on L, but we typically don't indicate L in the notation because N is always applied to ideals, which necessarily exist in the context of a particular ring (in our case the ring of integers of L). More generally, for any finite separable extension L/K where K is the fraction field of a Dedekind domain A, the ideal norm is defined as a map from ideals in the integral closure of A in L to A-ideals. In our setting $A = \mathbb{Z}$ is a PID, so we are effectively identifying the $\mathbb{Z}$-ideal $(N\mathfrak{a})$ with the integer $N\mathfrak{a}$. See [1, Ch. 4] for more details. Our definition here is also called the *absolute* norm.

7.3 Product formula for absolute values on number fields

Ostrowski's theorem for number fields classifies the absolute values on a number field up to equivalence. But in order to prove the product formula we need to properly normalize each absolute value appropriately, which we now do.

Let L be a number field with ring of integers $\mathcal{O}$. For each nonzero prime ideal $\mathfrak{p}$ in $\mathcal{O}$ we define the absolute value $|\alpha|_\mathfrak{p}$ on L by

$$|\alpha|_\mathfrak{p} = (N\mathfrak{p})^{-v_\mathfrak{p}(\alpha)},$$

where $v_\mathfrak{p}(\alpha)$ is the exponent of $\mathfrak{p}$ in the prime factorization of the ideal (α) for nonzero $\alpha \in \mathcal{O}$, and $v_\mathfrak{p}(\alpha/\beta) = v_\mathfrak{p}(\alpha) - v_\mathfrak{p}(\beta)$ for any nonzero $\alpha, \beta \in \mathcal{O}$ (recall that L is the fraction field of $\mathcal{O}$). As usual, we let $v_\mathfrak{p}(0) = \infty$ and define $(N\mathfrak{p})^{-\infty} = 0$.

This addresses all the nonarchimedean absolute values of L (by Ostrowski's theorem), we now consider the archimedean ones. As a number field of degree n, there are exactly n distinct embeddings of L into $\overline{\mathbb{Q}}$, hence into $\mathbb{C}$. But these n embeddings do not necessarily give rise to n distinct absolute values. Let f be a defining polynomial for L over $\mathbb{Q}$, that is, the minimal polynomial of a primitive element θ such that $L = \mathbb{Q}(\theta)$ (such a θ exists, by the primitive element theorem). Over $\mathbb{C}$, the roots of f are either real (let r be the number of real roots) or come in complex-conjugate pairs (let s be the number of such pairs). We then have $n = r + 2s$ distinct embeddings of L into $\mathbb{C}$, each sending θ to a different root of f (the roots are distinct because every finite extension of $\mathbb{Q}$ is separable). But there are only $r + s$ inequivalent archimedean absolute values on L, since complex-conjugate embeddings yield the same absolute value ($|z| = |\bar{z}|$).

As with $\mathbb{Q}$, it will be convenient to use the notation $|\ |_\mathfrak{p}$ to denote archimedean absolute values as well as nonarchimedean ones, and we may refer to the subscript $\mathfrak{p}$ as an archimedean or "infinite" prime and write $\mathfrak{p}|\infty$ to indicate this.[4] Using $\sigma_\mathfrak{p}$ to denote the embedding associated to a real archimedean prime $\mathfrak{p}$ and $\sigma_\mathfrak{p}, \bar{\sigma}_\mathfrak{p}$ to denote the conjugate pair of complex embeddings associated to a complex archimedean prime $\mathfrak{p}$, we now define

$$|\alpha|_\mathfrak{p} = \begin{cases} |\sigma_\mathfrak{p}(\alpha)| & \text{if } \mathfrak{p} \text{ is a real archimedean prime,} \\ |\sigma_\mathfrak{p}(\alpha)| \cdot |\bar{\sigma}_\mathfrak{p}(\alpha)| & \text{if } \mathfrak{p} \text{ is a complex archimedean prime.} \end{cases}$$

Of course $|\sigma_\mathfrak{p}(\alpha)| \cdot |\bar{\sigma}_\mathfrak{p}(\alpha)| = |\sigma_\mathfrak{p}(\alpha)|^2$, but it is more illuminating to write it as above.

We now prove the product formula for absolute values on number fields.

Theorem 7.6. *Let L be a number field. For every $\alpha \in L^\times$ we have*

$$\prod_\mathfrak{p} |\alpha|_\mathfrak{p} = 1,$$

where $\mathfrak{p}$ ranges over all the primes of L (both finite and infinite).

Proof. We first consider the archimedean primes. Let f_α be the characteristic polynomial of the linear operator on the $\mathbb{Q}$-vector space L corresponding to multiplication by α. If $\mathfrak{p}_1, \ldots, \mathfrak{p}_r$ and $\mathfrak{p}_{r+1}, \ldots, \mathfrak{p}_{r+s}$ are the real and complex archimedean primes of L, then the $n = r + 2s$ (not nescessarily distinct) roots of f_α are precisely

$$\sigma_{\mathfrak{p}_1}(\alpha), \ldots, \sigma_{\mathfrak{p}_r}(\alpha), \sigma_{\mathfrak{p}_{r+1}}(\alpha), \bar{\sigma}_{\mathfrak{p}_{r+1}}(\alpha), \ldots, \sigma_{\mathfrak{p}_{r+s}}(\alpha), \bar{\sigma}_{\mathfrak{p}_{r+s}}(\alpha).$$

[4] The finite and infinite primes of L are also often referred to as *places* of L and denoted by v.

We then have

$$\prod_{\mathfrak{p}|\infty} |\alpha|_{\mathfrak{p}} = \prod_{i=1}^{r} |\sigma_{\mathfrak{p}_i}(\alpha)| \prod_{i=r+1}^{s} |\sigma_{\mathfrak{p}_i}(\alpha)| \cdot |\bar{\sigma}_{\mathfrak{p}_i}(\alpha)| = |N_{L/\mathbb{Q}}(\alpha)|,$$

since $N_{L/\mathbb{Q}}(\alpha)$ is equal to the product of the roots of f_α.

Now let $(\alpha) = \mathfrak{q}_1^{a_1} \cdots \mathfrak{q}_t^{a_t}$ be the prime factorization of the principal ideal (α) in the ring of integers of L. Then

$$\prod_{\mathfrak{p}<\infty} |\alpha|_{\mathfrak{p}} = \prod_{i=1}^{t} (N\mathfrak{q}_i)^{-a_i} = N(\alpha)^{-1} = |N_{L/\mathbb{Q}}(\alpha)|^{-1},$$

by Theorem 7.5, and therefore $\prod_{\mathfrak{p}} |\alpha|_{\mathfrak{p}} = 1$, as desired. □

We now turn to a new topic, the *completion* of a field with respect to an absolute value.

7.4 Cauchy sequences and convergence

We begin with the usual definitions of convergence and Cauchy sequences, which apply to any field with an absolute value. Let k be a field equipped with an absolute value $\| \ \|$.

Definition 7.7. A sequence (x_n) of elements of k *converges* (to ℓ) if there is an element $\ell \in k$ such that for every $\epsilon > 0$ there is a positive integer N such that $\|x_n - \ell\| < \epsilon$ for all $n \geq N$. Equivalently, (x_n) converges to ℓ if $\|x_n - \ell\| \to 0$ as $n \to \infty$.[5]

The element ℓ is called the *limit* of the sequence, and if it exists, it is unique: if (x_n) converges to both ℓ and ℓ' then

$$\|\ell' - \ell\| = \|\ell' - x_n + x_n - \ell\| \leq \|\ell' - x_n\| + \|x_n - \ell\| = \|x_n - \ell'\| + \|x_n - \ell\| \to 0 + 0 = 0,$$

so $\|\ell' - \ell\| = 0$, and therefore $\ell' - \ell = 0$ and $\ell' = \ell$ (note that we used $\|-x\| = \|x\|$).

Sums and products of convergent sequences behave as expected.

Lemma 7.8. *Let (x_n) and (y_n) be sequences in k that converge to x and y respectively. Then the sequences $(x_n + y_n)$ and $(x_n y_n)$ convege to $x + y$ and xy respectively.*

Proof. Convergence of $(x_n y_n)$ to xy follows immediately from the multiplicativity of $\| \ \|$. To check $(x_n + y_n)$, for any $\epsilon > 0$ pick N so that $\|x - x_n\| < \epsilon/2$ and $\|y - y_n\| < \epsilon/2$ for all $n \geq N$. Then $\|(x_n + y_n) - (x + y)\| \leq \|x_n - x\| + \|y_n - y\| < \epsilon/2 + \epsilon/2 = \epsilon$ for all $n \geq N$. □

We now recall a necessary condition for convergence.

Definition 7.9. A sequence (x_n) in k is a *Cauchy sequence* if for every $\epsilon > 0$ there exists a positive integer N such that $\|x_m - x_n\| < \epsilon$ for all $m, n \geq N$.

Theorem 7.10. *Every convergent sequence is a Cauchy sequence.*

Proof. Suppose (x_n) is a convergent sequence. For any $\epsilon > 0$ there is a positive integer N for which $\|x_n - \ell\| < \epsilon/2$ for all $n \geq N$. For all $m, n \geq N$ we then have

$$\|x_m - x_n\| = \|x_m - \ell + \ell - x_n\| \leq \|x_m - \ell\| + \|\ell - x_n\| = \|x_m - \ell\| + \|x_n - \ell\| < \epsilon/2 + \epsilon/2 = \epsilon,$$

where we have again used $\|-x\| = \|x\|$. □

[5] The notation $\|x_n - \ell\| \to 0$ refers to convergence in $\mathbb{R}$ in the usual sense.

The converse of Theorem 7.10 is not necessarily true, it depends on the field k.

Definition 7.11. A field k is *complete* (with respect to $\| \ \|$) if every Cauchy sequence in k converges (to an element of k).

Every field is complete with respect to the trivial absolute value. The field $\mathbb{Q}$ is not complete with respect to the archimedean absolute value $| \ |$, but $\mathbb{R}$ is; indeed, $\mathbb{R}$ can be (and often is) defined as the smallest field containing $\mathbb{Q}$ that is complete with respect to $| \ |$, in other words, $\mathbb{R}$ is the completion of $\mathbb{Q}$. In order to formally define the completion of a field, we define an equivalence relation on sequences.

Definition 7.12. Two sequences (a_n) and (b_n) are *equivalent* if $\|a_n - b_n\| \to 0$ as $n \to \infty$.

It is easy to check that this defines an equivalence relation on the set of all sequences in k, and that any sequence equivalent to a Cauchy sequence is necessarily a Cauchy sequence. We may use the notation $[(x_n)]$ to denote the equivalence class of the sequence (x_n).

Definition 7.13. The *completion* of k (with respect to $\| \ \|$) is the field $\hat{k}$ whose elements are equivalence classes of Cauchy sequences in k, where

(1) $0_{\hat{k}} = [(0_k, 0_k, 0_k, \ldots)]$,
(2) $1_{\hat{k}} = [(1_k, 1_k, 1_k, \ldots)]$,
(3) $[(x_n)] + [(y_n)] = [(x_n + y_n)]$ and $[(x_n)][(y_n)] = [(x_n y_n)]$.

To verify that that actually defines a field, the only nontrivial thing to check is that every nonzero element has a multiplicative inverse. So let $[(x_n)]$ be a nonzero element of $\hat{k}$. The Cauchy sequence (x_n) must be eventually nonzero (otherwise it would be equivalent to zero), and if we consider the element $[(y_n)] \in \hat{k}$ defined by

$$y_n = \begin{cases} x_n^{-1} & \text{if } x_n \neq 0, \\ 0 & \text{if } x_n = 0, \end{cases}$$

we see that $[(x_n)][(y_n)] = 1$, since the sequence $(x_n y_n)$ is eventually 1.

The map $x \mapsto \hat{x} = [(x, x, x, \ldots)]$ is clearly a ring homomorphism from k to $\hat{k}$, and therefore a field embedding. We thus view $\hat{k}$ as an extension of k by identifying k with its image in $\hat{k}$.

We now extend the absolute value of k to $\hat{k}$ by defining

$$\|[(x_n)]\| = \lim_{n\to\infty} \|x_n\|.$$

This limit exists because $(\|x_n\|)$ is a Cauchy sequence of real numbers and $\mathbb{R}$ is complete, and we must get the same limit for any Cauchy sequence (y_n) equivalent to (x_n), so this definition does not depend on the choice of representative for the equivalence class $[(x_n)]$. Since $\|\hat{x}\| = \|x\|$ for any $x \in k$, this definition is compatible with our original $\| \ \|$.

We now note that any Cauchy sequence (x_n) in k can be viewed as a Cauchy sequence $(\hat{x}_n)$ in $\hat{k}$, since we view k as a subfield of $\hat{k}$, and $(\hat{x}_n)$ obviously converges to $[(x_n)]$ in $\hat{k}$. Thus every Cauchy sequence in $\hat{k}$ that consists entirely of elements of k converges. But what about other Cauchy sequences in $\hat{k}$? To show that these also converge we use the fact that k is dense in $\hat{k}$.

Definition 7.14. Let S be any subset of a field k with absolute value $\| \ \|$. The set S is *dense* in k if for every $x \in k$ and every $\epsilon > 0$ there exists $y \in S$ such that $\|x - y\| < \epsilon$.

Theorem 7.15. *Let k be a field with absolute value $\| \; \|$. Then k is dense in its completion $\hat{k}$.*

Proof. Let $x \in \hat{k}$ be the equivalence class of the Cauchy sequence (x_n) in k. For any $\epsilon > 0$ there is an x_m with the property that $\|x_m - x_n\|$ for all $n \geq m$. It follows that $\|x - \hat{x}_m\| < \epsilon$, where $\hat{x}_m \in k \subseteq \hat{k}$ is just the equivalence class of $(x_m, x_m, x_m, \ldots)$. □

Corollary 7.16. *Every Cauchy sequence in $\hat{k}$ is equivalent to a Cauchy sequence whose elements lie in k.*

Proof. Let (z_n) be a Cauchy sequence in $\hat{k}$. Since k is dense in $\hat{k}$, for each z_n we may pick $x_n \in k \subseteq \hat{k}$ so that $\|z_n - x_n\| < 1/n$. Then for any $\epsilon > 0$ we may pick N such that $\|z_m - x_m\| < \epsilon/3$, $\|z_n - x_n\| < \epsilon/3$ and $\|z_m - z_n\| < \epsilon/3$, for all $m, n \geq N$. It then follows from the triangle inequality that $\|x_m - x_n\| < \epsilon$ for all $m, n \geq N$, hence (x_n) is Cauchy. □

Corollary 7.17. *The completion $\hat{k}$ of k is complete. Moreover it is the smallest complete field containing k in the following sense: any embedding of k in a complete field k' can be extended to an embedding of $\hat{k}$ into k'.*

Proof. The first statement follows immediately from Corollary 7.16 and the discussion above. For the second, if $\pi\colon k \to k'$ is an embedding of k into a complete field k', then we can extend π to an embedding of $\hat{k}$ into k' by defining

$$\pi([(x_n)]) = \lim_{n \to \infty} \pi(x_n).$$

Such a limit always exists, since k' is complete, and the map $\pi\colon \hat{k} \to k'$ is a ring homomorphism (hence a field embedding) because taking limits commutes with addition and multiplication, by Lemma 7.8. □

Remark 7.18. We could have defined $\hat{k}$ more categorically as the field with the universal property that every embedding of k into a complete field can be extended to $\hat{k}$. Assuming it exists, such a $\hat{k}$ is unique up to a canonical isomorphism (map Cauchy sequences to their limits), but we still would have to prove existence.

Finally, we note that the absolute value on the completion of k with respect to $\| \; \|$ is nonarchimedean if and only if the absolute value on k is nonarchimedean.

Remark 7.19. Everything we have done here applies more generally to commutative rings. For example, $\mathbb{Z}_p$ is the completion of $\mathbb{Z}$ with respect to the p-adic absolute value $| \; |_p$ on $\mathbb{Z}$, as we will see in the next lecture.

References

[1] J. S. Milne, *Algebraic number theory*, 2013.

18.782 Introduction to Arithmetic Geometry **Fall 2013**
Lecture #8 **10/1/2013**

8.1 Completions of $\mathbb{Q}$

We already know that $\mathbb{R}$ is the completion of $\mathbb{Q}$ with respect to its archimedean absolute value $|\ |_\infty$. Now we consider the completion of $\mathbb{Q}$ with respect to any of its nonarchimedean absolute values $|\ |_p$.

Theorem 8.1. *The completion $\hat{\mathbb{Q}}$ of $\mathbb{Q}$ with respect to the p-adic absolute value $|\ |_p$ is isomorphic to $\mathbb{Q}_p$. More precisely, there is an isomorphism $\pi\colon \mathbb{Q}_p \to \hat{\mathbb{Q}}$ that satisifies $|\pi(x)|_p = |x|_p$ for all $x \in \hat{\mathbb{Q}}$.*

Proof. For any $x \in \mathbb{Q}_p$ either $x \in \mathbb{Z}_p$ or $x^{-1} \in \mathbb{Z}_p$, since $\mathbb{Z}_p = \{x \in \mathbb{Q}_p : |x|_p \le 1\}$, so to define π it is enough to give a ring homomorphism from $\mathbb{Z}_p$ to $\hat{\mathbb{Q}}$. Let us uniquely represent each $a \in \mathbb{Z}_p$ as a sequence of integers (a_n) with $a_n \in [0, p^n - 1]$, such that $a_{n+1} \equiv a_n \bmod \mathbb{Z}/p^n\mathbb{Z}$. For any $\epsilon > 0$ there is an integer N such that $p^{-N} < \epsilon$, and we then have $|a_m - a_n|_p < \epsilon$ for all $m, n \ge N$. Thus each $a \in \mathbb{Z}_p$ corresponds to a sequence of integers (a_n) that is Cauchy with respect to the p-adic absolute value on $\mathbb{Q}$ and we define $\pi(a)$ to be the equivalence class of (a_n) in $\hat{\mathbb{Q}}$. It follows immediately from the definition of addition and multiplication in both $\mathbb{Z}_p$ and $\hat{\mathbb{Q}}$ as element-wise operations on representative sequences that π is a ring homomorphism from $\mathbb{Z}_p$ to $\hat{\mathbb{Q}}$. Moreover, π preserves the absolute value $|\ |_p$, since

$$|a|_p = \lim_{n\to\infty} |a_n|_p = |\pi(a)|_p.$$

Here the first equality follows from the fact that if $v_p(a) = m$, then $a_n = 0$ for $n \le m$ and $v_p(a_n) = m$ for all $n > m$ (so the sequence $|a_n|_p$ eventually constant), and the second equality is the definition of $|\ |_p$ on $\hat{\mathbb{Q}}$.

We now extend π from $\mathbb{Z}_p$ to $\mathbb{Q}_p$ by defining $\pi(x^{-1}) = \pi(x)^{-1}$ for all $x \in \mathbb{Z}_p$ (this is necessarily consistent with our definition of π on $\mathbb{Z}_p^\times$, since π is a ring homomorphism). As a ring homomorphism of fields, $\pi\colon \mathbb{Q}_p \to \hat{\mathbb{Q}}$ must be injective, so we have an embedding of $\mathbb{Q}_p$ into $\hat{\mathbb{Q}}$. To show this it is an isomorphism, it suffices to show that $\mathbb{Q}_p$ is complete, since then we can embed $\hat{\mathbb{Q}}$ into $\mathbb{Q}_p$, by Corollary 7.17.

So let (x_n) be a Cauchy sequence in $\mathbb{Q}_p$. Then (x_n) is bounded (fix $\epsilon > 0$, pick N so that $|x_n - x_N|_p < \epsilon$ for all $n \ge N$ and note that $|x_n|_p \le \max_{n \le N}(|x_n|_p) + \epsilon$). Thus for some fixed power p^r of p the sequence $(y_n) = (p^r x_n)$ lies in $\mathbb{Z}_p$. We now define $a \in \mathbb{Z}_p$ as a sequence of integers $(a_1, a_2, \ldots)$ with $a_i \in [0, p^i - 1]$ and $a_{i+1} \equiv a_i \bmod \mathbb{Z}/p^i\mathbb{Z}$ as follows. For each integer $i \ge 1$ pick N so that $|y_n - y_N| < p^{-i}$ for all $n \ge N$. Then $v_p(y_n - y_N) \ge i$, and we let a_i be the unique integer in $[0, p^i - 1]$ for which $y_n \equiv a_i \bmod \mathbb{Z}/p^i\mathbb{Z}$ for all $n \ge N$. We necessarily have $a_{i+1} \equiv a_i \bmod p^i$, so this defines an element a of $\mathbb{Z}_p$, and by construction (y_n) converges to a and therefore (x_n) converges to a/p^r. Thus every Cauchy sequence in $\mathbb{Q}_p$ converges, so $\mathbb{Q}_p$ is complete. □

It follows from Theorem 8.1 that we could have defined $\mathbb{Q}_p$ as the completion of $\mathbb{Q}$, rather than as the fraction field of $\mathbb{Z}_p$, and many texts do exactly this. If we had taken this approach we would then define $\mathbb{Z}_p$ as the *the ring of integers* of $\mathbb{Q}_p$, that is, the ring

$$\mathbb{Z}_p = \{x \in \mathbb{Q}_p : |x|_p \le 1\}.$$

Alternatively, we could define $\mathbb{Z}_p$ as the completion of $\mathbb{Z}$ with respect to $|\ |_p$.

Andrew V. Sutherland

Remark 8.2. The use of the term "ring of integers" in the context of a p-adic field can be slightly confusing. The ring $\mathbb{Z}_p$ is the *topological closure* of $\mathbb{Z}$ in $\mathbb{Q}_p$ (in other words, the completion of $\mathbb{Z}$), but it is not the *integral closure* of $\mathbb{Z}$ in $\mathbb{Q}_p$ (the elements in $\mathbb{Q}_p$ that are roots of a monic polynomial with coefficients in $\mathbb{Z}$). The latter set is countable, since there are only countably many polynomials with integer coefficients, but we know that $\mathbb{Z}_p$ is uncountable. But it is true that $\mathbb{Z}_p$ is integrally closed in $\mathbb{Q}_p$, every element of $\mathbb{Q}_p$ that is the root of a monic polynomial with coefficients in $\mathbb{Z}_p$ lies in $\mathbb{Z}_p$, so $\mathbb{Z}_p$ certainly contains the integral closure of $\mathbb{Z}$ in $\mathbb{Q}_p$ (and is the completion of the integral closure).

8.2 Root-finding in p-adic fields

We now turn to the problem of finding roots of polynomials in $\mathbb{Z}_p[x]$. From Lecture 3 we already know how to find roots of polynomials in $(\mathbb{Z}/p\mathbb{Z})[x] \simeq \mathbb{F}_p[x]$. Our goal is to reduce the problem of root-finding over $\mathbb{Z}_p$ to root-finding over $\mathbb{F}_p$. To take the first step toward this goal we require the following compactness lemma.

Lemma 8.3. *Let (S_n) be an inverse system of finite non-empty sets with a compatible system of maps $f_n\colon S_{n+1} \to S_n$. The inverse limit $S = \varprojlim S_n$ is non-empty.*

Proof. If the f_n are all surjective, we can easily construct an element (s_n) of S: pick any $s_1 \in S_1$ and for $n \geq 1$ pick any $s_{n+1} \in f_n^{-1}(s_n)$. So our goal is to reduce to this case.

Let $T_{n,n} = S_n$ and for $m > n$, let $T_{m,n}$ be the image of S_m in S_n, that is

$$T_{m,n} = f_n(f_{n+1}(\cdots f_{m-1}(S_m)\cdots)).$$

For each n we then have an infinite sequence of inclusions

$$\cdots \subseteq T_{m,n} \subseteq T_{m-1,n} \subseteq \cdots \subseteq T_{n+1,n} \subseteq T_{n,n} = S_n.$$

The $T_{m,n}$ are all finite non-empty sets, and it follows that all but finitely many of these inclusions are equalities. Thus each infinite intersection $E_n = \bigcap_m T_{m,n}$ is a non-empty subset of S_n. Using the restriction of f_n to define a map $E_{n+1} \to E_n$, we obtain an inverse system (E_n) of finite non-empty sets whose maps are all surjective, as desired. □

Theorem 8.4. *For any $f \in \mathbb{Z}_p[x]$ the following are equivalent:*

(a) f has a root in $\mathbb{Z}_p$.

(b) $f \bmod p^n$ has a root in $\mathbb{Z}/p^n\mathbb{Z}$ for all $n \geq 1$.

Proof. $(a) \Rightarrow (b)$: apply the projection maps $\mathbb{Z}_p \to \mathbb{Z}/p^n\mathbb{Z}$ to the roots and coefficients of f. $(b) \Rightarrow (a)$: let S_n be the roots of f in $\mathbb{Z}/p^n\mathbb{Z}$ and consider the inverse system (S_n) of finite non-empty sets whose maps are the restrictions of the reduction maps from $\mathbb{Z}/p^{n+1}\mathbb{Z}$ to $\mathbb{Z}/p^n\mathbb{Z}$. By Lemma 8.3, the set $S = \varprojlim S_n \subseteq \varprojlim \mathbb{Z}/p^n\mathbb{Z} = \mathbb{Z}_p$ is non-empty, and its elements are roots of f. □

Theorem 8.4 reduces the problem of finding the roots of f in $\mathbb{Z}_p$ to the problem of finding roots of f modulo infinitely many powers of p. This might not seem like progress, but we will now show that under suitable conditions, once we have a root a_1 of $f \bmod p$, we can "lift" a_1 to a root a_n of $f \bmod p^n$, for each $n \geq 1$, and hence to a root of f in $\mathbb{Z}_p$.

A key tool in doing this is the Taylor expansion of f, which we now define in the general setting of a commutative ring R.[1]

[1] As always, our rings include a multiplicative identity 1.

Definition 8.5. Let $f \in R[x]$ be a polynomial of degree at most d and let $a \in R$. The (degree d) *Taylor expansion* of f about a is

$$f(x) = f_d(x-a)^d + f_{d-1}(x-a)^{d-1} + \cdots + f_1(x-a) + f_0,$$

with $f_0, f_1, \ldots, f_d \in R$.

The Taylor coefficients $f_0, f_1, \ldots, f_d$ are uniquely determined by the expansion of $f(y+z)$ in $R[y,z]$:

$$f(y+z) = f_d(y)z^d + f_{d-1}(y)z^{d-1} + \cdots + f_1(y)z + f_0(y).$$

Replacing y with a and z with $x-a$ yields the Taylor expansion of f with $f_i = f_i(a) \in R$.

This definition of the Taylor expansion agrees with the usual definition over $\mathbb{R}$ or $\mathbb{C}$ in terms of the derivatives of f.

Definition 8.6. Let $f(x) = \sum_{n=0}^{d} a_n x^n$ be a polynomial in $R[x]$. The *formal derivative* f' of f is the polynomial $f'(x) = \sum_{n=1}^{d} n a_n x^{n-1}$ in $R[x]$.

It is easy to check that the formal derivative satisfies the usual properties

$$\begin{aligned}(f+g)' &= f' + g',\\ (fg)' &= f'g + fg',\\ (f\circ g)' &= (f'\circ g)g'.\end{aligned}$$

Over a field of characteristic zero one then has the more familiar form of the Taylor expansion

$$f(x) = \frac{f^{(d)}(a)}{d!}(x-a)^d + \cdots + \frac{f^{(2)}(a)}{2}(x-a)^2 + f'(a)(x-a) + f(a),$$

where $f^{(n)}$ denotes the result of taking n successive derivatives ($f^{(n)}(a)$ is necessarily divisible by $n!$, so the coefficients lie in R). Regardless of the characteristic, the Taylor coefficients f_0 and f_1 always satisfy $f_0 = f(a)$ and $f_1 = f'(a)$.

Lemma 8.7. *Let $a \in R$ and $f \in R[x]$. Then $f(a) = f'(a) = 0$ if and only if a is (at least) a double root of f, that is, $f(x) = (x-a)^2 g(x)$ for some $g \in R[x]$.*

Proof. The reverse implication is clear: if $f(x) = (x-a)^2 g(x)$ then clearly $f(a) = 0$, and we have $f'(x) = 2(x-a)g(x) + (x-a)^2 g'(x)$, so $f'(a) = 0$ as well. For the forward implication, let $d = \max(\deg f, 2)$ and consider the degree d Taylor expansion of f about a:

$$f(x) = f_d(x-a)^d + f_{d-1}(x-a)^{d-1} + \cdots + f_2(x-a)^2 + f_1(x)(x-a) + f_0.$$

If $f(a) = f'(a) = 0$ then $f_0 = f(a) = 0$ and $f_1 = f'(a) = 0$ and we can put

$$f(x) = (x-a)^2 \left(f_d(x-a)^{d-2} + f_{d-2}(x-a)^{d-3} + \cdots + f_2\right),$$

in the desired form. □

8.3 Hensel's lemma

We are now ready to prove Hensel's lemma, which allows us to lift any simple root of $f \bmod p$ to a root of f in $\mathbb{Z}_p$.

Theorem 8.8 (Hensel's lemma)**.** *Let $a \in \mathbb{Z}_p$ and $f \in \mathbb{Z}_p[x]$. Suppose $f(a) \equiv 0 \bmod p$ and $f'(a) \not\equiv 0 \bmod p$. Then there is a unique $b \in \mathbb{Z}_p$ such that $f(b) = 0$ and $b \equiv a \bmod p$.*

Our strategy for proving Hensel's lemma is to apply Newton's method, regarding a as an approximate root of f that we can iteratively improve. Remarkably, unlike the situation over an archimedean field like $\mathbb{R}$ or $\mathbb{C}$, this is guaranteed to always work.

Proof. Let $a_1 = a$, and for $n \geq 1$ define

$$a_{n+1} = a_n - f(a_n)/f'(a_n).$$

We will prove by induction on n that the following hold

$$f'(a_n) \not\equiv 0 \bmod p, \tag{1}$$

$$f(a_n) \equiv 0 \bmod p^n, \tag{2}$$

Note that (1) ensures that $f'(a_n) \in \mathbb{Z}_p^\times$, so a_{n+1} is well defined and an element of $\mathbb{Z}_p$. Together with the definition of a_{n+1}, (1) and (2) imply $a_{n+1} \equiv a_n \bmod p^n$, which means that the sequence $(a_n \bmod p^n)$ defines an element of $b \in \mathbb{Z}_p$ for which $f(b) = 0$ and $b \equiv a_1 \equiv a$ modulo p (equivalently, the sequence (a_n) is a Cauchy sequence in $\mathbb{Z}_p$ with limit b).

The base case $n = 1$ is clear, so assume (1) and (2) hold for a_n. Then $a_{n+1} \equiv a_n \bmod p^n$, so $f'(a_{n+1}) \equiv f'(a_n) \bmod p^n$. Reducing mod p gives $f'(a_{n+1}) \equiv f'(a_n) \not\equiv 0 \bmod p$. So (1) holds for a_{n+1}. To show (2), let $d = \max(\deg f, 2)$ and consider the Taylor expansion of f about a_n:

$$f(x) = f_d(x - a_n)^d + f_{d-1}(x - a_n)^{d-1} + \cdots + f_1(x - a_n) + f_0.$$

Reversing the order of the terms and noting that $f_0 = f(a_n)$ and $f_1 = f'(a_n)$ we can write

$$f(x) = f(a_n) + f'(a_n)(x - a_n) + (x - a_n)^2 g(x),$$

for some $g \in \mathbb{Z}_p[x]$. Substituting a_{n+1} for x yields

$$f(a_{n+1}) = f(a_n) + f'(a_n)(a_{n+1} - a_n) + (a_{n+1} - a_n)^2 g(a_{n+1}).$$

From the definition of a_{n+1} we have $f'(a_n)(a_{n+1} - a_n) = -f(a_n)$, thus

$$f(a_{n+1}) = (a_{n+1} - a_n)^2 g(a_{n+1}).$$

As noted above, $a_{n+1} \equiv a_n \bmod p^n$, so $f(a_{n+1}) \equiv 0 \bmod p^{2n}$. Since $2n \geq n + 1$, we have $f(a_{n+1}) \equiv 0 \bmod p^{n+1}$, so (2) holds for a_{n+1}.

For uniqueness we argue that each a_{n+1} (and therefore b) is congruent modulo p^{n+1} to the *unique* root of $f \bmod p^{n+1}$ that is congruent to $a_n \bmod p^n$. There can be only one such root because a_n is a *simple* root of $f \bmod p^n$, since (1) implies $f'(a_n) \not\equiv 0 \bmod p^n$. □

There are stronger version of Hensel's lemma than we have given here. In particular, the hypothesis $f'(a) \not\equiv 0 \bmod p$ can be weakened so that the lemma can be applied even in situations where a is not a simple root. Additionally, the sequence (a_n) actualy converges to a root of f more rapidly than indicated by inductive hypothesis (2). You will prove stronger and more effective versions of Hensel's lemma on the problem set, as well as exploring several applications.

18.782 Introduction to Arithmetic Geometry **Fall 2013**
Lecture #9 **10/3/2013**

9.1 Quadratic forms

We assume throughout k is a field of characteristic different from 2.

Definition 9.1. The four equivalent definitions below all define a *quadratic form* on k.

1. A *homogeneous quadratic polynomial* $f \in k[x_1, \ldots, x_n]$.
2. Associated to f is a *symmetric matrix* $A \in k^{n\times n}$ whose entries (a_{ij}) correspond to the coefficients of x_ix_j in f via $f(x_1,\ldots,x_n) = \sum_{i,j} a_{ij}x^ix^j$.[1] Conversely, every symmetric matrix defines a homogeneous quadratic polynomial.
3. Each symmetric matrix A defines a *symmetric bilinear form* $B\colon k^n \times k^n \to k$ via $B_f(x,y) = x^tAy$, where x and $y$denote column vectors. It is symmetric, since
$$B(x,y) = x^tAy = (x^tAy)^t = y^tA^tx = y^tAx = B(y,x),$$
and it is bilinear, since for any $a \in k$ and $x,y,z \in k^n$ we have
$$B(ax+y,z) = (ax+y)^tAz = (ax^t + y^t)Az = ax^tAz + y^tAz = aB_f(x,z) + B(y,z).$$
Conversely, if B is a symmetric bilinear form, and $e_1,\ldots,e_n$ are basis vectors, the matrix $A = (a_{ij})$ defined by $a_{ij} = B(e_i,e_j)$ is symmetric.
4. The function $f\colon k^n \to k$ obtained by evaluating a homogeneous quadratic polynomial is a *homogeneous quadratic function.* In terms of the corresponding bilinear form $B(x,y)$, we have $f(x) = B(x,x)$. Conversely, we can recover $B(x,y)$ from $f(x)$ via
$$B(x,y) = \frac{f(x+y) - f(x) - f(y)}{2}.$$

We thus have canonical isomorphisms between four sets of objects: homogeneous quadratic polynomials, symmetric matrices, symmetric bilinear forms, and homogeneous quadratic functions. We use the symbol f to refer to both a homogeneous quadratic polynomial and its evaluation function, and we use the symbols A and B to refer to the associated matrix and bilinear form.

The definition of a symmetric bilinear form $B\colon V \times V \to k$ makes sense over any finite dimensional k-vector space V, and we can define the corresponding homogeneous function $f\colon V \to k$ abstractly as $f(v) = B(v,v)$. If we then choose a basis for V we can compute the symmetric matrix A whose coefficients define a homogeneous quadratic polynomial.

Symmetric bilinear forms can be viewed as a generalization of inner products to arbitrary fields. Inner products are also required to satisfy $B(v,v) > 0$ for any nonzero vector v, but this only makes sense if k is an *ordered field.*[2] In general, symmetric bilinear forms are allowed to vanish on nonzero vectors (indeed, the zero map is a symmetric bilinear form).

[1] Note that for $i \neq j$ this means that if f_{ij} is the coefficient of x_ix_j then $a_{ij} = a_{ji} = f_{ij}/2$, so that $f_{ij}x_ix_j = a_{ij}x_ix_j + a_{ji}x_jx_i$. This is slightly unpleasant but makes everything else work nicely.

[2] An ordered field is a field with a total ordering $\leq$ that satisfies $a \leq b \Rightarrow a{+}c \leq b{+}c$ and $a,b > 0 \Rightarrow ab > 0$. In such a field 0 cannot be written as a sum of nonzero squares. This is a severe restriction; it rules out all fields of positive characterstic, all p-adic fields, the complex numbers, and most number fields.

The group $\mathrm{GL}_n(k)$ of invertible $n \times n$ matrices over k acts on the space of quadratic forms as a linear change of variables. If T is any invertible linear transformation on V, and A is the matrix of a quadratic form f on V, then we have

$$f(Tv) = (Tv)^t A(Tv) = v^t(T^t AT)v$$

where $T^t AT$ is a symmetric matrix that defines another quadratic form.

Definition 9.2. Two quadratic forms f and g are *equivalent* if $g(v) = f(Tv)$ for some $T \in \mathrm{GL}_n(k)$. This defines an equivalence relation on the set of all quadratic forms of the same dimension over the field k.

Note that, in general, the matrices $T^t AT$ and $T^{-1}AT$ are not the same, this $\mathrm{GL}_n(k)$ action is not the same as its action by conjugation. In particular, equivalent symmetric matrices need not be similar, as can be seen by the fact that equivalent matrices may have different determinants:

$$\det(T^t AT) = \det(T^t)\det(A)\det(T) = \det(T)^2\det(A).$$

Definition 9.3. The *rank* of a quadratic form is the rank of its matrix; rank is clearly preserved under equivalence. A quadratic form is *non-degenerate* if it has full rank, equivalently, the determinant of its matrix is nonzero.

If B is the symmetric bilinear form associated to a non-degenerate quadratic form on V, then each nonzero $v \in V$ defines a nonzero linear map $w \to B(v, w)$ (otherwise the matrix of the form with respect to a basis including v would have a zero row).

Definition 9.4. The *discriminant* of a nondegenerate quadratic form with matrix A is the image of $\det A$ in $k^\times/k^{\times 2}$; it is clearly preserved by equivalence.

Inequivalent forms may have the same discriminant; over $\mathbb{C}$ for example, every non-degenerate form has the same discriminant (in fact all nondegenerate forms of the same dimension are equivalent). However, quadratic forms with different discriminants cannot be equivalent; this implies that over $\mathbb{Q}$, for example, there are infinitely many distinct equivalence classes of quadratic forms in every dimension $n > 0$.

A quadratic form is said to be *diagonal* if its matrix is diagonal.

Theorem 9.5. *Every quadratic form is equivalent to a diagonal quadratic form.*

Proof. We proceed by induction on the dimension n. The base cases $n \leq 1$ are trivial. Let f be a quadratic form on a vector space V, and let B be the corresponding symmetric bilinear form. If f is the zero function then its matrix is zero, hence diagonal, so assume otherwise and pick $v \in k^n$ so that $f(v) \neq 0$. The map $x \to B(x, v)$ is a nonzero linear map from k^n to k, hence surjective, so its kernel $v^\perp = \{x \in V : B(x, v) = 0\}$ has dimension $n - 1$. We know that $v \notin v^\perp$, since $B(v, v) = f(v) \neq 0$, so $V \simeq \langle v \rangle \oplus v^\perp$. Thus any $y \in V$ can be written as $y = y_1 + y_2$ with $y_1 \in \langle v \rangle$ and $y_2 \in v^\perp$. We then have

$$f(y_1 + y_2) = B(y_1 + y_2, y_1 + y_2) = B(y_1, y_1) + B(y_2, y_2) + 2B(y_1, y_2) = f(y_1) + f(y_2),$$

since $B(y_1, y_2) = 0$ for any $y_1 \in \langle v \rangle$ and $y_2 \in v^\perp$. By the inductive hypothesis, the restriction $f|_{v^\perp}$ of f to $v^\perp$ can be diagonalized (that is, there is a diagonal quadratic form on $v^\perp$ that is equivalent to $f|_{v^\perp}$), and the same is certainly true for the restriction of f to the 1-dimensional subspace $\langle v \rangle$, thus f can be diagonalized. □

So to understand equivalence classes of quadratic forms we can restrict our attention to diagonal quadratic forms.

Example 9.6. The quadratic form $x^2 + y^2$ is equivalent to $2x^2 + 2y^2$, since

$$(x + y)^2 + (x - y)^2 = 2x^2 + 2y^2,$$

but it is not equivalent to $3x^2 + 3y^2$. Indeed, for $x^2 + y^2$ to be equivalent to $\alpha x^2 + \beta y^2$ we must have

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^t \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a^2 + c^2 & ab + cd \\ ab + cd & b^2 + d^2 \end{pmatrix} = \begin{pmatrix} \alpha & 0 \\ 0 & \beta \end{pmatrix},$$

and in particular, α and β must both be sums of squares, which 3 is not.

Thus equivalence of quadratic forms depends on arithmetic properties of the field k.

Definition 9.7. A quadratic form f on V *represents* $a \in k$ if a lies in the image of $f\colon V \to k$. Equivalent forms necessarily represent the same elements (but the converse need not hold).

Example 9.8. The form $x^2 - 2y^2$ represents -7 but not 0.

The constraint that $x \neq 0$ is critical, otherwise every quadratic form would represent 0; the quadratic forms that represent 0 are of particular interest to us.

Theorem 9.9. *If a nondegenerate quadratic form f represents* 0 *then it represents every element of k.*

Proof. Assume $f(v) = 0$ for some $v \in V$. Since f is nondegenerate, there exists $w \in V$ with $B(v, w) \neq 0$, and v and w must be independent, since $B(v, v) = f(v) = 0$ and therefore $B(v, xv) = cB(v, v) = 0$ for any $c \in k$. For any $x \in k$ we have

$$f(xv + w) = B(xv + w, xv + w) = B(v, v)x^2 + 2B(v, w)x + B(w, w) = ax + b,$$

with $a = 2B(v, w) \neq 0$ and $b = f(w)$. For any $c \in k$ we can solve $ax + b = c$ for x, proving that f represents $c = f(xv + w)$. □

Our main goal is to prove the following theorem of Minkowski, which was generalized to number fields by Hasse.

Theorem 9.10 (Hasse-Minkowski)**.** *A quadratic form over $\mathbb{Q}$ represents* 0 *if and only if it represents* 0 *over every completion of $\mathbb{Q}$, that is, over $\mathbb{Q}_p$ for all primes $p \leq \infty$.*

This is an example of a *local-global* principle. We have an object f (in this case a quadratic form) defined over a "global" field (in this case $\mathbb{Q}$) and a certain property of interest (in this case representing 0). Since f is defined over the global field, we can also consider f as an object over any of the "local fields" associated to the global field (in this case the completions of $\mathbb{Q}$). If f satisfies the property of interest over the global field then it typically must satisfy this property over every local field (this is certainly true in our case), but the question is whether the converse holds. In the case of quadratic forms representing 0, the answer is "yes", but in many other cases we will see later in this the answer is "no," and it is a major point of interest in arithmetic geometry to understand exactly when and how various local-global principles can fail.

18.782 Introduction to Arithmetic Geometry **Fall 2013**
Lecture #10 **10/8/2013**

In this lecture we lay the groundwork needed to prove the Hasse-Minkowski theorem for $\mathbb{Q}$, which states that a quadratic form over $\mathbb{Q}$ represents 0 if and only if it represents 0 over every completion of $\mathbb{Q}$ (as proved by Minkowski). The statement still holds if $\mathbb{Q}$ is replaced by any number field (as proved by Hasse), but we will restrict our attention to $\mathbb{Q}$.

Unless otherwise indicated, we use p througout to denote any prime of $\mathbb{Q}$, including the archimedean prime $p = \infty$. We begin by defining the Hilbert symbol for p.

10.1 The Hilbert symbol

Definition 10.1. For $a, b \in \mathbb{Q}_p^\times$ the *Hilbert symbol* $(a, b)_p$ is defined by

$$(a,b)_p = \begin{cases} 1 & ax^2 + by^2 = 1 \text{ has a solution in } \mathbb{Q}_p, \\ -1 & \text{otherwise.} \end{cases}$$

It is clear from the definition that the Hilbert symbol is symmetric, and that it only depends on the images of a and b in $\mathbb{Q}_p^\times/\mathbb{Q}_p^{\times 2}$ (their *square classes*). We note that

$$\mathbb{Q}_p^\times/\mathbb{Q}_p^{\times 2} \simeq \begin{cases} \simeq \mathbb{Z}/2\mathbb{Z} & \text{if } p = \infty, \\ \simeq (\mathbb{Z}/2\mathbb{Z})^2 & \text{if } p \text{ is odd}, \\ \simeq (\mathbb{Z}/2\mathbb{Z})^3 & \text{if } p = 2. \end{cases}$$

The case $p = \infty$ is clear, since $\mathbb{R}^\times = \mathbb{Q}_\infty^\times$ has just two square classes (positive and negative numbers), and the cases with $p < \infty$ were proved in Problem Set 4. Thus the Hilbert symbol can be viewed as a map $(\mathbb{Q}^\times/\mathbb{Q}^{\times 2}) \times (\mathbb{Q}^\times/\mathbb{Q}^{\times 2}) \to \{\pm 1\}$ of finite sets.

We say that a solution $(x_0, \ldots, x_n)$ to a homogeneous polynomial equation over $\mathbb{Q}_p$ is *primitive* if all of its elements lie in $\mathbb{Z}_p$ and at least one lies in $\mathbb{Z}_p^\times$. The following lemma gives several equivalent definitions of the Hilbert symbol.

Lemma 10.2. *For any $a, b \in \mathbb{Q}_p^\times$, the following are equivalent:*

(i) $(a, b)_p = 1$.

(ii) *The quadratic form $z^2 - ax^2 - by^2$ represents* 0.

(iii) *The equation $ax^2 + by^2 = z^2$ has a primitive solution.*

(iv) *$a \in \mathbb{Q}_p$ is the norm of an element in $\mathbb{Q}_p(\sqrt{b})$.*

Proof. (i)⇒(ii) is immediate (let $z = 1$). The reverse implication is clear if $z^2 - ax^2 - by^2 = 0$ represents 0 with z nonzero (divide by z^2), and otherwise the non-degenerate quadratic form $ax^2 + by^2$ represents 0, hence it represents every element of $\mathbb{Q}_p$ including 1, so (ii)⇒(i).

To show (ii)⇒(iii), multiply through by p^r, for a suitable integer r, and rearrange terms. The reverse implication (iii)⇒(ii) is immediate.

If b is square then $\mathbb{Q}_p(\sqrt{b}) = \mathbb{Q}_p$ and $N(a) = a$ so (iv) holds, and the form $z^2 - by^2$ represents 0, hence every element of $\mathbb{Q}_p$ including ax_0^2 for any x_0, so (ii) holds. If b is not square then $N(z + y\sqrt{b}) = z^2 - by^2$. If a is a norm in $\mathbb{Q}(\sqrt{b})$ then $z^2 - ax^2 - by^2$ represents 0 (set $x = 1$), and if $z^2 - ax^2 - by^2$ represents 0 then dividing by x^2 and adding a to both sides shows that a is a norm. So (ii)⇔(iv). □

Andrew V. Sutherland

Corollary 10.3. *For all $a, b, c \in \mathbb{Q}_p^\times$, the following hold:*

(i) $(1, c)_p = 1$.

(ii) $(-c, c)_p = 1$.

(iii) $(a, c)_p = 1 \implies (a, c)_p(b, c)_p = (ab, c)_p$.

(iv) $(c, c)_p = (-1, c)_p$.

Proof. Let N denote the norm map from $\mathbb{Q}_p(\sqrt{c})$ to $\mathbb{Q}_p$. For (i) we have $N(1) = 1$. For (ii), $-c = N(-c)$ for $c \in \mathbb{Q}^{\times 2}$ and $-c = N(\sqrt{c})$ otherwise. For (iii), If a and b are both norms in $\mathbb{Q}(\sqrt{c})$, then so is ab, by the multiplicativity of the norm map; conversely, if a and ab are both norms, so is $1/a$, as is $(1/a)ab = b$. Thus if $(a, c)_p = 1$, then $(b, c)_p = 1$ if and only if $(ab, c)_p = 1$, which implies $(a, c)_p(b, c)_p = (ab, c)_p$. For (iv), $(-c, c)_p = 1$ by (ii), so by (iii) we have $(c, c)_p = (-c, c)_p(c, c)_p = (-c^2, c)_p = (-1, c)_p$. □

Theorem 10.4. $(a, b)_\infty = -1$ *if and only if* $a, b < 0$

Proof. We can assume $a, b \in \{\pm 1\}$, since $\{\pm 1\}$ is a complete set of representatives for $\mathbb{R}^\times/\mathbb{R}^{\times 2}$. If either a or b is 1 then $(a, b)_\infty = 1$, by Corollary 10.3.(i), and $(-1, -1)_\infty = -1$, since -1 is not a norm in $\mathbb{C} = \mathbb{Q}_\infty(\sqrt{-1})$. □

Lemma 10.5. *If p is odd, then $(u, v)_p = 1$ for all $u, v \in \mathbb{Z}_p^\times$.*

Proof. Recall from Lecture 3 (or the Chevalley-Warning theorem on problem set 2) that every plane projective conic over $\mathbb{F}_p$ has a rational point, so we can find a non-trivial solution to $z^2 - ux^2 - vy^2 = 0$ modulo p. If we then fix two of x, y, z so that the third is nonzero, Hensel's lemma gives a solution over $\mathbb{Z}_p$. □

Remark 10.6. Lemma 10.5 does not hold for $p = 2$; for example, $(3, 3)_2 = -1$.

Theorem 10.7. *Let p be an odd prime, and write $a, b \in \mathbb{Q}_p^\times$ as $a = p^\alpha u$ and $b = p^\beta v$, with $\alpha, \beta \in \mathbb{Z}$ and $u, v \in \mathbb{Z}_p^\times$. Then*

$$(a, b)_p = (-1)^{\alpha\beta\frac{p-1}{2}} \left(\frac{u}{p}\right)^\beta \left(\frac{v}{p}\right)^\alpha,$$

where $(\frac{x}{p})$ denotes the Legendre symbol $(\frac{x \bmod p}{p})$.

Proof. Since $(a, b)_p$ depends only on the square classes of a and b, we assume $\alpha, \beta \in \{0, 1\}$.

Case $\alpha = 0, \beta = 0$: We have $(u, v)_p = 1$, by Lemma 10.5, which agrees with the formula.

Case $\alpha = 1, \beta = 0$: We need to show that $(pu, v)_p = (\frac{v}{p})$. Since $(u^{-1}, v)_p = 1$, we have $(pu, v)_p = (pu, v)_p(u^{-1}, v)_p = (p, v)_p$, by Corollary 10.3.(iii). If v is a square then we have $(p, v)_p = (p, 1)_p = (1, p)_p = 1 = (\frac{v}{p})$. If v is not a square then $z^2 - px^2 - vy^2 = 0$ has no non-trivial solutions modulo p, hence no primitive solutions. This implies $(p, v)_p = -1 = (\frac{v}{p})$.

Case $\alpha = 1, \beta = 1$: We must show $(pu, pv)_p = (-1)^{\frac{p-1}{2}} \left(\frac{u}{p}\right) \left(\frac{v}{p}\right)$. Applying Corollary 10.3 we have

$$(pu, pv)_p = (pu, pv)_p(-pv, pv)_p = (-p^2uv, pv)_p = (-uv, pv)_p = (pv, -uv)_p$$

Applying the formula in the case $\alpha = 1, \beta = 0$ already proved, we have

$$(pv, -uv)_p = \left(\frac{-uv}{p}\right) = \left(\frac{-1}{p}\right)\left(\frac{u}{p}\right)\left(\frac{v}{p}\right) = (-1)^{\frac{p-1}{2}}\left(\frac{u}{p}\right)\left(\frac{v}{p}\right).$$ □

Lemma 10.8. *Let $u, v \in \mathbb{Z}_2^\times$. The equations $z^2 - ux^2 - vy^2 = 0$ and $z^2 - 2ux^2 - vy^2 = 0$ have primitive solutions over $\mathbb{Z}_2$ if and only if they have primitive solutions modulo 8.*

Proof. Without loss of generality we can assume that u and v are odd integers, since every square class in $\mathbb{Z}_2^\times/\mathbb{Z}_2^{\times 2}$ is represented by an odd integer (in fact one can assume $u, v \in \{\pm 1, \pm 5\}$) The necessity of having a primitive solution modulo 8 is clear. To prove sufficiency we apply the strong form of Hensel's lemma proved in Problem Set 4. In both cases, if we have a non-trivial solution (x_0, y_0, z_0) modulo 8 we can fix two of x_0, y_0, z_0 to obtain a quadratic polynomial $f(w)$ over $\mathbb{Z}_2$ and $w_0 \in \mathbb{Z}_2^\times$ that satisfies $v_2(f(w_0)) = 3 > 2 = 2v_2(f'(w_0))$. In the case of the second equation, note that a primitive solution (x_0, y_0, z_0) modulo 8 must have y_0 or z_0 odd; if not, then z_0^2 and vy_0^2, and therefore $2ux_0^2$, are divisible by 4, but this means x_0 is also divisible by 2, which contradicts the primitivity of (x_0, y_0, z_0). Lifting w_0 to a root of $f(w)$ over $\mathbb{Z}_2$ yields a solution to the original equation. □

Theorem 10.9. *Write $a, b \in \mathbb{Q}_2^\times$ as $a = 2^\alpha u$ and $b = 2^\beta v$ with $\alpha, \beta \in \mathbb{Z}$ and $u, v \in \mathbb{Z}_2^\times$. Then*

$$(a, b)_2 = (-1)^{\epsilon(u)\epsilon(v) + \alpha\omega(v) + \beta\omega(u)},$$

where $\epsilon(u)$ and $\omega(u)$ denote the images in $\mathbb{Z}/2\mathbb{Z}$ of $(u-1)/2$ and $(u^2-1)/8$, respectively.

Proof. Since $(a, b)_2$ only depends on the square classes of a and b, It suffices to verify the formula for $a, b \in S$, where $S = \{\pm 1, \pm 3, \pm 2, \pm 6\}$ is a complete set of representatives for $\mathbb{Q}_2^\times/\mathbb{Q}_2^{\times 2}$. As in the proof of Theorem 10.7, we can use $(pu, pv)_2 = (pv, -uv)_2$ to reduce to the case where one of a, b lies in $\mathbb{Z}_p^\times$. By Lemma 10.8, to compute $(a, b)_2$ with one of a, b in $\mathbb{Z}_2^\times$, it suffices to check for primitive solutions to $z^2 - ax^2 - by^2 = 0$ modulo 8, which reduces the problem to a finite verification which performed by Sage worksheet. □

We now note the following corollary to Theorems 10.4, 10.7, and 10.9.

Corollary 10.10. *The Hilbert symbol $(a, b)_p$ is a nondegenerate bilinear map. This means that for all $a, b, c \in \mathbb{Q}_p^\times$ we have*

$$(a, c)_p(b, c)_p = (ab, c) \qquad \text{and} \qquad (a, b)_p(a, c)_p = (a, bc)_p,$$

and that for every non-square c we have $(b, c)_p = -1$ for some b.

Proof. Both statements are clear for $p = \infty$ (there are only 2 square classes and 4 combinations to check). For p odd, let $c = p^\gamma w$ and fix $\varepsilon = (-1)^{\gamma\frac{p-1}{2}}$. Then for $a = p^\alpha u$ and $b = p^\beta v$, we have

$$\begin{aligned} (a, c)_p(b, c)_p &= \varepsilon^\alpha \left(\frac{u}{p}\right)^\gamma \left(\frac{w}{p}\right)^\alpha \varepsilon^\beta \left(\frac{v}{p}\right)^\gamma \left(\frac{w}{p}\right)^\beta \\ &= \varepsilon^{\alpha+\beta} \left(\frac{uv}{p}\right)^\gamma \left(\frac{w}{p}\right)^{\alpha+\beta} \\ &= (ab, c)_p. \end{aligned}$$

To verify non-degeneracy, we note that if c is not square than either $\gamma = 1$ or $(\frac{w}{p}) = -1$. If $\gamma = 1$ we can choose $b = v$ with $(\frac{v}{p}) = -1$, so that $(b, c)_p = (\frac{v}{p})^\gamma = -1$. If $\gamma = 0$, then $\varepsilon = 1$ and $(\frac{w}{p}) = -1$, so with$b = p$ we have $(b, c)_p = (\frac{w}{p}) = -1$.

For $p = 2$, we have

$$\begin{aligned}(a,c)_2(b,c)_2 &= (-1)^{\epsilon(u)\epsilon(w)+\alpha\omega(w)+\gamma\omega(u)}(-1)^{\epsilon(v)\epsilon(w)+\beta\omega(w)+\gamma\omega(v)}\\ &= (-1)^{(\epsilon(u)+\epsilon(v))\epsilon(w)+(\alpha+\beta)\omega(w)+\gamma(\omega(u)+\omega(v))}\\ &= (-1)^{\epsilon(uv)\epsilon(w)+(\alpha+\beta)\omega(w)+\gamma\omega(uv)}\\ &= (ab,c)_2,\end{aligned}$$

where we have used the fact that ϵ and ω are group homomorphisms from $\mathbb{Z}_2^\times$ to $\mathbb{Z}/2\mathbb{Z}$. To see this, note that the image of $\epsilon^{-1}(0)$ in $(\mathbb{Z}/4\mathbb{Z})^\times$ is $\{1\}$, a subgroup of index 2, and the image of $\omega^{-1}(0)$ in $(\mathbb{Z}/8\mathbb{Z})^\times$ is $\{\pm 1\}$, which is again a subgroup of index 2.

We now verify non-degeneracy for $p = 2$. If c is not square then either $\gamma = 1$, or one of $\epsilon(w)$ and $\omega(w)$ is nonzero. If $\gamma = 1$, then $(5,c)_2 = -1$. If $\gamma = 0$ and $\omega(w) = 1$, then $(2,c)_2 = -1$. If $\gamma = 0$ and $\omega(w) = 0$, then we must have $\epsilon(w) = 1$, so $(-1,c)_2 = -1$. □

We now prove Hilbert's reciprocity law, which may be regarded as a generalization of quadratic reciprocity.

Theorem 10.11. *Let $a, b \in \mathbb{Q}^\times$. Then $(a,b)_p = 1$ for all but finitely many primes p and*

$$\prod_p (a,b)_p = 1.$$

Proof. We can assume without loss of generality that $a, b \in \mathbb{Z}$, since multiplying each of a and b by the square of its denominator will not change $(a,b)_p$ for any p. The theorem holds if either a or b is 1, and by the bilinearity of the Hilbert symbol, we can assume that

$$a, b \in \{-1\} \cup \{q \in \mathbb{Z}_{>0} : q \text{ is prime}\}.$$

The first statement of the theorem is clear, since $a, b \in \mathbb{Z}_p^\times$ for $p < \infty$ not equal to a or b, and $(u,v)_p = 1$ for all $u, v \in \mathbb{Z}_p^\times$ when p is odd, by Lemma 10.5. To verify the product formula, we consider 5 cases.

Case 1: $a = b = -1$. Then $(-1,-1)_\infty = (-1,-1)_2 = -1$ and $(-1,-1)_p = 1$ for p odd.

Case 2: $a = -1$ and b is prime. If $b = 2$ then $(1,1)$ is a solution to $-x^2 + 2y^2 = 1$ over $\mathbb{Q}_p$ for all p, thus $\prod_p(-1,2) = 1$. If b is odd, then $(-1,b)_p = 1$ for $p \notin \{2,b\}$, while $(-1,b)_2 = (-1)^{\epsilon(b)}$ and $(-1,b)_b = (\frac{-1}{b})$, both of which are equal to $(-1)^{(b-1)/2}$.

Case 3: a and b are the same prime. Then by Corollary 10.3, $(b,b)_p = (-1,b)_p$ for all primes p, and we are in case 2.

Case 4: $a = 2$ and b is an odd prime. Then $(2,b)_p = 1$ for all $p \notin \{2,b\}$, while $(2,b)_2 = (-1)^{\omega(b)}$ and $(2,b)_b = (\frac{2}{p})$, both of which are equal to $(-1)^{(b^2-1)/8}$.

Case 5: a and b are distinct odd primes. Then $(a,b)_p = 1$ for all $p \notin \{2,a,b\}$, while

$$(a,b)_p = \begin{cases} (-1)^{\epsilon(a)\epsilon(b)} & \text{if } p = 2,\\ \left(\frac{a}{b}\right) & \text{if } p = b,\\ \left(\frac{b}{a}\right) & \text{if } p = a.\end{cases}$$

Since $\epsilon(x) = (x-1)/2 \bmod 2$, we have

$$\prod_p (a,b)_p = (-1)^{\frac{a-1}{2}\frac{b-1}{2}}\left(\frac{a}{b}\right)\left(\frac{b}{a}\right) = 1,$$

by quadratic reciprocity. □

18.782 Introduction to Arithmetic Geometry **Fall 2013**
Lecture #11 **10/10/2013**

11.1 Quadratic forms over $\mathbb{Q}_p$

The Hasse-Minkowski theorem reduces the problem of determining whether a quadratic form f over $\mathbb{Q}$ represents 0 to the problem of determining whether f represents zero over $\mathbb{Q}_p$ for all $p \le \infty$. At first glance this might not seem like progress, since there are infinitely many p to check, but in fact we only need to check $p = 2$, $p = \infty$ and a finite set of odd primes.

Theorem 11.1. *Let p be an odd prime and let f be a diagonal quadratic form of dimension $n > 2$ with coefficients $a_1, \ldots, a_n \in \mathbb{Z}_p^\times$. Then f represents 0 over $\mathbb{Q}_p$.*

Proof. The equation $f(x_1, \ldots, x_n) \equiv 0 \bmod p$ is a homogeneous equation of degree 2 in $n > 2$ variables over $\mathbb{F}_p$. It follows from the Chevalley-Warning theorem that it has a non-trivial solution $(y_1, \ldots, y_n)$ over $\mathbb{F}_p \simeq \mathbb{Z}/p\mathbb{Z}$. Assume without loss of generality that $y_1 \ne 0$ and let $g(z)$ be the univariate polynomial $g(y) = f(y, y_2, \ldots, y_n)$ over $\mathbb{Z}_p$. Then $g(y_1) \equiv 0 \bmod p$ and $g'(y_1) = 2a_1y_1 \not\equiv 0 \bmod p$, so by Hensel's lemma there is a root z_1 of $g(y)$ over $\mathbb{Z}_p$. We then have $f(z_1, y_2, \ldots, y_n) = g(z_1) = 0$, so f represents 0 over $\mathbb{Q}_p$. □

Corollary 11.2. *Every quadratic form of dimension $n > 2$ over $\mathbb{Q}$ represents 0 over $\mathbb{Q}_p$ for all but finitely many primes p.*

Proof. In diagonal form the coefficients $a_1, \ldots, a_n$ lie in $\mathbb{Z}_p^\times$ for all odd $p \nmid a_1 \cdots a_n$. □

For quadratic forms of dimension $n \le 2$, we note that a nondegenerate unary form never represents 0, and the nondegenerate form $ax^2 + by^2$ represents 0 if and only if $-ab$ is square (this holds over any field). But when $-ab$ is not square it may still be the case that $ax^2 + by^2$ represents a given nonzero element t, and having a criterion for identifying such t will be useful in our proof of the Hasse-Minkowski theorem.

Lemma 11.3. *The nondegenerate quadratic form $ax^2 + by^2$ over $\mathbb{Q}_p$ represents $t \in \mathbb{Q}_p^*$ if and only if $(a, b)_p = (t, -ab)_p$.*

Proof. Since $t \ne 0$, the equation $ax^2 + by^2 = t$ has a non-trivial solution in $\mathbb{Q}_p$ if and only if $(a/t)x^2 + (b/t)y^2 = 1$ has a solution, which is equivalent to $(a/t, b/t)_p = 1$. We have

$$\begin{aligned}(a/t, b/t)_p &= (at, bt)_p = (a, bt)_p(t, bt)_p = (a, b)_p(a, t)_p(t, bt) = (a, b)_p(t, abt)_p\\ &= (a, b)_p(t, abt)_p(t, -t)_p = (a, b)_p(t, -ab)_p,\end{aligned}$$

where we have used that the Hilbert symbol is symmetric, bilinear, invariant on square classes, and satisfies $(x, -x)_p = 1$. Thus $(a/t, b/t)_p = 1$ if and only if $(a, b)_p(t, -ab)_p = 1$, which is equivalent to $(a, b)_p = (t, -ab)_p$ since both are ± 1. □

Corollary 11.4. *The nondegenerate form $ax^2 + by^2 + cz^2$ over $\mathbb{Q}_p$ represents 0 if and only if $(a, b)_p = (-c, -ab)_p$*

Proof. By the lemma, if suffices to show that $ax^2 + by^2 + cz^2$ represents 0 if and only if the binary form $ax^2 + by^2$ represents $-c$. The reverse implication is clear (set $z = 1$). For the forward implication, if $ax_0^2 + by_0^2 + cz_0^2 = 0$ then either $z_0 \ne 0$, in which case $a(x_0/z_0^2) + b(y_0/z_0)^2 = -c$ or $z_0 = 0$ in which case $ax^2 + by^2$ represents 0 and therefore every element of $\mathbb{Q}_p$, including $-c$. □

Andrew V. Sutherland

Corollary 11.5. *A ternary quadratic form over $\mathbb{Q}$ that represents 0 over all but at most one completion of $\mathbb{Q}$ represents 0 over every completion of $\mathbb{Q}$.*

Proof. The corollary is trivially true if the form is degenerate and otherwise it follows from the product formula for Hilbert symbols and the corollary above. □

11.2 Approximation

We now prove two *approximation theorems* that we will need to prove the Hasse-Minkowski theorem for $\mathbb{Q}$. These are quite general theorems that have many applications, but we will state them in a particularly simple form that suffices for our purposes here. Before proving them we first note/recall that $\mathbb{Q}$ is dense in $\mathbb{Q}_p$ and $\mathbb{Z}$ is dense in $\mathbb{Z}_p$.

Theorem 11.6. *Let $p \le \infty$ be any prime of $\mathbb{Q}$. Under the metric $d(x,y) = |x-y|_p$, the set $\mathbb{Q}$ is dense in $\mathbb{Q}_p$ and the set $\mathbb{Z}$ is dense in $\mathbb{Z}_p$.*

Proof. We know that $\mathbb{Q}_\infty = \mathbb{R}$ is the completion of $\mathbb{Q}$ and we proved that $\mathbb{Q}_p$ is (isomorphic to) the completion of $\mathbb{Q}$ for $p < \infty$, and any field is dense in its completion (this follows immediately from the definition). We note that the completion $\mathbb{Z}_\infty = \mathbb{Z}$ (any Cauchy sequence of integers must be eventually constant), and for $p < \infty$ the we can apply the fact that $\mathbb{Z}_p = \{x \in \mathbb{Q}_p : |x|_p \le 1\}$ and $\mathbb{Z} = \{x \in \mathbb{Q} : |x|_p \le 1\}$. □

Theorem 11.7 (Weak approximation)**.** *Let S be a finite set of primes $p \le \infty$, and for each $p \in S$ let $x_p \in \mathbb{Q}_p$ be given. Then for every $\epsilon > 0$ there exists $x \in \mathbb{Q}$ such that*

$$|x - x_p|_p < \epsilon$$

for all $p \in S$. Equivalently, the image of $\mathbb{Q}$ in $\prod_{p\in S} \mathbb{Q}_p$ dense under the product topology.

Proof. If S has cardinality 1 we can apply Theorem 11.6, so we assume S contains at least 2 primes. For any particular prime $p \in S$, we claim that there is a $y_p \in \mathbb{Q}$ such that $|y_p|_p > 1$ and $|y_p|_q < 1$ for $q \in S - \{p\}$. Indeed, let P be the product of the finite primes in S, and for each $p < \infty$ choose $r \in \mathbb{Z}_{>0}$ so that $p^{-r}P < 1$. Then define

$$y_p = \begin{cases} P & \text{if } p = \infty, \\ p^{-r}P & \text{otherwise.} \end{cases}$$

We now note that for any $q \in S$,

$$\lim_{n\to\infty} |y_p^n|_q = \begin{cases} \infty & \text{if } q = p, \\ 0 & \text{if } q \ne p. \end{cases}$$

It follows that for each $q \in S$

$$\lim_{n\to\infty} \frac{y_p^n}{1+y_p^n} = \begin{cases} 1 \text{ with respect to } |\ |_q \text{ for } q = p, \\ 0 \text{ with respect to } |\ |_q \text{ for } q \ne p, \end{cases}$$

since $\lim_{n\to\infty} |1 - y_p^n/(1+y_p^n)|_p = \lim_{n\to\infty} |1/(1+y_p^n)|_p = 0$ and $\lim_{n\to\infty} |y_p^n/(1+y_p^n)|_q = 0$ for $q \ne p$. For each $n \in \mathbb{Z}_{>0}$ define

$$z_n = \sum_{p\in S} \frac{x_p y_p^n}{1+y_p^n}.$$

Then $\lim_{n\to\infty} z_n = x_p$ with respect to $|\ |_p$ for each $p \in S$. So for any $\epsilon > 0$ there is an n for which $x = z_n$ satisfies $|x - x_p|_p < \epsilon$ for all $p \in S$. □

Theorem 11.8 (Strong approximation)**.** *Let S be a finite set of primes $p < \infty$, and for each $p \in S$ let $x_p \in \mathbb{Z}_p$ be given. Then for every $\epsilon > 0$ there exists $x \in \mathbb{Z}$ such that*

$$|x - x_p|_p < \epsilon$$

for all $p \in S$. Equivalently, the image of $\mathbb{Z}$ in $\prod_{p \in S} \mathbb{Z}_p$ is dense under the product topology.

Proof. Fix $\epsilon > 0$. By Theorem 11.6, for each x_p we can pick $y_p \in \mathbb{Z}_{\geq 0}$ so that $|y_p - x_p|_p < \epsilon$. Let n be a positive integer such that $p^n > y_p$ for all $p \in S$. By the Chinese remainder theorem, there exists $x \in \mathbb{Z}$ such that $x \equiv y_p \bmod p^n$ for all $p \in S$, and for this x we have $|x - x_p|_p < \epsilon$ for all $p \in S$. □

Remark 11.9. In more general settings it is natural to consider the infinite product of *all* the rings of p-adic integers

$$\hat{\mathbb{Z}} = \prod_{p<\infty} \mathbb{Z}_p.$$

Recall that for infinite products, the product topology is defined using a basis of open sets that consists of sequences (U_p), where each U_p is an open subset of $\mathbb{Z}_p$, and for all but finitely many p we have $U_p = \mathbb{Z}_p$. It follows from Theorem 11.8 that the image of $\mathbb{Z}$ in $\hat{\mathbb{Z}}$ is dense.

There is another way to define $\hat{\mathbb{Z}}$, which is to consider the inverse system of rings $(\mathbb{Z}/n\mathbb{Z})$, where n ranges over all positive integers n and we have reduction maps from $\mathbb{Z}/m\mathbb{Z}$ to $\mathbb{Z}/n\mathbb{Z}$ whenever $n|m$ (note that we now have an infinite acyclic graph of maps, not just a linear chain). The inverse limit

$$\hat{\mathbb{Z}} = \varprojlim \mathbb{Z}/n\mathbb{Z}$$

is called the *profinite completion* of $\mathbb{Z}$. One can show that these two definitions of $\hat{\mathbb{Z}}$ are canonically isomorphic. So a more pithy statement of Theorem 11.8 is that $\mathbb{Z}$ is dense in its profinite completion (this statement applies to profinite completions in general).

Remark 11.10. Note the difference between weak and strong approximation. With weak approximation we obtain a rational number x that is p-adically close to x_p for each p in a finite set S, but we have no control on $|x|_p$ for $p \notin S$. With strong approximation we obtain a rational number (in fact an integer) x that is p-adically close to x_p for each $p \in S$ and also satisfies $|x|_p \leq 1$ for all $p \notin S$, *except* the prime $p = \infty$; in order to apply the CRT we may need to make $|x|_\infty$ very large. More generally, we could allow $\infty \in S$ if we grant ourselves the freedom to make $|x|_{p_0}$ large for one prime $p_0 \notin S$; in this case x would be a rational number, not an integer, but its denominator would be divisible by no primes other than p_0, so that $x \in \mathbb{Z}_p$ for all $p \neq p_0$. This is characteristic of strong approximation theorems, we obtain an element whose absolute value is bounded at all but one prime.

The following lemma follows from the strong approximation theorem and Dirichlet's theorem on primes in arithmetic progressions: for any relative prime integers a and b there are infinitely many primes congruent to a mod b.

Lemma 11.11. *Let S be a finite set of primes $p \leq \infty$, and for each $p \in S$ let $x_p \in \mathbb{Q}_p^\times$ be given. Then there exists an $x \in \mathbb{Q}$ such that*

(i) *$x \in x_p \mathbb{Q}_p^{\times 2}$ for each $p \in S$.*

(ii) *$|x|_p = 1$ for all but at most one finite prime $p_0 \notin S$.*

Proof. Let $S_0 = S - \{\infty\}$, and define the rational number

$$y = \pm \prod_{p \in S_0} p^{v_p(x_p)},$$

where the sign of y is negative if $\infty \in S$ and $x_\infty < 0$, and positive otherwise. Then $|y|_p = |x_p|_p$ for all $p \in S_0$, and it follows that for each $p \in S_0$ we have $y = u_p x_p$ for some $u_p \in \mathbb{Z}_p^\times$. By the strong approximation theorem there exists an integer $z \equiv u_p \bmod p^{e_p}$, for all $p \in S_0$, where $e_p = 1$ for odd p and $e_p = 3$ for $p = 2$. It follows that $z \in u_p\mathbb{Q}_p^{\times 2}$ for all $p \in S_0$, since the square class of u_p depends only on its reduction mod p^{e_p}.

The integers z and $m = \prod_{p \in S_0} p^{e_p}$ are relatively prime, so it follows from Dirichlet's theorem that there are infinitely many primes congruent to $z \bmod m$. Let p_0 be the least such prime. Then $p_0 \in z\mathbb{Q}_p^{\times 2}$ for all $p \in S_0$, and $x = p_0 y$ satisfies both (i) and (ii). □

11.3 Proof of the Hasse-Minkowski theorem

Before proving the Hasse-Minkowski theorem for $\mathbb{Q}$ we make one final remark. The definition of the Hilbert symbol we gave in the last lecture makes sense over any field, in particular $\mathbb{Q}$, and the proofs of Lemma 10.2 and Corollary 10.3 still apply. In the proof below we use (a, b) to denote the Hilbert symbol of $a, b \in \mathbb{Q}^\times$.

Theorem 11.12 (Hasse-Minkowski). *A quadratic form over $\mathbb{Q}$ represents 0 if and only if it represents 0 over every completion of $\mathbb{Q}$.*

Proof. The forward implication is clear, we only need to prove the reverse implication. So let f be a quadratic form over $\mathbb{Q}$ that represents 0 over every completion of $\mathbb{Q}$. We may assume without loss of generality that f is a diagonal form $a_1x_1^2 + \cdots + a_nx_n^2$, which we may denote $\langle a_1, \ldots, a_n\rangle$. We write $\langle a_1, \ldots, a_n\rangle_p$ to denote the same form over $\mathbb{Q}_p$. If any $a_i = 0$, then f clearly represents 0 over $\mathbb{Q}$ (set $x_i = 1$ and $x_j = 0$ for $i \neq j$), so we assume f is nondegenerate and proceed by induction on its dimension n.

Case $n = 1$: The theorem holds trivially (f cannot represent 0 over any $\mathbb{Q}_p$).

Case $n = 2$: The form $\langle a, b\rangle_p$ represents 0 if and only if $-ab$ is square in $\mathbb{Q}_p$. Thus $v_p(-ab) \equiv 0 \bmod 2$ for all $p < \infty$ and $-ab > 0$. It follows that $-ab$ is square in $\mathbb{Q}$, and therefore $\langle a, b\rangle$ represents 0.

Case $n = 3$: Let $f(x, y, z) = z^2 - ax^2 - by^2$, where a and b are nonzero square-free integers with $|a| \le |b|$. We know $(a, b)_p = 1$ for all $p \le \infty$ and wish to show $(a, b) = 1$. We proceed by induction on $m = |a| + |b|$. The base case $m = 2$ has $a = \pm 1$ and $b = \pm 1$, in which case $(a, b)_\infty = 1$ implies that either a or b is 1 and therefore $(a, b) = 1$.

We now suppose $m \ge 3$, and that the result has been proven for all smaller m. For each prime $p|b$ there is a primitive solution $(x_0, y_0, z_0) \in \mathbb{Z}_p^3$ to $z^2 - ax^2 - by^2 = 0$. We must have $p|(z_0^2 - ax_0^2)$, since $p|b$, but we cannot have $p|x_0$ since then we would have $p|z_0$, contradicting primitivity. So $x_0 \in \mathbb{Z}_p^\times$ and $a = (z_0/x_0)^2$ is a square modulo p. This holds for every prime $p|b$, and b is square-free, so a is a square modulo b.

It follows that $a + bb' = t^2$ for some $t, b' \in \mathbb{Z}$ with $t \le |b/2|$. This implies $(a, bb') = 1$, since $bb' = t^2 - a$ is the norm of $t + \sqrt{a}$ in $\mathbb{Q}(\sqrt{a})$. Therefore

$$(a, b) = (a, b)(a, bb') = (a, b^2b') = (a, b').$$

We also have $(a, bb')_p = 1$, and therefore $(a, b')_p = (a, b)_p = 1$, for all $p \le \infty$. But

$$|b'| = \left|\frac{t^2 - a}{b}\right| \le \left|\frac{t^2}{b}\right| + \left|\frac{a}{b}\right| \le \frac{|b|}{4} + 1 < |b|,$$

so $|a|+|b'| < m$ and the inductive hypothesis implies $(a,b') = 1$. Thus $(a,b) = 1$, as desired.

Case $n = 4$: Let $f = \langle a_1, a_2, a_3, a_4\rangle$ and let S consist of the primes $p|2a_1a_2a_3a_4$ and ∞. Then $a_i \in \mathbb{Z}_p^\times$ for all $p \notin S$. For each $p \in S$ there exists $t_p \in \mathbb{Q}_p^\times$ such that $\langle a_1, a_2\rangle_p$ represents t_p and $\langle a_3, a_4\rangle_p$ represents $-t_p$ (we can assume $t_p \neq 0$: if 0 is represented, by both forms, so is every element of $\mathbb{Q}_p$). By Lemma 11.11, there is a rational number t and a prime $p_0 \notin S$ such that $t \in t_p\mathbb{Q}_p^{\times 2}$ for all $p \in S$ and $|t|_p = 1$ for all $p \notin S \cup \{p_0\}$.

The forms $\langle a_1, a_2, -t\rangle_p$ and $\langle a_3, a_4, t\rangle_p$ represent 0 for all $p \notin S \cup \{p_0\}$ because all such p are odd, and $a_i, \pm t \in \mathbb{Z}_p^\times$, so $(a_1, a_2)_p = 1 = (t, -a_1a_2)_p$ and $(a_3, a_4)_p = 1 = (-t, -a_3a_4)_p$, and we may apply Corollary 11.4. Since $t \in t_p\mathbb{Q}_p^{\times 2}$ for all $p \in S$, the forms $\langle a_1, a_2, -t\rangle_p$ and $\langle a_3, a_4, t\rangle_p$ also represent 0 for all $p \in S$. Thus $\langle a_1, a_2, -t\rangle_p$ and $\langle a_3, a_4, t\rangle_p$ represent 0 for all $p \neq p_0$, and by Corollary 11.5, also for $p = p_0$. By the inductive hypothesis $\langle a_1, a_2, -t\rangle$ and $\langle a_3, a_4, t\rangle$ both represent 0, therefore $\langle a_1, a_2, a_3, a_4\rangle$ represents 0.

Case $n \geq 5$: Let $f = \langle a_1, \ldots, a_n\rangle$. Let S be the set of primes for which $\langle a_3, \ldots, a_n\rangle_p$ does not represent 0. The set S is finite, by Corollary 11.2. If S is empty then $\langle a_3, \ldots, a_n\rangle$, and therefore f, represents 0, by the inductive hypothesis, so we assume S is not empty. For each $p \in S$ pick $t_p \in \mathbb{Q}_p^\times$ represented by $\langle a_1, a_2\rangle$, say $a_1x_p^2 + a_2y_p^2 = t_p$, such that $\langle a_3, \ldots, a_n\rangle_p$ represents $-t_p$ (such a t_p exists since f represents 0 over $\mathbb{Q}_p$ and, as above, we can always pick $t_p \neq 0$).

By the weak approximation theorem there exists $x, y \in \mathbb{Q}$ that are simultaneously close enough to all the $x_p, y_p \in \mathbb{Q}_p$ so that $t = a_1x^2 + a_2y^2$ is close enough to all the t_p to guarantee that $t \in t_p\mathbb{Q}_p^{\times 2}$ for all $p \in S$ (for $p < \infty$ the square class only depends on at most the first three nonzero p-adic digits, and over $\mathbb{R} = \mathbb{Q}_\infty$ we can ensure that x and y have the same signs as x_∞ and y_∞).[1] It follows that $\langle t, a_3, \ldots, a_n\rangle_p$ represents 0 for all $p \in S$, and since $\langle a_3, \ldots, a_n\rangle_p$ represents 0 for all $p \notin S$, so does $\langle t, a_3, \ldots, a_n\rangle_p$. Thus $\langle t, a_3, \ldots, a_n\rangle_p$ represents 0 for all p, and by the inductive hypothesis, $\langle t, a_3, \ldots, a_n\rangle$ represents 0. Therefore $\langle a_3, \ldots, a_n\rangle$ represents $-t = -a_1x^2 - a_2y^2$, hence $\langle a_1, \ldots, a_n\rangle$ represents 0. □

[1] Equivalently, the set of squares $\mathbb{Q}_p^{\times 2}$ is an open subset of $\mathbb{Q}_p^\times$, hence so is every square class $t_p\mathbb{Q}_p^{\times 2}$.

12.1 Field extensions

Before beginning our introduction to algebraic geometry we recall some standard facts about field extensions. Most of these should be familiar to you and can be found in any standard introductory algebra text, such as [1, 2]. We will occasionally need to results in slightly greater generality than you may have seen before, and here we may reference [3, 4].[1]

We start in the general setting of an arbitrary field extension L/k with no restrictions on k or L. The fields k and L necessarily have the same prime field (the subfield of k generated by the multiplicative identity), and therefore the same characteristic. The *degree* of the extension L/k, denoted $[L\colon k]$, is the dimension of L as a k-vector space, a not necessarily finite cardinal number. If have a tower of fields $k \subseteq L \subseteq M$, then

$$[M : k] = [M : L][L : k],$$

where the RHS is a product of cardinals.[2] When $[L : k]$ is finite we say that L/k is a *finite extension.*

An element $\alpha \in L$ is said to be *algebraic* over k if it is the root of a polynomial in $k[x]$, and otherwise it is *transendental* over k. The extension L/k is algebraic if every element of L is algebraic over k, and otherwise it is transcendental. If M/L and L/k are both algebraic extensions, so is M/k. A necessary and sufficient condition for L/k to be algebraic is that L be equal to the union of all finite extensions of k contained in L; in particular, every finite extension is algebraic.

The subset of L consisting of the elements that are algebraic over k forms a field called the *algebraic closure* of k in L. A field k is *algebraically closed* if every every non-constant polynomial in $k[x]$ has a root in k; equivalently, k has no non-trivial algebraic extensions. For every field k there exists an extension $\bar{k}/k$ with $\bar{k}$ algebraically closed; such a $\bar{k}$ is called an *algebraic closure* of k, and all such $\bar{k}$ are isomorphic (but this isomorphism is not unique in general). Any algebraic extension L/k can be embedded into any algebraic closure of k, since every algebraic closure of L is also an algebraic closure of k.

Remark 12.1. When working with algebraic extensions of k it is convenient to view them all as subfields of a some fixed algebraic closure $\bar{k}$ (there is in general no canonical choice). The key point is that we can always (not necessarily uniquely) embed any algebraic extension of L/k in our chosen $\bar{k}$, and if we have another extension M/L, our embedding of L into $\bar{k}$ can always be extended to an embedding of M into $\bar{k}$.

A set $S \subseteq L$ is said to be *algebraically independent* (over k) if for every finite subset $\{s_1, \ldots, s_n\}$ of S and every nonzero polynomial $f \in k[x_1, \ldots, x_n]$ we have

$$f(s_1, \ldots, s_n) \neq 0.$$

[1] With the exception of [1], which you should be familiar to you from 18.701/18.702, these references are all available online through the MIT library system (just click the title links in the references section at the end of these notes). I encourage you to consult them for further details on anything that is unfamiliar to you. One note of caution: when jumping into the middle of a textbook (or, especially, the results of a web search), be wary of assumptions that may have been stated much earlier (e.g. at the beginning of a chapter).

[2] Recall that a cardinal number is an equivalence class of equipotent sets (sets that can be put in bijection). The product of $n_1 = \#S_1$ and $n_2 = \#S_2$ is $n_1n_2 = \#(S_1 \times S_2)$ and the sum is the cardinality of the disjoint union: $n_1 + n_2 = \#(S_1 \sqcup S_2)$. But we shall be primarily interested in finite cardinals (natural numbers).

Note that this means the empty set is algebraically independent (just as the empty set is linearly independent in any vector space). An algebraically independent set $S \subseteq L$ for which $L/k(S)$ is algebraic is called a *transcendence basis* for the extension L/k.

Theorem 12.2. *Every transcendence basis for L/k has the same cardinality.*

Proof. We will only prove this in the case that L/k has a finite transcendence basis (which includes all extensions of interest to us); see [3, Theorem 7.9] for the general case. Let $S = \{s_1, \ldots, s_m\}$ be a smallest transcendence basis and let $T = \{t_1, \ldots, t_n\}$ be any other transcendence basis, with $n \geq m$. The set $\{t_1, s_1, \ldots, s_m\}$ must then algebraically dependent, since $t_1 \in L$ is algebraic over $k(S)$, and since t_1 is transcendental over k, some s_i, say s_1, must be algebraic over $k(t_1, s_2, \ldots, s_m)$. It follows that L is algebraic over $k(t_1, s_2, \ldots, s_m)$, and the set $T_1 = \{t_1, s_2, \ldots, s_m\}$ must be algebraically independent, otherwise it would contain a transcendence basis for L/k smaller than S. So T_1 is a transcendence basis for L/k of cardinality m that contains t_1.

Continuing in this fashion, for $i = 2, \ldots, m$ we can iteratively construct transcendence bases T_i of cardinality m that contain $\{t_1, \ldots, t_i\}$, until $T_m \subseteq T$ is a transcendence basis of cardinality m; but then we must have $T_m = T$, so $n = m$. □

Definition 12.3. The *transcendence degree* of a field extension L/K is the cardinality of any (hence every) transcendence basis for L/k.

Unlike extension degrees, which multiply in towers, transcendence degrees add in towers: for any fields $k \subseteq L \subseteq M$, the transcendence degree of M/k is the sum (as cardinals) of the transcendence degrees of M/L and L/k.

We say that the extension L/k is *purely transcendental* if $L = k(S)$ for some transcendence basis S for L/k. All purely transcendental extensions of k with the same transcendence degree are isomorphic. Every field extension L/k can be viewed as an algebraic extension of a purely transcendental extension: if S is a transcendence basis of L/k then $L/k(S)$ is an algebraic extension of the purely transcendental extension $k(S)/k$.

Remark 12.4. It is not the case that every field extension is a purely transcendental extension of an algebraic extension. Indeed, there are already plenty of counterexamples with transcendence degree 1, as we shall soon see.

The field extension L/k is said to be *simple* if $L = k(x)$ for some $x \in L$. A purely transcendental extension of transcendence degree 1 is obviously simple, but, less trivially, so is any finite separable extension (see below for the definition of separable); this is known as the primitive element theorem.

Remark 12.5. The notation $k(x)$ can be slightly confusing. If $x \in L$ is transcendental over k then $k(x)$ is isomorphic to the field of rational functions over k, in which case we may as well regard x as a variable. But if $x \in L$ is algebraic over k, then every rational expression $r(x)$ with nonzero denominator can be simplified to a polynomial in x of degree less than $n = [k(x) : k]$ by reducing modulo the minimal polynomial f of x (note that we can invert nonzero denominators modulo f); indeed, this follows from the fact that $\{1, x, \ldots, x^{n-1}\}$ is a basis for the n-dimensional k-vector space $k(x)$.

12.1.1 Algebraic extensions

We now assume that L/k is algebraic and fix $\bar{k}$ so that $L \in \bar{k}$. The extension L/k is *normal* if it satisfies either of the equivalent conditions:

- every irreducible polynomial in $k[x]$ with a root in L splits completely in L;
- $\sigma(L) = L$ for all $\sigma \in \operatorname{Aut}(\overline{k}/k)$ (every automorphism of $\overline{k}$ that fixes k also fixes L).[3]

Even if L/k is not normal, there is always an algebraic extension M/L for which M/k is normal. The minimal such extension is called the *normal closure* of L/k; it exists because intersections of normal extensions are normal. It is not true in general that if L/k and M/L are normal extensions then so is M/k, but if $k \subseteq L \subseteq M$ is a tower of fields with M/k normal, then M/L is normal (but L/k need not be).

A polynomial $f \in k[x]$ is *separable* if any of the following equivalent conditions hold:

- the factors of f in $\bar{k}[x]$ are all distinct;
- f and f' have no common root in $\bar{k}$;
- $\gcd(f, f') = 1$ in $k[x]$.

An element $\alpha \in L$ is separable over k if any of the following equivalent conditions hold:

- α is a root of a separable polynomial $f \in k[x]$;
- the minimal polynomial of α is separable;
- $\operatorname{char}(k) = 0$ or $\operatorname{char}(k) = p > 0$ and the minimal polynomial of α is not of the form $g(x^p)$ for some $g \in k[x]$.

The elements of L that are separable over k form a field called the *separable closure* of k in L. The separable closure of k in its algebraic closure $\bar{k}$ is denoted k^{sep} and is simply called the separable closure of k. If $k \subseteq L \subseteq M$ then M/k is separable if and only if both M/L and L/k are separable.

A field k is said to be *perfect* if any of the following equivalent conditions hold:

- $\operatorname{char}(k) = 0$ or $\operatorname{char}(k) = p > 0$ and $k = \{x^p : x \in k\}$ (k is fixed by Frobenius);
- every finite extension of k is separable over k;
- every algebraic extension of k is separable over k.

Note that finite fields and all fields of characteristic 0 are perfect.

Example 12.6. The rational function field $k = \mathbb{F}_p(t)$ is not perfect. If we consider the finite extension $L = k(t^{1/p})$ obtained by adjoining a pth root of t to k, the minimal polynomial of $t^{1/p}$ is $x^p - t$, which is irreducible over k but not separable (its derivative is 0).

An algebraic extension L/k is*Galois* if it is both normal and separable, and in this case we call $\operatorname{Gal}(L/k) = \operatorname{Aut}(L/k)$ the *Galois group* of L/k. The extension k^{sep}/k is always normal: if an irreducible polynomial $f \in k[x]$ has a root α in k^{sep}, then (up to scalars) f is the minimal polynomial of α over k, hence separable over k, so all its roots lie in k^{sep}. Thus k^{sep}/k is a Galois extension and its Galois group

$$G_k = \operatorname{Gal}(k^{\mathrm{sep}}/k)$$

[3] Some authors write $\operatorname{Gal}(L/k)$ for $\operatorname{Aut}(L/k)$, others only use $\operatorname{Gal}(L/k)$ when L/k is known to be Galois; we will use the later convention.

is the *absolute Galois group* of k (we could also define G_k as $\operatorname{Aut}(\bar{k}/k)$, the restriction map from $\operatorname{Aut}(\bar{k}/k)$ to $\operatorname{Gal}(k^{\text{sep}}/k)$ is always an isomorphism).

The *splitting field* of a polynomial $f \in k[x]$ is the extension of k obtained by adjoining all the roots of f (which lie in $\bar{k}$). Every splitting field is normal, and every finite normal extension of k is the splitting field of some polynomial over k; when k is a perfect field we can go further and say that L/k is a finite Galois extension if and only if it is the splitting field of some polynomial over k.

For finite Galois extensions M/k we always have $\#\operatorname{Gal}(M/k) = [M : k]$, and the fundamental theorem of Galois theory gives an inclusion-reversing bijection between subgroups $H \subseteq \operatorname{Gal}(M/k)$ and intermediate fields $k \subseteq L \subseteq M$ in which $L = M^H$ and $H = \operatorname{Gal}(M/L)$ (note that M/L is necessarily Galois). Beware that none of the statements in this paragraph necessarily applies to infinite Galois extensions, some modifications are required (this will be explored further on the next problem set).

12.2 Affine space

Let k be a perfect field and fix an algbebraic closure $\bar{k}$.

Definition 12.7. n-dimensional *affine space* over k is the set

$$\mathbb{A}^n_k = \{(x_1, \ldots, x_n) \in \bar{k}^n\},$$

equivalently $\mathbb{A}^n_k$ is the vector space $\bar{k}^n$ regarded as a set. When k is clear from context we may just write $\mathbb{A}^n$. If $k \subseteq L \subseteq \bar{k}$, the set of *$L$-rational points* (or just L-points) in $\mathbb{A}^n$ is

$$\mathbb{A}^n(L) = \{(x_1, \ldots, x_n) \in L^n\} = \mathbb{A}^n(\bar{k})^{G_L},$$

where $A^n(\bar{k})^{G_L}$ denotes the set of points in $\mathbb{A}^n(\bar{k})$ fixed by $G_L = \operatorname{Gal}(L^{\text{sep}}/L) = \operatorname{Gal}(\bar{k}/L)$. In particular, $\mathbb{A}^n(k) = \mathbb{A}^n(\bar{k})^{G_k}$.

Definition 12.8. If S is a set of polynomials in $A = \bar{k}[x_1, \ldots, x_n]$, the set of points

$$Z_S = \{P \in \mathbb{A}^n : f(P) = 0 \text{ for all } f \in S\},$$

is called an (affine) *algebraic set.* If $k \subseteq L \subseteq \bar{k}$, the set of L-rational points in Z_S is

$$Z_S(L) = Z_S \cap \mathbb{A}^n(L).$$

When S is a singleton $\{f\}$ we may write Z_f in place of $Z_{\{f\}}$.

Note that if I is the A-ideal generated by S, then $Z_I = Z_S$, since $f(P) = g(P) = 0$ implies $(f + g)(P) = 0$ and $f(P) = 0$ implies $(fg)(P) = 0$. Thus we can always replace S by the ideal (S) that it generates, or by any set of generators for (S).

Example 12.9. We have $Z_\emptyset = Z_{(0)} = \mathbb{A}^n$ and $Z_{\{1\}} = Z_{(1)} = \emptyset$.

For any $S, T \subseteq A$ we have

$$S \subseteq T \quad \Longrightarrow \quad Z_T \subseteq Z_S,$$

but the converse need not hold, even if S and T are ideals: consider $T = (x_1)$ and $S = (x_1^2)$.

We now recall the notion of a Noetherian ring and the Hilbert basis theorem.

Definition 12.10. A commutative ring R is *noetherian* if every R-ideal is finitely generated.[4] Equivalently, every infinite ascending chain of R-ideals

$$I_1 \subseteq I_2 \subseteq \cdots$$

eventually stabilizes, that is, $I_{n+1} = I_n$ for all sufficiently large n.

Theorem 12.11 (Hilbert basis theorem)**.** *If R is a noetherian ring, then so is $R[x]$.*

Proof. See [1, Theorem 14.6.7] or [2, Theorem 8.32]. □

Note that we can apply the Hilbert basis theorem repeatedly: if R is noetherian then so is $R[x_1]$, and so is $(R[x_1])[x_2] = R[x_1, x_2]$, …, and so is $R[x_1, \ldots, x_n]$. Like every field, $\bar{k}$ is a noetherian ring (it has just two ideals, so it certainly satisfies the ascending chain condition). Thus $A = \bar{k}[x_1, \ldots, x_n]$ is noetherian, so every A-ideal is finitely generated. It follows that every algebraic set can be written in the form Z_S with S finite.

Definition 12.12. For an algebraic set $Z \subseteq \mathbb{A}^n$, the *ideal of* Z is the set

$$I(Z) = \{f \in A : f(P) = 0 \text{ for all } P \in Z\},$$

where A is the polynomial ring $\bar{k}[x_1, \ldots, x_n]$.

The set $I(Z)$ is clearly an A-ideal (it is closed under addition and under multiplication by elements of A), and we note that

$$Y \subseteq Z \quad \Longrightarrow \quad I(Z) \subseteq I(Y)$$

and

$$I(Y \cup Z) = I(Y) \cap I(Z)$$

(both statements are immediate from the definition).

We have $Z = Z_{I(Z)}$ for every algebraic set Z, but it is not true that $I = I(Z_I)$ for every ideal I. As a counterexample, consider $I = (f^2)$ for some polynomial $f \in A$. In this case

$$I(Z_{(f^2)}) = (f) \neq (f^2).$$

In order to avoid this situation, we want to restrict our attention to *radical* ideals.

Definition 12.13. Let R be a commutative ring. For any R-ideal I we define

$$\sqrt{I} = \{x \in R : x^r \in I \text{ for some integer } r > 0\},$$

and say that I is a *radical ideal* if $I = \sqrt{I}$.

Lemma 12.14. *For any ideal I in a commutative ring R, the set $\sqrt{I}$ is an ideal.*

Proof. Let $x \in \sqrt{I}$ with $x^r \in I$. For any $y \in R$ we have $y^r x^r = (xy)^r \in I$, so $xy \in \sqrt{I}$. If $y \in \sqrt{I}$ with $y^s \in I$, then every term in the sum

$$(x+y)^{r+s} = \sum_i \binom{r+s}{i} x^i y^{r+s-i}$$

is a multiple of either $x^r \in I$ or $y^s \in I$, hence lies in I, so $(x+y)^{r+s} \in I$ and $(x+y) \in \sqrt{I}$. □

[4] The term "noetherian" refers to the mathematician Emmy Noether. The word noetherian is used so commonly in algebraic geometry (and elsewhere) that it is typically no longer capitalized (like abelian).

Theorem 12.15 (Hilbert's *Nullstellensatz*)**.** *For every ideal $I \subseteq \bar{k}[x_1, \dots, x_n]$ we have*

$$I(Z_I) = \sqrt{I}.$$

Proof. See [3, Theorem 7.1]. □

Nullstellensatz literally means "zero locus theorem." The theorem above is the strong of the *Nullstellensatz*; it implies the weak *Nullstellensatz*:

Theorem 12.16 (weak *Nullstellensatz*)**.** *For any proper ideal $I \subseteq \bar{k}[x_1, \dots, x_n]$ the variety Z_I is nonempty.*

Proof. Suppose I is an ideal for which Z_I is the empty set. Then $I(Z_I) = (1)$, and by the strong *Nullstellensatz*, $\sqrt{I} = (1)$. But then $1^r = 1 \in I$, so I is not proper. □

Note the importance of working over $\bar{k}$. It is easy to find proper ideals I for which $Z_I(k) = \emptyset$ when k is not algebraically closed; consider $Z_{(x^2+y^2+1)}(\mathbb{Q})$ in $\mathbb{A}^2$. A useful corollary of the weak *Nullstellensatz* is the following.

Corollary 12.17. *The maximal ideals of the ring $\bar{k}[x_1, \dots, x_n]$ are all of the form*

$$m_P = (x_1 - P_1, \dots, x_n - P_n)$$

for some point $P = (P_1, \dots, P_n)$ in $\mathbb{A}^n(\bar{k})$.

Proof. The evaluation map that sends $f \in \bar{k}[x_1, \dots, x_n]$ to $f(P) \in \bar{k}$ is a surjective ring homomorphism with kernel m_P. Thus $\bar{k}[x_1, \dots, x_n]/m_P \simeq \bar{k}$ is a field, hence m_P is a maximal ideal. If m is any maximal ideal in $\bar{k}[x_1, \dots, x_n]$, then it is a proper ideal, and by the weak *Nullstellensatz* the algebraic set Z_m is nonempty and contains a point $P \in \mathbb{A}^n$. So $m_P \subseteq I(Z_m)$, but also $m \subseteq I(Z_m)$. The ideal $I(Z_m)$ is a proper ideal (since Z_m is nonempty) and the ideals m and m_P are both maximal, so $m = I(Z_m) = m_P$. □

We also have the following corollary of the strong *Nullstellensatz*.

Corollary 12.18. *There is a one-to-one inclusion-reversing correspondence between radical ideals $I \subseteq \bar{k}[x_1, \dots, x_n]$ and algebraic sets $Z \subseteq \mathbb{A}^n(\bar{k})$ in which $I = I(Z)$ and $Z = Z_I$.*

Remark 12.19. It is hard to overstate the importance of Corollary 12.18; it is the basic fact that underlies nearly all of algebraic geometry. It tells us that the study of algebraic sets (geometric objects) is the same thing as the study of radical ideals (algebraic objects). It also suggests ways in which we might generalize our notion of an algebraic set: there is no reason to restrict ourselves to radical ideals in the ring $\bar{k}[x_1, \dots, x_n]$, there are many other rings we might consider. This approach eventually leads to the much more general notion of a *scheme*, but for our first foray into algebraic geometry we will stick to algebraic sets (in particular, varieties, which we will define momentarily).

Definition 12.20. A algebraic set is *irreducible* if it is nonempty and not the union of two smaller algebraic sets.

Theorem 12.21. *An algebraic set is irreducible if and only if its ideal is prime.*

Proof. (⇒) Let Y be an irreducible algebraic set and suppose $fg \in I(Y)$ for some $f, g \in A$. We will show that either $f \in I(Y)$ or $g \in I(Y)$ (and therefore $I(Y)$ is prime).

$$\begin{aligned} Y \subseteq Z_{fg} &= Z_f \cup Z_g \\ &= (Y \cap Z_f) \cup (Y \cap Z_g), \end{aligned}$$

and since Y is irreducible we must have either $Y = (Y \cap Z_f) = Z_f$ or $Y = (Y \cap Z_g) = Z_g$), hence either $f \in I(Y)$ or $g \in I(Y)$. Therefore $I(Y)$ is a prime ideal.

(⇐) Now suppose $I(Y)$ is prime and that $Y = Y_1 \cup Y_2$. We will show that either $Y = Y_1$ or $Y = Y_2$. This will show that Y is irreducible, since Y must be nonempty ($I(Y) \neq A$ because $I(Y)$ is prime). We have

$$I(Y) = I(Y_1 \cup Y_2) = I(Y_1) \cap I(Y_2) \supseteq I(Y_1)I(Y_2),$$

and therefore $I(Y)$ divides/contains either $I(Y_1)$ or $I(Y_2)$, since $I(Y)$ is a prime ideal, but it is also contained in both $I(Y_1)$ and $I(Y_2)$, so either $I(Y) = I(Y_1)$ or $I(Y) = I(Y_2)$. Thus either $Y = Y_1$ or $Y = Y_2$, since algebraic sets with the same ideal must be equal. □

References

[1] M. Artin, *Algebra*, 2nd edition, Pearson Education, 2011.

[2] A. Knapp, *Basic Algebra*, Springer, 2006.

[3] A. Knapp, *Advanced Algebra*, Springer, 2007.

[4] J.S. Milne, *Fields and Galois Theory*, 2012.

18.782 Introduction to Arithmetic Geometry **Fall 2013**
Lecture #13 **10/22/2013**

As before, k is a perfect field, $\bar{k}$ is a fixed algebraic closure of k, and $\mathbb{A}^n = \mathbb{A}^n(\bar{k})$ is n-dimensional affine space.

13.1 Affine varieties

Definition 13.1. An algebraic set $Z \in \mathbb{A}^n$ is said to be *defined over* k if its ideal is generated by polynomials in $k[x_1, \ldots, k_n]$, that is, $I(Z)$ is equal to the ideal generated by $I(Z) \cap k[x_1, \ldots, x_n]$ in $\bar{k}[x_1, \ldots, k_n]$. We write Z/k to indicate that Z is an algebraic set that is defined over k and define the ideal

$$I(Z/k) = I(Z) \cap k[x_1, \ldots, x_n].$$

When Z is defined over k the action of the absolute Galois group G_k on $\mathbb{A}^n$ induces an action on Z, since for any $\sigma \in G_k$, any $f \in k[x_1, \ldots, x_n]$, and any $P \in \mathbb{A}^n$ we have

$$f(P^\sigma) = f(P)^\sigma.$$

In this case we have $Z(k) = \{P \in Z : P^\sigma = P \text{ for all } \sigma \in G_k\} = Z^{G_k}$.

Definition 13.2. Let Z be an algebraic set defined over k. The *affine coordinate ring* of Z/k is the ring

$$k[Z] = \frac{k[x_1, \ldots, x_n]}{I(Z/k)}.$$

We similarly define

$$\bar{k}[Z] = \frac{\bar{k}[x_1, \ldots, x_n]}{I(Z)}.$$

The coordinate ring $k[Z]$ may have zero divisors; it is an integral domain if and only if $I(Z/k)$ is a prime ideal. Even if $k[Z]$ has no zero divisors, $\bar{k}[Z]$ may still have zero divisors (the fact that $I(Z/k)$ is a prime ideal does not guarantee that $I(Z)$ is a prime ideal; the principal ideal (x^2+1) is prime in $\mathbb{Q}$ but not in $\overline{\mathbb{Q}}$, for example). We want $k[Z]$ to be an integral domain so that we can work with its fraction field. Recall from last lecture that $I(Z)$ is a prime ideal if and only if Z is irreducible. This motivates the following definition.

Definition 13.3. An *affine variety* V is an irreducible algebraic set in $\mathbb{A}^n$.[1]

An algebraic set Z is a variety if and only if $I(Z)$ is a prime ideal; the one-to-one correspondence between algebraic sets and radical ideals restricts to a one-to-one correspondence between varieties and prime ideals (note that every prime ideal is necessarily a radical ideal). The set $\mathbb{A}^n$ is a variety since $I(\mathbb{A}^n)$ is the zero ideal, which is prime in the ring $\bar{k}[x_1, \ldots, k_n]$ because it is an integral domain (the zero ideal is prime in any integral domain).

Definition 13.4. Let V/k be an affine variety defined over k. The *function field* $k(V)$ of V is the fraction field of the coordinate ring $k[V]$.

We similarly define the function field of V over any extension of k on which V is defined. Every variety is defined over $\bar{k}$, so we can always refer to the function field $\bar{k}(V)$.

[1] Not all authors require varieties to be irreducible (but many do).

13.1.1 Dimension

Definition 13.5. The *dimension* of an affine variety V is the transcendence degree of the field extension $\bar{k}(V)/\bar{k}$.

Lemma 13.6. *The dimension of $\mathbb{A}^n$ is n, and the dimension of any point $P \in \mathbb{A}^n$ is 0.*

Proof. We have $\bar{k}[\mathbb{A}^n] = \bar{k}[x_1, \dots, x_n]/(0) = \bar{k}[x_1, \dots, x_n]$, so $\bar{k}(\mathbb{A}^n) = \bar{k}(x_1, \dots, x_n)$ is a purely transcendental extension of $\bar{k}$ with transcendence degree n. For the point P, the ideal $I(P) = m_P$ is maximal, so the coordinate ring $\bar{k}[P] = \bar{k}[x_1, \dots, x_n]/m_P$ is a field isomorphic to $\bar{k}$, as is $\bar{k}(P)$, and the transcendence degree of $\bar{k}/\bar{k}$ is obviously 0. □

Let us note an alternative definition of dimension using the *Krull dimension* of a ring.

Definition 13.7. The *Krull dimension* of a commutative ring R is the supremum of the set of integers d for which there exists a chain of distinct prime R-ideals

$$\mathfrak{p}_0 \subsetneq \mathfrak{p}_1 \subsetneq \cdots \subsetneq \mathfrak{p}_d.$$

The Krull dimension of a ring need not be finite, even when the ring is noetherian, but the Krull dimension of $\bar{k}[x_1, \dots, x_n]$ is finite, equal to n, and this bounds the Krull dimension of the coordinate ring of any variety $V \subseteq \mathbb{A}^n$. The following theorem implies that dimension of a V is equal to the Krull dimension of $\bar{k}[V]$.

Theorem 13.8. *Let k be a field and let R be an integral domain finitely generated as a k-algebra. The Krull dimension of R is the transcendence degree of its fraction field over k.*

Proof. See [1, Theorem 7.22]. □

Now consider a chain of distinct prime ideals in $\bar{k}[V]$ of length d equal to the Krull dimension of $\bar{k}[V]$.

$$\mathfrak{p}_0 \subsetneq \mathfrak{p}_1 \subsetneq \cdots \subsetneq \mathfrak{p}_d.$$

Since $\bar{k}[V]$ is an integral domain, the zero ideal is prime, so $\mathfrak{p}_0 = (0)$ (otherwise the chain would not be maximal). There is a one-to-one correspondence between ideals of the quotient ring $\bar{k}[V] = \bar{k}[x_1, \dots, x_n]$ and ideals of $\bar{k}[x_1, \dots, x_n]$ that contain $I(V)$, and this correspondence preserves prime ideals (this follows from the third ring isomorphism theorem). Thus we have a chain of distinct prime ideals in $\bar{k}[x_1, \dots, x_n]$:

$$I(V) = I_0 \subsetneq I_1 \subsetneq \cdots \subsetneq I_d,$$

This corresponds to a chain of distinct varieties (with inclusions reversed):

$$V_d \subsetneq V_1 \subsetneq \cdots \subsetneq V_0 = V.$$

Conversely, we could have started with a chain of distinct varieties V and obtained a chain of distinct prime ideals in $\bar{k}[V]$. This one-to-one correspondence yields an alternative definition of the dimension of V.

Definition 13.9. The *geometric dimension* of a variety V is the largest integer d for which there exists a chain

$$V_0 \subsetneq \cdots \subsetneq V_d = V$$

of distinct varieties contained in V.

The discussion above shows that this agrees with our earlier definition. This notion of dimension also works for algebraic sets: the dimension of an algebraic set Z is the largest integer d for which there exists a chain of distinct varieties (irreducible algebraic sets) contained in Z.

13.1.2 Singular points

Definition 13.10. Let $V \subseteq \mathbb{A}^n$ be a variety, and let $f_1, \ldots, f_m \in \bar{k}[x_1, \ldots, x_n]$ be a set of generators for $I(V)$. A point $P \in V$ is a *nonsingular* (or *smooth*) if the $m \times n$ Jacobian matrix $M(P)$ with entries

$$M_{ij}(P) = \frac{\partial f_i}{\partial x_j}(P)$$

has rank $n - \dim V$; otherwise P is a *singular point* of V. If V has no singular points than we say that V is *smooth.*

A useful fact that we will not prove is that if one can show that the rank of $M(P)$ is equal to $n - d$ for every point $P \in \mathbb{A}^n$, then V is a smooth variety of dimension d.

13.2 Projective space

Definition 13.11. n-dimensional *projective space* $\mathbb{P}^n$ over k is the set of all points in $\mathbb{A}^{n+1} - \{\mathbf{0}\}$ modulo the equivalence relation

$$(a_0, \ldots, a_n) \sim (\lambda a_0, \ldots, \lambda a_n)$$

for all $\lambda \in \bar{k}^\times$. We use the ratio notation $(a_0 : \ldots : a_n)$ to denote the equivalence class of $(a_0, \ldots, a_n)$, and call it a *projective point* or a *point* in $\mathbb{P}^n$. The set of k-rational points in $\mathbb{P}^n$ is

$$\mathbb{P}^n(k) = \{(a_0 : \ldots : a_n) \in \mathbb{P}^n : a_0, \ldots, a_n \in k\}$$

(and similarly for any extension of k in $\bar{k}$).

Remark 13.12. Note that $(a_0 : \ldots : a_n) \in \mathbb{P}^n(L)$ does not necessarily imply that all a_i lie in L, it simply means that there exists some $\lambda \in \bar{k}^\times$ for which all λa_i lie in L. However we do have $a_i/a_j \in L$ for all $0 \le i, j \le n$.

The absolute Galois group G_k acts on $\mathbb{P}^n$ via

$$(a_0 : \ldots : a_n)^\sigma = (a_0^\sigma : \ldots : a_n^\sigma).$$

This action is well defined, since $(\lambda P)^\sigma = \lambda^\sigma P^\sigma \sim P^\sigma$ for any $\lambda \in \bar{k}^\times$ and $P \in \mathbb{A}^{n+1} - \{\mathbf{0}\}$. We then have

$$\mathbb{P}^n(k) = (\mathbb{P}^n)^{G_k}.$$

13.3 Homogeneous polynomials

Definition 13.13. A polynomial $f \in \bar{k}[x_0, \ldots, x_n]$ is *homogenous of degree d* if

$$f(\lambda x_0, \ldots, \lambda x_n) = \lambda^d f(x_0, \ldots, x_n)$$

for all $\lambda \in \bar{k}$. Equivalently, every monomial in f has total degree d. We say that f is *homogeneous* if it is homogeneous of some degree.

Fix an integer $i \in [0, n]$. Given any polynomial $f \in \bar{k}[x_0, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n]$ in n variables, let d be the total degree of f and define the *homegenization* of f (with respect to x_i) to be the polynomial

$$F(x_0, \ldots, x_n) = x_i^d f\left(\frac{x_0}{x_i}, \ldots, \frac{x_{i-1}}{x_i}, \frac{x_{i+1}}{x_i}, \ldots, \frac{x_n}{x_i}\right).$$

Conversely, given any homogenous polynomial $F \in \bar{k}[x_0, \ldots, x_n]$, the polynomial

$$f(x_0, \ldots, x_{i-1}, x_{i+1}, \ldots) = f(x_0, \ldots, x_{i-1}, 1, x_{i+1}, \ldots, x_n)$$

is the *dehomegenization* of F (with respect to x_i).

Let $P = (a_0 : \ldots : a_n)$ be a point in $\mathbb{P}^n$ and let and f be a homogeneous polynomial in $\bar{k}[x_0, \ldots, x_n]$. The value $f(a_0, \ldots, a_n)$ will depend, in general, on our choice of representative $(a_0, \ldots, a_n)$ for P. However,

$$f(a_0, \ldots, a_n) = 0 \quad \Longleftrightarrow \quad f(\lambda a_0, \ldots, \lambda a_n) = 0 \text{ for all } \lambda \in \bar{k}^{\times}.$$

Thus it makes sense to write $f(P) = 0$ (or $f(P) \neq 0$), and the zero locus of a homogeneous polynomial is a well-defined subset of $\mathbb{P}^n$.

13.3.1 Affine covering of projective space

For $0 \leq i \leq n$, the zero locus of the homogeneous polynomial x_i is the *hyperplane*

$$H_i = \{(a_0 : \ldots : a_{i-1} : 0 : a_{i+1} : \ldots : a_n) \in \mathbb{P}^n\},$$

which corresponds to a copy of $\mathbb{P}^{n-1}$ embedded in $\mathbb{P}^n$.

Definition 13.14. The complement of H_i in $\mathbb{P}^n$ is the *affine patch* (or *affine chart*)

$$U_i = \{(a_0 : \ldots : a_{i-1} : 1 : a_{i+1} : \ldots : a_n) \in \mathbb{P}^n\},$$

which corresponds to a copy of $\mathbb{A}^n$ embedded in $\mathbb{P}^n$ (note that fixing $a_i = 1$ fixes a choices of representative for the projective point $(a_0 : \ldots : a_{i-1} : 1 : a_{i+1} : \ldots : a_n)$).

If we pick a hyperplane, say H_0, we can partition $\mathbb{P}^n$ as

$$\mathbb{P}^n = U_0 \sqcup H_0 \simeq \mathbb{A}^n \sqcup \mathbb{P}^{n-1}.$$

We can now apply the same procedure to $H_0 \simeq \mathbb{P}^{n-1}$, and repeating this yields

$$\mathbb{P}^n \simeq \mathbb{A}^n \sqcup \mathbb{A}^{n-1} \sqcup \cdots \sqcup \mathbb{A}^1 \sqcup \mathbb{P}^0,$$

where the final $\mathbb{P}^0$ corresponds a single projective point in $\mathbb{P}^n$.

Alternatively, we can view $\mathbb{P}^n$ as the union of $n + 1$ (overlapping) affine patches, each corresponding to a copy of $\mathbb{A}^n$ embedded in $\mathbb{P}^n$. Note that every projective point P lies in at least one affine patch.

Remark 13.15. Just as a manifold is locally defined in terms of an atlas of overlapping charts (each of which maps the neighborhood of a point to an open set in Euclidean space), we can view $\mathbb{P}^n$ as being locally defined in terms of its overlapping affine patches, viewing each as mapping a neighborhood of $\mathbb{P}^n$ to $\mathbb{A}^n$ (this viewpoint can be made quite rigorous, but we will not do so here).

13.4 Projective varieties

For any set S of polynomials in $\bar{k}[x_0, \ldots, x_n]$ we define the (*projective*) *algebraic set*

$$Z_S = \{P \in \mathbb{P}^n : f(P) = 0 \text{ for all homogeneous } f \in S\}.$$

Definition 13.16. A *homogeneous ideal* in $\bar{k}[x_0, \ldots, x_n]$ is an ideal that is generated by a set of homogeneous polynomials.

Note that not every polynomial in a homogeneous ideal I is homogeneous (the sum of homogeneous polynomials of different degrees is not homogeneous), but this has no impact on the algebraic set Z_I, since our definition of Z_I ignores elements of I that are not homogeneous.

Definition 13.17. Let Z be an algebraic set in $\mathbb{P}^n$, the (*homogeneous*) *ideal of* Z is the ideal $I(Z)$ generated by all the homogeneous polynomials in $\bar{k}[x_0, \ldots, x_n]$ that vanish at every point in Z.

We say that Z is *defined over* k if its ideal can be generated by homogeneous polynomials in $k[x_0, \ldots, x_n]$, and write Z/k to indicate this. If Z is defined over k the set of *k-rational points* on Z is

$$Z(k) = Z \cap \mathbb{P}^n(k) = Z^{G_k},$$

and similarly for any extension of k in $\bar{k}$.

As with affine varieties, we say that an algebraic set in $\mathbb{P}^n$ is irreducible if it is nonempty and not the union of two smaller algebraic sets in $\mathbb{P}^n$.

Definition 13.18. A (*projective*) *variety* is an irreducible algebraic set in $\mathbb{P}^n$.

As you will show on the problem set, an algebraic set $Z \subseteq \mathbb{P}^n$ is irreducible if and only if $I(Z)$ is prime. One can then define the coordinate ring $k[V]$ and function field $k(V)$ of a projective variety exactly as in the affine case. Here we take a different approach using affine patches, which yields the same result.

Definition 13.19. Let V be a projective variety with homogeneous ideal $I = (f_1, \ldots, f_m)$. Let I_i be the ideal generated by the dehomegenizations of $f_1, \ldots, f_m$ at x_i. Then I_i is a prime ideal (since I is) and the ith *affine part* of V is the affine variety $V_i = V \cap U_i$ whose ideal is I_i. We can then write $V = \bigcup_i V_i$ as the union of its affine parts.

Definition 13.20. The *dimension* of a projective variety V is the maximum of the dimensions of its affine parts, and V is *smooth* if and only if all its affine parts are.

Finally, we define the *coordinate ring* $k[V]$ of a projective variety V/k to be the coordinate ring of any of its nonempty affine parts (we will prove below that it doesn't matter which one we pick), and the *function field* $k(V)$ of V is the fraction field of its coordinate ring, and similarly for any extension of k in $\bar{k}$.

13.5 Projective closure

Definition 13.21. If $Z \subseteq \mathbb{A}^n$ is any affine algebraic set, we can embed it in $\mathbb{P}^n$ by identifying $\mathbb{A}^n$ with the affine patch U_0 of $\mathbb{P}^n$; we write $Z \subseteq \mathbb{A}^n \subset \mathbb{P}^n$ to indicate this embedding. The *projective closure* of Z in $\mathbb{P}^n$, denoted $\overline{Z}$, is the projective algebraic set defined by the ideal generated by all the homogenizations (with respect to x_0) of all the polynomials in $I(Z)$.

When the ideal of an algebraic set $Z \subseteq \mathbb{A}^n$ is principal, say $I(Z) = (f)$, then $I(\overline{Z})$ is generated by the homogenization of f. But in general the homegenizations of a set of generators for $I(Z)$ *do not* generate $I(\overline{Z})$, as shown by the following example.

Example 13.22. Consider the *twisted cubic* $C = \{(t, t^2, t^3) : t \in \bar{k}\} \subseteq \mathbb{A}^3 \subset \mathbb{P}^3$. It is the zero locus of the ideal

$$(x^2 - y, x^3 - z)$$

in $\bar{k}[x, y, z]$, hence an algebraic set, in fact, an affine variety of dimension 1 (an *affine curve*). To see this note that $\bar{k}[C] = \bar{k}[x, y, z]/I(C) \simeq \bar{k}[x]$ is obviously an integral domain, so $I(C)$ is prime, and the function field $\bar{k}(C) \simeq \bar{k}(x)$ has transcendence degree 1.

If we homogenize the generators of $I(C)$ by introducing a new variable w, we get the homogeneous ideal $I = (x^2 - wy, x^3 - w^2 z)$. The zero locus of this ideal in $\mathbb{P}^3$ is

$$\{(1 : t : t^2 : t^3) : t \in \bar{k}\} \cup \{(0 : 0 : y : z) : y, z \in \bar{k}\},$$

which ought to strike you as a bit too large to be the projective closure of C; indeed, the homogeneous polynomial $y^2 - xz$ is not in I even though $y^2 - x$ is in $I(C)$, so this cannot be $\overline{C}$. But if we instead consider the homogeneous ideal

$$(x^2 - wy,\ xy - wz,\ y^2 - xz),$$

we see that its zero locus is

$$\{(1 : t : t^2 : t^3) : t \in \bar{k}\} \cup \{(0 : 0 : 0 : 1)\},$$

and we claim this is $\overline{C}$. There are many ways to prove this, but here is completely elementary argument: Suppose that $f \in \bar{k}[w, x, y, z]$ is homogeneous of degree d, with C in its zero locus. Then the polynomial $g(t) = f(1, t, t^2, t^3)$ must be the zero polynomial (here we use that $\bar{k}$ is infinite). If $f(0, 0, 0, 1) \neq 0$, then f must contain a term of the form cz^d with $c \in \bar{k}^\times$. But then $g(t) = ct^{3d} + h(t)$ with $\deg h \leq 3(d - 1) + 2 = 3d - 1 < 3d$, which means that g cannot be the zero polynomial, a contradiction. The claim follows.

Theorem 13.23. *If $V \in \mathbb{A}^n \subset \mathbb{P}^n$ is an affine variety then its projective closure $\overline{V}$ is a projective variety, and $V = \overline{V} \cap \mathbb{A}^n$ is an affine part of $\overline{V}$.*

Proof. For any polynomial $f \in \bar{k}[x_1, \ldots, x_n]$, let $\overline{f} \in \bar{k}[x_0, x_1, \ldots, x_n]$ denote its homogenization with respect to x_0. For any $f \in \bar{k}[x_1, \ldots, x_n]$ and any point $P \in \mathbb{A}^n$, we have $f(P) = 0$ if and only if $\overline{f}(\overline{P}) = 0$, where $\overline{P} = (1 : a_1 : \ldots : a_n)$ is the projective closure of P (viewing points as singleton algebraic sets). It follows that $V = \overline{V} \cap \mathbb{A}^n$.

To show that $\overline{V}$ is a projective variety, we just need to show that it is irreducible, equivalently (by Problem Set 6), that its ideal is prime. So let $fg \in I(\overline{V})$. Then fg vanishes on $\overline{V}$, hence it vanishes on V, as does the dehomegenization $f(1, x_1, \ldots, x_n)g(1, x_1, \ldots, x_n)$ But $I(V)$ is prime (since V is a variety), so either $f(1, x_1, \ldots, x_n)$ of $g(1, x_1, \ldots, x_n)$ lies in $I(V)$, and therefore one of f and g lies in $I(\overline{V})$. Thus $I(\overline{V})$ is prime. □

Theorem 13.24. *Let V be a projective variety and let V_i be any of its nonempty affine parts. Then V_i is an affine variety and V is its projective closure.*

Proof. Without loss of generality we assume $i = 0$ and use the notation introduced in the proof above, identifying $\mathbb{A}^n$ with U_0. As above, for any $f \in \bar{k}[x_1, \ldots, x_n]$ and any point $P \in \mathbb{A}^n$, we have $f(P) = 0$ if and only if $\overline{f}(\overline{P}) = 0$. It follows that V_0 is an algebraic set defined by the ideal generated by the dehomegenization of all the homogeneous polynomials in $I(V)$, and therefore $V = \overline{V_0}$.

To show that V_0 is an affine variety, we just need to check that $I(V_0)$ is a prime ideal. So let $fg \in I(V_0)$. Then $\overline{fg} \in I(V)$ and therefore either $\overline{f}$ or $\overline{g}$ is in $I(V)$ (since $I(V)$ is prime), and then either f or g must lie in $I(V_0)$. Thus $I(V_0)$ is prime. □

Remark 13.25. Theorem 13.23 is still true if "variety" is replaced by "algebraic set", but Theorem 13.24 is not.

Corollary 13.26. *The dimension, coordinate ring, and function field of an affine variety are equal to those of its projective closure. The dimension, coordinate ring, and function field of a projective variety are equal to those of each of its nonempty affine parts.*

Remark 13.27. One can define the function field of a projective variety V directly in terms of its homogeneous ideal $I(V)$ rather than identifying it with the function field of its nonempty affine pieces (all of which are isomorphic), but some care is required. The function field $\bar{k}(V)$ is *not* the fraction field of $\bar{k}[x_0, \ldots, x_n]/I(V)$, it is the subfield of $\bar{k}[x_0, \ldots, x_n]/I(V)$ consisting of all fractions g/h where g and h are both homogeneous polynomials (modulo $I(V)$) of the *same degree*, with $h \neq 0$. This restriction is necessary in order for us to sensibly think of elements of $\bar{k}(V)$ as functions from V to $\bar{k}$. In order to evaluate a function $f(x_0, \ldots, x_n)$ at a projective point $P = (a_0 : \ldots : a_n)$ in a well-defined way we must require that

$$f(\lambda a_0, \ldots, \lambda a_n) = f(a_0, \ldots, a_n)$$

for any $\lambda \in \bar{k}^\times$. If $f = g/h$ with g and h homogeneous of degree d, then

$$f(\lambda a_0, \ldots, \lambda a_n) = \frac{g(\lambda a_0, \ldots, \lambda a_n)}{h(\lambda a_0, \ldots, \lambda a_n)} = \frac{\lambda^d g(a_0, \ldots, a_n)}{\lambda^d h(a_0, \ldots, a_n)} = \frac{g(a_0, \ldots, a_n)}{h(a_0, \ldots, a_n)} = f(a_0, \ldots, a_n),$$

as required. With this definition the function field $\bar{k}(V)$ is isomorphic to the function field of each of its nonempty affine parts.

References

[1] A. Knapp, *Advanced Algebra*, Springer, 2007.

18.782 Introduction to Arithmetic Geometry **Fall 2013**
Lecture #14 **10/24/2013**

As usual, k is a perfect field and $\bar{k}$ is a fixed algebraic closure of k. Recall that an affine (resp. projective) variety is an irreducible alebraic set in $\mathbb{A}^n = \mathbb{A}^n(\bar{k})$ (resp. $\mathbb{P}^n = \mathbb{P}^n(\bar{k})$).

14.1 Affine morphisms

We begin our discussion of maps between varieties with the simplest case, morphisms of affine varieties.

Definition 14.1. Let $X \subseteq \mathbb{A}^m$ and $Y \subseteq \mathbb{A}^n$ be affine varieties. A *morphism* $f\colon X \to Y$ is a map $f(P) := (f_1(P), \ldots, f_n(P))$ defined by polynomials $f_1, \ldots, f_n \in \bar{k}[x_1, \ldots, x_m]$ such that $f(P) \in Y$ for all points $P \in X$. We may regard $f_1, \ldots, f_n$ as representatives of elements of the coordinate ring $\bar{k}[X] = \bar{k}[x_1, \ldots, x_m]/I(X)$; we are evaluating $f_1, \ldots, f_n$ only at points in X, so there is no reason to distinguish them modulo the ideal $I(X)$.

As befits their name, morphisms can be composed: if $f\colon X \to Y$ and $g\colon Y \to Z$ are morphisms of varieties $X \subseteq \mathbb{A}^m$, $Y \subseteq \mathbb{A}^n$, and $Z \subseteq \mathbb{A}^r$, then $(g \circ f)\colon X \to Z$ is defined by

$$(g \circ f)(P) := g(f(P)) = \big(g_1(f_1(P), \ldots, f_n(P)),\ \ldots,\ g_r(f_1(P), \ldots, f_n(P))\big).$$

Notice that in order for this composition to actually make sense, we need to pick particular representatives $g_1, \ldots, g_r \in \bar{k}[y_1, \ldots, y_m]$ modulo $I(Y)$ (of course it doesn't matter which). The rings $\bar{k}[x_1, \ldots, x_m]$ and $\bar{k}[X]$ are $\bar{k}$ algebras, so it makes sense to evaluate a polynomial with coefficients in $\bar{k}$ in either of these rings (depending on our perspective), but it does not make sense to "evaluate" an element of $\bar{k}[Y]$ at elements of $\bar{k}[X]$. We also have the identity morphism $f\colon X \to X$, which is defined by letting f_i be the polynomial x_i.[1]

Thus we have a category whose objects are affine varieties and whose morphisms are (no surpise) morphisms. Contrary to what you might expect (if you happened to be thinking of morphisms in the category of groups or rings), the image of a morphism is not necessarily a variety, or even an algebraic set.

Example 14.2. Consider the morphism $f\colon \mathbb{A}^2 \to \mathbb{A}^2$ defined by $f(x_1, x_2) = (x_1, x_1x_2)$. Its image is the entire affine plane except for the points on the x_1-axis with $x_2 \neq 0$. This is not an algebraic set; this is obvious if $\bar{k} = \mathbb{C}$, and in general, if $g(y_1, y_2)$ vanishes on the image of f, then for any infinitely many $c \in \bar{k}$ the polynomial $h(t) = g(t, c)$ has infinitely many zeroes, hence is the zero polynomial, and this implies that g is the zero polynomial. Thus $I(\operatorname{im} f)$ is the zero ideal and the only algebraic set containing $\operatorname{im} f$ is all of $\mathbb{A}^2$.

On the other hand, if you were thinking of morphisms in the category of topological spaces (which is the better analogy), then morphisms of varieties behave as expected; indeed, they are continuous maps (and more), we just need to put the right topology on our varieties.

Definition 14.3. In the *Zariski topology* on $\mathbb{A}^n$ (resp. $\mathbb{P}^n$), the closed sets are precisely the algebraic sets. Any algebraic set in $\mathbb{A}^n$ (resp. $\mathbb{P}^n$) then inherits the subspace topology.

[1] Note that we are using the symbol x_i in three different ways: as an indeterminate used to define the polynomial ring $R = \bar{k}[x_1, \ldots, x_m]$, as an element of R (i.e., a polynomial), and as the function $x_i\colon \mathbb{A}^m(\bar{k}) \to \bar{k}$ that evaluates the polynomial x_i on a given input.

Andrew V. Sutherland

Let us verify that this actually defines a topology: the empty set and $\mathbb{A}^n$ are algebraic sets defined by the ideals (1) and (0), respectively (and similarly for $\mathbb{P}^n$), and algebraic sets are closed under arbitrary intersections (we can take the zero locus of an arbitrary sum of ideals), and finite unions (we can take the zero locus of a finite product of ideals).[2]

With a topology in place we can now use words like *open, closed, dense*, etc., when referring to subsets of $\mathbb{A}^n$ or $\mathbb{P}^n$, with the understanding that they refer to the Zariski topology. Note that our definition of the projective closure of an affine variety V embedded in $\mathbb{P}^n$ is consistent with this; we proved last time that the projective closure of V in $\mathbb{P}^n$ is a variety (hence closed), and it is clearly the smallest closed set that contains V: a homogeneous polynomial in $\bar{k}[x_0, \dots, x_n]$ vanishes on V in $\mathbb{P}^n$ if and only if its dehomogenization vanishes on V in $\mathbb{A}^n$.

It should be noted that the Zariski topology is extremely coarse. In $\mathbb{A}^1$, for example, every nonempty open set is the complements of finite sets, and in general every nonempty open set is dense in $\mathbb{A}^n$ (and in $\mathbb{P}^n$); the same applies in the subspace of a variety. And the Zariski topology is definitely *not* a Hausdorff topology; indeed, the intersection of any pair of nonempty open sets is not only nonempty, it must be dense!

Theorem 14.4. *Every morphism $f : X \to Y$ of affine varieties is continuous. That is, the inverse image $f^{-1}(Z)$ of any algebraic subset $Z \subseteq Y$ is an algebraic subset of X.*

Proof. Showing that the inverse image of a closed set is closed is the same thing as showing that the inverse of an open set is open, which is the definition of a continuous map. So let Z be an algebraic subset of Y defined by the ideal $(g_1, \dots, g_r)$ (we include generators for $I(Y)$ in this list). Then $f^{-1}(Z)$ is the zero locus of $g_1(f_1, \dots, f_n), \dots, g_r(f_1, \dots, f_n)$ (as compositions of polynomials) in X, hence an algebraic of subset of X. □

Remark 14.5. It is not true that every continuous map between affine varieties is a morphism; the coarseness of the Zariski topology simply makes it too easy for a function to be continuous. The additional requirement that a morphism must satisfy is that it must also be a rational map, as we will see in the next lecture.

For affine varieties, an isomorphism is a bijective morphism whose inverse is a morphism, but we will use the more formal definition that applies in any category.

Definition 14.6. We say that two varieties $X \simeq Y$ are *isomorphic* if there exist morphisms $f\colon X \to Y$ and $g\colon Y \to X$ such that both $f \circ g$ and $g \circ f$ are the identity morphisms on X and Y, respectively. In this case we may refer to both f and g as *isomorphisms.*

Just as not every continuous map is an morphism, not every bicontinuous map (homemorphism) is an isomorphism. Indeed, not even a bicontinuous morphism is necessarily an isomorphism.

Example 14.7. Consider the map from $\mathbb{A}^1$ to $\mathbb{A}^2$ defined by $t \mapsto (t^2, t^3)$. The image of this map is a variety V (the polynomial $y^2 - x^3$ is irreducible in $\bar{k}[x, y]$, so the principal ideal $(y^2 - x^3)$ is prime). Thus we have a morphism $f\colon \mathbb{A}^1 \to V$, and it is clearly bijective; the inverse map can be defined as

$$f^{-1}(x, y) = \begin{cases} y/x & \text{if } x \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

[2] One needs to check that this also works for projective varieties and homogeneous ideals, but this is straight-forward; sums and products of homogeneous ideals are again homogeneous ideals and the rest follows from Problem 2 of Problem Set 6.

Moreover, f is a closed map (the only closed sets in $\mathbb{A}^1$ are points and $\mathbb{A}^1$ itself, and these are all mapped to closed sets in V), so it is bicontinuous and thus both a morphism and a homeomorphism. But it is not an isomorphism because its inverse is not a morphism; the function f^{-1} cannot be defined as a polynomial map.

The example above shows that two varieties may be isomorphic as topological spaces without being isomorphic as varieties; this should not be too surprising, the Zariski topology makes it very easy for varieties to be homoemorphic (indeed, every affine curve is homeomorphic to $\mathbb{A}^1$). On the other hand, in the example of the twisted cubic (see Lecture 13) we actually have an isomorphism of affine varieties.

We now come to a very important theorem that gives a one-to-one correspondence between morphisms of affine varieties $\phi\colon X \to Y$ and homomorphisms of their coordinate rings $\phi^*\colon \bar{k}[Y] \to \bar{k}[X]$ (note that the directions of the arrows are reversed). Actually, we want ϕ^* to be more than just a ring homomorphism, we also want it to fix the field $\bar{k}$. A compact way of saying this is to regard $\bar{k}[X]$ and $\bar{k}[Y]$ not as rings, but as algebras over $\bar{k}$, and require ϕ^* to be a homomorphism of $\bar{k}$-algebras. This means that that ϕ^* must commute with sums and products (in our setting this makes ϕ^* a ring homomorphism), and it must fix elements of $\bar{k}$.

In order to obtain an actual equivalence of categories, we want to specify objects that correspond to coordinate rings in a purely algebraic way that does not involve varieties. So consider an arbitrary integral domain R that is also a finitely generated $\bar{k}$-algebra; let us call such an R an *affine algebra.* If we denote the generators of R by $x_1, \ldots, x_n$, there is a canonical ring homomorphism

$$\bar{k}[x_1, \ldots, x_n] \to R$$

from the polynomial ring with indeterminates $x_1, \ldots, x_n$ onto R, and the kernel of this homomorphism is an ideal I for which $R = \bar{k}[x_1, \ldots, x_n]/I$. The ideal I is prime (since R is an integral domain), hence a radical ideal. Let V be the variety it defines in $\mathbb{A}^n$. Then by Hilbert's *Nullstellensatz*, we have $I = I(V)$ (note that here we use that $\bar{k}$ is algebraically closed). The coordinate ring of V is then $\bar{k}[V] \simeq \bar{k}[x_1, \ldots, x_n]/I \simeq R$.

Thus we have a one-to-one correspondence between affine varieties and affine algebras in which varieties correspond to their coordinate rings and affine algebras correspond to varieties as described above. By taking the morphisms to be k-algebra homomorphisms, we can consider the category of affine algebras. In order to prove that the category of affine varieties is equivalent to the category of affine algebras, we need to understand how their morphisms correspond.

Theorem 14.8. *The following hold:*

(i) Every morphism $\phi\colon X \to Y$ of affine varieties induces a morphism $\phi^*\colon \bar{k}[Y] \to \bar{k}[X]$ of affine algebras such that $\phi^*(g) = g \circ \phi$.

(ii) Every morphism $\theta\colon R \to S$ of affine algebras induces a morphism $\theta^*\colon X \to Y$ of affine varieties with $R \simeq \bar{k}[Y]$ and $S \simeq \bar{k}[X]$ such that the image of $\theta(g)$ in $\bar{k}[X]$ is $g \circ \theta^*$.

(iii) If $\phi\colon X \to Y$ and $\psi\colon Y \to Z$ are morphisms of affine varieties, then $(\psi \circ \phi)^* = \phi^* \circ \psi^*$.

Before proving the theorem, let us comment on the notation $\phi^*(g) = g \circ \phi$. In order for this to make sense, we need to interpret it as follows: given $g \in \bar{k}[Y] = \bar{k}[y_1, \ldots, y_n]/I(Y)$, we pick a representative $\hat{g} \in \bar{k}[y_1, \ldots, y_n]$ (so g is the coset $\hat{g} + I(Y)$) and then $\phi^*(g)$ is the reduction of the polynomial $\hat{g}\circ\phi = \hat{g}(\phi_1,\ldots,\phi_n) \in \bar{k}[x_1,\ldots,x_m]$ modulo $I(X)$ (i.e., its image under the quotient map). In short, $g\circ\phi$ means lift/compose/reduce.[3]

The key point is that for any f in $I(Y)$, the composition $f\circ\phi$ yields an element of $I(X)$, because ϕ maps points in X to points in Y and f vanishes at points in Y. Thus it does not matter which lift $\hat{g}$ we pick and our interpretation of $g\circ\phi$ is well defined.

Proof. We assume throughout that X and Y are varieties in $\mathbb{A}^m$ and $\mathbb{A}^n$, respectively.

(i) We first note that the operations of lifting from $\bar{k}[Y]$ to $\bar{k}[y_1,\ldots,y_n]$ and reducing from $\bar{k}[x_1,\ldots,x_m]$ to $\bar{k}[X]$ are both compatible with ring operations, and when lifting or reducing an element of $\bar{k}$ it remains fixed. Now if $g\in\bar{k}$ is a constant polynomial, then $g\circ\phi = g$, and for any $f,g\in\bar{k}[y_1,\ldots,y_n]$ we have

$$(f+g)\circ\phi = (f+g)(\phi_1,\ldots,\phi_n) = f(\phi_1,\ldots,\phi_n) + g(\phi_1,\ldots,\phi_n) = (f\circ\phi)+(g\circ\phi)$$

and

$$(fg)\circ\phi = (fg)(\phi_1,\ldots,\phi_n) = f(\phi_1,\ldots,\phi_n)g(\phi_1,\ldots,\phi_n) = (f\circ\phi)(g\circ\phi).$$

Thus ϕ^* is a ring homomorphism that fixes $\bar{k}$, hence a homomorphism of affine algebras.

(ii) Let $\theta\colon R\to S$ be a morphism of affine algebras. As described above, there exist varieties X and Y for which $R\simeq\bar{k}[Y]$ and $S\simeq\bar{k}[X]$, and any morphism $R\to S$ induces a morphism $\bar{k}[Y]\to\bar{k}[X]$ that commutes with these isomorphisms.[4] So without loss of generality we assume $\theta\colon\bar{k}[Y]\to\bar{k}[X]$ is an affine algebra morphism of coordinate rings. We now define a morphism $\theta^*\colon X\to Y$ by letting $\theta^* = (\theta(y_1),\ldots,\theta(y_n))$, where $\theta(y_i)$ denotes the image under θ of the image of the polynomial y_i in $\bar{k}[Y]$ under the quotient map from $\bar{k}[y_1,\ldots,y_n]$. For any $g\in\bar{k}[Y]$ we have

$$g\circ\theta^* = \hat{g}(\theta(y_1),\ldots,\theta(y_n)) = \theta(g(y_1,\ldots,y_n)) = \theta(g),$$

where the middle equality follows from the fact that θ is a ring homomorphism that fixes $\bar{k}$. We also note that for any $f\in I(Y)$ we have $f\circ\theta^* = \theta(f) = \theta(0) = 0$, so $f(\theta^*(P)) = 0$ for all $f\in I(Y)$ and $P\in X$, which implies that the image of θ^* lies in Y. Thus θ^* is indeed a morphism from X to Y as claimed.

(iii) For any $g\in\bar{k}[Z]$ we have

$$(\psi\circ\phi)^*(g) = g\circ(\psi\circ\phi) = (g\circ\psi)\circ\phi = \phi^*(g\circ\psi) = (\phi^*(\psi^*(g)) = (\phi^*\circ\psi^*)(g). \qquad\square$$

Corollary 14.9. *The categories of affine varieties and affine algebras are contravariantly equivalent.*[5]

Proof. The only thing that remains to be shown is that the two functors arising from (i) and (ii) of Theorem 14.8 are inverses, that is, we need to show that $(\phi^*)^* = \phi$ and $(\theta^*)^* = \theta$, up to isomorphism.[6] The second equality is clear from the statement of the theorem and the first is clear from its proof. □

[3] If we view $\phi_1,\ldots,\phi_n$ as elements of $\bar{k}[X]$, we also need to lift the ϕ_i in order to compute $g\circ\phi$.

[4] One says that the induced morphism is *natural*; more precisely, the functor from the category of function fields to the category of function fields of varieties is a *natural transformation* (in fact, a natural isomorphism). If you think this is just a fancy way of stating the obvious, you are right; but the same phenomenon occurs in more general situations where it is not always so obvious.

[5] Contravariantly equivalent categories are also called *dual* categories; they are also said to be *anti-equivalent*, but we won't use this term.

[6] Up to isomorphism means that the domains and codmains of the morphisms on either side of the equality need not be precisely equal, they just need to be isomorphic, and the isomorphisms and the morphisms must form a commutative diagram; in other words, $(\phi^*)^*$ is naturally isomorphic to ϕ (and similarly for θ).

Corollary 14.10. *All the nonempty affine parts of a projective variety are isomorphic.*

Proof. We proved in Lecture 13 (see Corollary 13.26) that the nonempty affine parts of a projective variety all have the same coordinate ring (up to isomorphism). □

Definition 14.11. If $\phi = (\phi_1, \ldots, \phi_n)$ is a morphism of varieties $X \to Y$ that are defined over k, we say that ϕ is *defined over* k if $\phi_1, \ldots, \phi_n \in k[Y]$. Equivalently, ϕ is defined over k if $\phi^\sigma = (\phi_1^\sigma, \ldots, \phi_n^\sigma) = \phi$ for all $\sigma \in G_k$.[7] If ϕ is an isomorphism defined over k and it has an inverse isomorphism defined over k, then we say that X and Y are *isomorphic over* k.

Corollary 14.12. *Let X and Y be affine varieties defined over k. If $\phi\colon X \to Y$ is a morphism defined over k then the affine algebra morphism $\phi^*\colon \bar{k}[Y] \to \bar{k}[X]$ restricts to an affine algebra morphism from $k[Y]$ to $k[X]$.*

Proof. This follows immediately from the definition $\phi^*(g) = g \circ \phi$. □

References

[1] J.H. Silverman, *The arithemetic of elliptic curves*, 2nd edition, Springer, 2009.

[7] Proving this equivalence is not completely trivial; see [1, Ex. I.1.12a].

18.782 Introduction to Arithmetic Geometry **Fall 2013**
Lecture #15 **10/29/2013**

As usual, k is a perfect field and $\bar{k}$ is a fixed algebraic closure of k. Recall that an affine (resp. projective) variety is an irreducible alebraic set in $\mathbb{A}^n = \mathbb{A}^n(\bar{k})$ (resp. $\mathbb{P}^n = \mathbb{P}^n(\bar{k})$).

15.1 Rational maps of affine varieties

Before defining rational maps we want to nail down two points on which we we were intentional vague in the last lecture. We defined a morphism $\phi\colon X \to Y$ of varieties $X \subseteq \mathbb{A}^m$ and $Y \subseteq \mathbb{A}^n$ as a "map defined by a tuple of polynomials $(\phi_1, \ldots, \phi_n)$." This definition is vague in two ways. First, is a morphism a map between two sets X and Y, or is it a tuple of polynomials? We shall adopt the second view; we still get a function by evaluating the polynomials at points in X.

Given that we regard ϕ as a tuple $(\phi_1, \ldots, \phi_n)$, the next question is in which ring do its components ϕ_i lie? Are they elements of $\bar{k}[x_1, \ldots, x_m]$ or $\bar{k}[X]$. The function they define is the same in either case, but we shall regard the ϕ_i as elements of $\bar{k}[X]$. This means that in order to evaluate ϕ_i at a point $P \in X$, we need to lift $\phi_i = \hat{\phi}_i + I(X)$ to a representative $\hat{\phi}_i \in \bar{k}[x_1, \ldots, x_m]$ and then compute $\hat{\phi}_i(P)$. Of course it does not matter which representative $\hat{\phi}_i$ we pick; we define $\phi_i(P)$ to be the value $\hat{\phi}_i(P)$ for any/every choice of $\hat{\phi}$, and thereby define $\phi(P)$ for $P \in X$ (but note that $\phi(P)$ is *not defined* for any $P \notin X$).

One advantage of this approach is that there is then a one-to-one correspondence between morphisms and the functions they define. If $\phi = (\phi_1, \ldots, \phi_n)$ and $\psi = (\psi_1, \ldots, \psi_n)$ define the same function from X to Y then each of the polynomials $\hat{\phi}_i - \hat{\psi}_i$ in $\bar{k}[x_1, \ldots, x_m]$ contains X in its zero locus and therefore lies in the ideal $I(X)$. This implies, by definition, that in $\bar{k}[X] = \bar{k}[x_i, \ldots, x_m]/I(X)$ we have $\phi_i = \psi_i$ for $1 \le i \le m$ and therefore $\phi = \psi$.

We now want to extend these ideas to the function field $\bar{k}(X)$. Elements of $\bar{k}(X)$ have the form $r = f/g$, with $f, g \in \bar{k}[X]$ and $g \ne 0$, and are called *rational functions* (or even just *functions*), on X, even though they are formally elements of the fraction field of $\bar{k}[X]$ and typically do *not* define a function from X to $\bar{k}$; indeed, this is precisely the issue we must now address.

It seems natural to say that for a point $P \in X$ we should define $r(P)$ to be $f(P)/g(P)$ whenever $g(P) \ne 0$ and call it undefined otherwise. But there is a problem with this approach: the representation $r = f/g$ is not necessarily unique.[1] We also have $r = p/q$ whenever $pg = fq$ holds in $\bar{k}[X]$ (recall that this is part of the definition of a fraction field, it is a set of equivalence classes of fractions). The values $f(P)/g(P)$ and $p(P)/q(P)$ are necessarily equal wherever both denominators are nonzero, but it may be that $q(P) \ne 0$ at points where $g(P) = 0$ (and vice versa).

Example 15.1. Consider the the zero locus X of $x_1x_2 - x_3x_4$ in $\mathbb{A}^4$ (which is in fact a variety) and the rational function $r = x_1/x_3 = x_4/x_2$. At the point $P = (0, 1, 0, 0) \in X$ the value $x_1(P)/x_3(P)$ is not defined, but $x_4(P)/x_2(P) = 0$ is defined, and the reverse occurs for $P = (0, 0, 1, 0) \in X$. But we can assign a meaningful value to $r(P)$ at both points; the only points in X where r is not defined are those with $x_2 = x_3 = 0$.

This motivates the following definition.

[1] If $\bar{k}[X]$ is a UFD then we can put $r = f/g$ in lowest terms to get a unique representation. However, the coordinate ring $\bar{k}[X]$ is usually *not* a UFD, even though $\bar{k}[x_1, \ldots, x_m]$ is. A quotient of a UFD is typically not a UFD, even when it is an integral domain; consider $\mathbb{Z}[x]/(x^2 + 5)$, for example.

Andrew V. Sutherland

Definition 15.2. A function $r \in \bar{k}(X)$ is said to be *regular* at a point $P \in X$ if $gr \in \bar{k}[X]$ for some $g \in \bar{k}[X] \subseteq \bar{k}(X)$[2] for which $g(P) \neq 0$ (we then have $r = f/g$ for some $f \in \bar{k}[X]$).

The set of points at which a function $r \in \bar{k}(X)$ is regular form a nonempty open (hence dense) subset $\operatorname{dom}(r)$ of the subspace X: the complement of $\operatorname{dom}(r)$ in X is the closed subset of X defined by the *denominator ideal* $\{g \in \bar{k}[X] : gr \in \bar{k}[X]\}$, which we note is not the zero ideal, since $r = f/g$ for some nonzero g.[3]

We now associate to r the function from $\operatorname{dom}(r)$ to $\bar{k}$ that maps P to

$$r(P) = (f/g)(P) = f(P)/g(P),$$

where g is chosen so that $g(P) \neq 0$ and $gr = f \in \bar{k}[X]$. Now if r is actually an element of $\bar{k}[X]$, then r is regular at every point in X and we have $\operatorname{dom}(r) = X$. The following lemma says that the converse holds.

Lemma 15.3. *A rational function $r \in \bar{k}(X)$ lies in $\bar{k}[X]$ if and only if $\operatorname{dom}(r) = X$.*

Proof. The forward implication is clear, and if $\operatorname{dom}(r) = X$ then the complement of $\operatorname{dom}(r)$ in X is the empty set and the denominator ideal is (1), which implies $r \in \bar{k}[X]$. □

Definition 15.4. Let $X \subseteq \mathbb{A}^m$ and $Y \subseteq \mathbb{A}^n$ be affine varieties. We say that a tuple $(\phi_1, \ldots, \phi_n)$ with $\phi_i \in \bar{k}(X)$ is *regular* at $P \in X$ if the ϕ_i are all regular at P. A *rational map* $\phi\colon X \to Y$ is a tuple $(\phi_1, \ldots, \phi_n)$ with $\phi_i \in \bar{k}(X)$ such that $\phi(P) := (\phi_1(P), \ldots, \phi_n(P)) \in Y$ for all points $P \in X$ where ϕ is regular. If ϕ is regular at every point $P \in X$ then we say that ϕ is *regular*.

The set of points where ϕ is regular form an open subset $\operatorname{dom}(\phi) = \cap_i \operatorname{dom}(\phi_i)$ of X. Thus ϕ defines a function from $\operatorname{dom}(\phi)$ to Y, which we may also regard as a partial function from X to Y. We get a complete function from X to Y precisely when $\operatorname{dom}(\phi) = X$, that is, when ϕ is regular. This occurs precisely when ϕ is a morphism.

Theorem 15.5. *A rational map of affine varieties is a morphism if and only if it is regular.*

Proof. A morphism is clearly a regular rational map. For the converse, apply Lemma 15.3 to each component of $\phi = (\phi_1, \ldots, \phi_n)$. □

We now want to generalize the categorical equivalence between affine varieties and their function fields, analogous to what we proved in the last lecture for affine varieties and their coordinate rings (affine algebras), but with morphisms of varieties replaces by the more general notion of a rational map. But there is a problem with this. In order to even define a category of varieties with rational maps, we need to be able to compose rational maps. But this is not always possible!

Example 15.6. Let $X = Y = Z = \mathbb{A}^2$, and let $\phi\colon X \to Y$ be the rational map $(1/x_1, 0)$ and let $\psi\colon Y \to Z$ be the rational map $(0, 1/x_2)$. Then the image of ϕ is disjoint from $\operatorname{dom}(\psi)$. There is no rational function that corresponds to the composition of ϕ and ψ (or ψ with ϕ). Even formally, the fractions $1/0$ that we get by naively composing ϕ with ψ are not elements of $\bar{k}(X)$, and the function defined by the composition of the functions defined by ϕ and ψ has the empty set as its domain, which is not true of any $r \in \bar{k}(X)$.

[2] Recall that an integral domain can always be embedded in its fraction field by identifying g with the equivalence class of $g/1$, so we assume $\bar{k}[X] \subseteq \bar{k}(X)$ henceforth.

[3] When restricting our attention to a variety X in $\mathbb{A}^n$ it is simpler to work with ideals in $\bar{k}[X] = \bar{k}[x_1, \ldots, x_m]/I(X)$ rather than $\bar{k}[x_1, \ldots, x_m]$. The one-to-one corrsespondence between radical ideals and closed sets still holds, as does the correspondence between prime ideals and (sub-) varieties.

To fix this problem we want to restrict our attention to rational maps whose image is dense in its codomain.

Definition 15.7. A rational map $\phi\colon X \to Y$ is *dominant* if $\overline{\phi(\mathrm{dom}(\phi))} = Y$.

If $\phi\colon X \to Y$ and $\psi\colon Y \to Z$ are dominant rational maps then the intersection of $\phi(\mathrm{dom}(\phi))$ and $\mathrm{dom}(\psi)$ must be nonempty; the complement of $\mathrm{dom}(\psi)$ in Y is a proper closed subset of Y and therefore contains no sets that are dense in Y, including $\phi(\mathrm{dom}(\phi))$. It follows that we can always compose dominant rational maps, and since the identity map is also dominant rational map, we can now speak of the category of affine varieties and rational maps. Not every morphism is a dominant rational map, but affine varieties and dominant morphisms form a subcategory of affine varieties and dominant rational maps. As you will show on the problem set, the closure of the image of a morphism of varieties is a variety, so one can always make a morphism dominant be restricting its codomain.

We now prove the analog of Theorem 14.8, replacing morphisms with dominant rational maps, and coordinate rings with function fields. We now use the term function field to refer to any finitely generated extension of $\bar{k}$, and we require morphisms of function fields to fix $\bar{k}$ (we could also call them $\bar{k}$-algebra homomorphisms). Field homomorphisms are always injective, so a morphism of function fields is just a field embedding that fixes $\bar{k}$. Note that the all the interesting function fields $F/\bar{k}$ are transcendental. If $F/\bar{k}$ is algebraic then $F = \bar{k}$; this corresponds to the function field of a zero-dimensional variety (a point).

Given a function field F generated by elements $\alpha_1, \ldots, \alpha_n$ over $\bar{k}$, let R be the $\bar{k}$-algebra generated by $\alpha_1, \ldots, \alpha_n$ in F; this means that R is equal to the set of all polynomial expressions in $\alpha_1, \ldots, \alpha_n$, but there may be algebraic relations between the α_i that make many of these expressions equivalent. In any case, R is isomorphic to the quotient of the polynomial ring $\bar{k}[x_1, \ldots, x_n]$ by an ideal I corresponding to all the algebraic relations that exist among the α_i. The ring R is an integral domain (since it is contained in a field), therefore I is a prime ideal that defines a variety X whose coordinate ring is isomorphic to R and whose function field is isomorphic to F, the fraction field of R.

Theorem 15.8. *The following hold:*

(i) Every dominant rational map $\phi\colon X \to Y$ of affine varieties, induces a morphism $\phi^*\colon \bar{k}(Y) \to \bar{k}(X)$ of function fields such that $\phi^*(r) = r \circ \phi$.

(ii) Every morphism $\theta\colon K \to L$ of function fields induces a dominant rational map of affine varieties $\theta^*\colon X \to Y$, with $K \simeq \bar{k}(Y)$ and $L \simeq \bar{k}(X)$, such that the image of $\theta(r)$ in $\bar{k}(X)$ is $r \circ \theta^*$.

(iii) If $\phi\colon X \to Y$ and $\psi\colon Y \to Z$ are dominant rational maps of affine varieties then $(\psi \circ \phi)^* = \phi^* \circ \psi^*$.

As in the analogous Theorem 14.8, to compute $r \circ \phi$ one needs to lift/compose/reduce, that is, pick representatives for $\phi_1, \ldots, \phi_n$ that are rational functions in $\bar{k}(x_1, \ldots, x_m)$, pick a representative of r in $\bar{k}(y_1, \ldots, y_n)$, then compose and reduce the numerator and denominator modulo $I(X)$. The fact that ϕ is dominant ensures that the denominator of the composition does not lie in $I(X)$, so $r \circ \phi$ is an element of $\bar{k}(X)$.

Proof. The proofs of parts (i) and (iii) follow the proof of Theorem 14.8 verbatim, with coordinate rings replaced by function fields, only part (ii) merits further discussion. As discussed above, we can write $K \simeq \bar{k}(Y)$ and $L \simeq \bar{k}(X)$ for some varieties X and Y, and any morphism $K \to L$ induces a morphism $\bar{k}(Y) \to \bar{k}(Y)$ that is compatible with these isomorphisms, so let us assume $\theta\colon \bar{k}(Y) \to \bar{k}(X)$.

As in the proof of Theorem 14.8 we define $\theta^*\colon X \to Y$ by $\theta^* = (\theta(y_1), \ldots, \theta(y_n))$, where we now regard the coordinate functions y_i as elements of $\bar{k}(Y)$. For any $r \in \bar{k}(Y)$ we have

$$r \circ \theta^* = \hat{r}(\theta(y_1), \ldots, \theta(y_n)) = \theta(r(y_1, \ldots, y_n)) = \theta(r).$$

The fact that $r \in \bar{k}(Y)$ ensures that the denominator of $\hat{r} \in \bar{k}[y_1, \ldots, y_n]$ is not in $I(Y)$, so this composition is well defined. The proof that the image of θ^* actually lies in Y is the same as in Theorem 14.8: for any $f \in I(Y)$ we have $f \circ \theta^* = \theta(f) = 0$, so certainly $f(\theta^*(P)) = 0$ for all $P \in \operatorname{dom}(\theta^*)$. Thus θ^* is a rational map from X to Y.

But we also need to check that θ^* is dominant (this is the only new part of the proof). This is equivalent to showing that the only element of $\bar{k}[Y]$ that vanishes on the image of θ^* is the zero element, which is in turn equivalent to showing that the only element of $\bar{k}(Y)$ that vanishes on the intersection of its domain and the image of θ^* is the zero element. This is in turn equivalent to showing that if $r \circ \theta^* = \theta(r)$ vanishes at every point in its domain then $r = 0$. But the only element of $\bar{k}(X)$ that vanishes at every point in its domain is the zero element, and θ is injective, so we are done. □

Corollary 15.9. *The category of affine varieties with dominant rational maps and the category of function fields are contravariantly equivalent.*

Proof. As in the proof of Corollary 14.9, the only thing left to show is that $(\phi^*)^* = \phi$ and $(\theta^*)^* = \theta$, up to isomorphism, but both follow from Theorem 15.8 and its proof. □

Definition 15.10. Two affine varieties X and Y are said to be *birationally equivalent* if there exist dominant rational maps $\phi\colon X \to Y$ and $\psi\colon Y \to X$ such that $(\phi \circ \psi)(P) = P$ for all $P \in \operatorname{dom}(\phi \circ \psi)$ and $(\psi \circ \phi)(P) = P$ for all $P \in \operatorname{dom}(\psi \circ \phi)$.

Corollary 15.11. *Two affine varieties are birationally equivalent if and only if their function fields are isomorphic.*

As with morphisms, if $\phi\colon X \to Y$ is a rational map of varieties that are defined over k, we say that $\phi = (\phi_1, \ldots, \phi_n)$ is defined over k if the ϕ_i all lie in $k(X)$.

Corollary 15.12. *Let X and Y be affine varieties defined over k. If $\phi\colon X \to Y$ is a dominant rational map defined over k then $\phi^*\colon \bar{k}(Y) \to \bar{k}(X)$ restricts to a morphism $k(Y) \to k(X)$.*

15.2 Morphisms and rational maps of projective varieties

We now want to generalize everything we have done for maps between affine varieties to maps between projective varieties. This is completely straight-forward, we just need to account for the equivalence relation on $\mathbb{P}^n$.

Recall from Lecture 13 that although we defined the function field $\bar{k}(X)$ of a projective variety $X \subseteq \mathbb{P}^n$ as the function field of any of its non-empty affine parts, we can always represent r by an element $\hat{r} \in \bar{k}(x_0, \ldots, x_n)$ whose numerator and denominator are homogeneous polynomials of the same degree, and for any point $P \in X$ where $\hat{r}(P)$ is defined (has nonzero denominator), we can unambiguously define $r(P) = \hat{r}(P)$.

Definition 15.13. Let X be a projective variety. We say that $r \in \bar{k}(X)$ is *regular* at a point $P \in X$ if it has a representation $\hat{r}$ that is defined at P. The set of points $P \in X$ at which r is regular form an open subset of X that we denote $\operatorname{dom}(r)$.[4] For any point $P \in \operatorname{dom}(r)$ we define $r(P) = \hat{r}(P)$, where $\hat{r}$ is chosen so that $\hat{r}$ is defined at P.

Definition 15.14. Let $X \subseteq \mathbb{P}^m$ and $Y \subseteq \mathbb{P}^n$ be projective varieties. A *rational map* $\phi\colon X \to Y$ is an equivalence class of tuples $\phi = (\phi_0 : \ldots : \phi_n)$ with $\phi_i \in \bar{k}(X)$ not all zero such that at any point $P \in X$ where all the ϕ_i are regular and at least one is nonzero, the point $(\phi_0(P) : \ldots : \phi_n(P))$ lies in Y. The equivalence relation is given by

$$(\phi_0 : \ldots : \phi_n) = (\lambda\phi_0 : \ldots : \lambda\phi_n)$$

for any $\lambda \in \bar{k}(X)^\times$. We say that ϕ is *regular* at P if there is a tuple $(\lambda\phi_0 : \ldots : \lambda\phi_n)$ in its class with each component regular at P and at least one nonzero at P. The open subset of X at which ϕ is regular is denoted $\operatorname{dom}(\phi)$.

Remark 15.15. We can alternatively represent the rational map $\phi\colon X \to Y$ as a tuple of homogeneous polynomials in $\bar{k}[x_0, \ldots, x_m]$ that all have the same degree and not all of which lie in $I(X)$. To ensure that the image lies in Y one requires that for all $f \in I(Y)$ we have $f(\phi_0, \ldots, \phi_n) \in I(X)$. The equivalence relation is then $(\phi_0 : \ldots : \phi_n) = (\psi_0 : \ldots : \psi_n)$ if and only if $\phi_i\psi_j - \phi_j\psi_i \in I(X)$ for all $i.j$.

Remark 15.16. One can also define rational maps $X \to Y$ where one of X, Y is an affine variety and the other is a projective variety. When Y is projective the definition is exactly the same as in the case that both are projective (but we don't use homegenized functions to represent elements of $\bar{k}(X)$ when X is affine). When X is projective and Y is affine, a rational map is no longer an equivalence class of tuples, it is a particular tuple of rational functions on X. Of course there is still a choice of representation for each rational function (and the choice may vary with P), but note that in this case Remark 15.15 *no longer applies.*

Now that we have defined rational maps for projective varieties we can define morphisms and dominant rational maps; the definitions are exactly the same as in the affine case, so we can now state them generically.

Definition 15.17. A *morphism* is a regular rational map. A rational map is *dominant* if its image is dense in its codomain.

The analogs of Theorem 15.8 and Corollary 15.9 both apply to dominant rational maps between projective varieties. The proofs are exactly the same, modulo the equivalence relations for projective points and rational maps between projective varieties. Alternatively, one can simply note that any dominant rational map $X \to Y$ of projective varieties restricts to a dominant rational map between any pair of the nonempty affine parts of X and Y (the nonempty affine parts of X (resp. Y) are all dense in X (resp. Y), and they are all isomorphic as affine varieties; see Corollary 14.10). Conversely, any dominant rational map of affine varieties can be extended to a dominant rational map of their projective closures (but this is *not* true of morphisms; see Example 15.20 below).

Theorem 15.18. *Theorem 15.8 and Corollary 15.9 hold for projective varieties as well as affine varieties.*

[4] It is clear that $\operatorname{dom}(r)$ is open in X; its intersection with each affine patch is an open subset of X.

Let us now look at a couple of examples.

Example 15.19. Let $X \subseteq \mathbb{A}^2$ be the affine variety defined by $x^2 + y^2 = 1$, and let P be the point $(-1, 0) \in X$. The rational map $\phi\colon X \to \mathbb{A}^1$ defined by

$$\phi(x, y) = \left(\frac{y}{x+1}\right) = \left(\frac{1-x}{y}\right)$$

sends each point $Q = (x, y) \in X$ different from P to the slope of the line $\overline{PQ}$. The map ϕ is not regular (hence not a morphism), because it is not regular at P, but it is dominant (even surjective). The rational map $\phi^{-1}\colon \mathbb{A}^1 \to X$ defined by

$$\phi^{-1}(t) = \left(\frac{1-t^2}{1+t^2}, \frac{2t}{1+t^2}\right)$$

is an inverse to ϕ. Note that ϕ^{-1} is also not regular (it is not defined at $\sqrt{-1}$), but it is dominant (but not surjective). Thus X is birationally equivalent to $\mathbb{A}^1$, but not isomorphic to $\mathbb{A}^1$, as expected. The function fields of X and $\mathbb{A}^1$ are both isomorphic to $\bar{k}(t)$.

Now let us consider the corresponding map of projective varieties. The projective closure $\overline{X}$ of X in $\mathbb{P}^2$ is defined by $x^2 + y^2 = z^2$. We now define the rational map $\varphi\colon \overline{X} \to \mathbb{P}^1$ by

$$\varphi(x : y : z) = \left(\frac{y}{x+z} : 1\right) = \left(\frac{z-x}{y} : 1\right) = \left(1 : \frac{y}{z-x}\right)$$

Per Remark 15.15, we could also write φ as

$$\varphi(x : y : z) = (y : x + z) = (z - x : y)$$

The first RHS is defined everywhere except $(1 : 0 : -1)$ and the second RHS is defined everywhere except $(1 : 0 : 1)$, thus φ is regular everywhere, hence a morphism.

We also have the rational map $\varphi\colon \mathbb{P}^1 \to \overline{X}$ defined by

$$\varphi^{-1}(s : t) = \left(\frac{s^2-t^2}{s^2+t^2} : \frac{2st}{s^2+t^2} : 1\right) = \left(1 : \frac{2st}{s^2-t^2} : \frac{s^2+t^2}{s^2-t^2}\right)$$

which can also be written as

$$\varphi^{-1}(s : t) = (s^2 - t^2 : 2st : s^2 + t^2).$$

The map φ^{-1} is regular everywhere, hence a morphism, and the compositions $\varphi \circ \varphi^{-1}$ and $\varphi^{-1} \circ \varphi$ are both the identity maps, thus $\overline{X}$ and $\mathbb{P}^1$ are isomorphic.

Example 15.20. Recall the morphism $\phi\colon \mathbb{A}^2 \to \mathbb{A}^2$ defined by $\phi(x, y) = (x, xy)$ from Lecture 14, where we noted that the image of ϕ is not closed (but it is dense in $\mathbb{A}^2$, so ϕ is dominant). Let us now consider the corresponding rational map $\varphi\colon \mathbb{P}^2 \to \mathbb{P}^2$ defined by

$$\varphi(x : y : z) = \left(\frac{x}{z} : \frac{xy}{z^2} : 1\right) = (xz : xy : z^2).$$

We might expect φ to be a morphism, but this is not the case! It is not regular at $(0 : 1 : 0)$.

This is not an accident. As we will see in the next lecture, morphisms of projective varieties are *proper*, and in particular this means that they are closed maps (so unlike the affine case, the image of a morphism of projective varieties *is* a variety). But there is clearly no way to extend the morphism $\phi\colon \mathbb{A}^2 \to \mathbb{A}^2$ to a proper morphism $\varphi\colon \mathbb{P}^2 \to \mathbb{P}^2$ (the image of φ in the affine patch $z \neq 0$ must be dense but not surjective), and this means that φ cannot be a morphism.

Our goal for this lecture is to prove that morphisms of projective varieties are closed maps. In fact we will prove something stronger, that projective varieties are *complete*, a property that plays a role comparable to compactness in topology. For varieties, compactness as a topological space does not mean much because the Zariski topology is so coarse. Indeed, *every* subset of $\mathbb{A}^n$ (and hence of $\mathbb{P}^n$) is compact (or quasicompact, if your definition of compactness includes Hausdorff).

Theorem 16.1. *Let S be a subset of $\mathbb{A}^n$, and let $\{U_a\}_{a\in A}$ be any collection of open sets of $\mathbb{A}^n$ whose union contains S. Then there exists a finite set $B \subseteq A$ for which $S \subseteq \{U_b\}_{b\in B}$.*

Proof. By enumerating the index set A in some order (which we can do, via the axiom of choice), we can construct a chain of properly nested open sets $\{V_b\}_{b\in B}$, where each V_b is the union of the sets U_a over $a \in B$ with $a \leq b$ (in our arbitrary ordering), and $B \subseteq A$ is constructed so that each $S \cap V_a$ is properly contained in $S \cap V_b$ for every pair $a \leq b$ in B. The complements of the sets V_b then form a strictly descending chain of closed sets whose ideals form a strictly ascending chain of nested ideals $\{I_b\}_{b\in B}$ in $R = k[x_1, \ldots, x_n]$. The ring R is Noetherian, so B must be finite, and $\{U_b\}_{b\in B}$ is the desired finite subcover. □

In order to say what it means for a variety to be complete, we first need to define the product of two varieties. Throughout this lecture k denotes a fixed algebraically closed field.

16.1 Products of varieties

Definition 16.2. Let $X \subseteq \mathbb{A}^m$ and $Y \subseteq \mathbb{A}^n$ be algebraic sets. Let $k[\mathbb{A}^m] = k[x_1, \ldots, x_m]$, $k[\mathbb{A}^n] = k[y_1, \ldots, y_n]$, and $k[\mathbb{A}^{m+n}] = k[x_1, \ldots, x_m, y_1, \ldots, y_n]$, so that we can identity $k[\mathbb{A}^m]$ and $k[\mathbb{A}^n]$ as subrings of $k[\mathbb{A}^{m+n}]$ whose intersection is k. The *product* $X \times Y$ is the zero locus of the ideal $I(X)k[\mathbb{A}^n] + I(Y)k[\mathbb{A}^m]$ in $k[\mathbb{A}^{m+n}]$.

If $I(X) = (f_1, \ldots, f_s)$ and $I(Y) = (g_1, \ldots, g_t)$, then $I(X \times Y) = (f_1, \ldots, f_s, g_1, \ldots, g_t)$ is just the ideal generated by the f_i and g_j when regarded as elements of $k[\mathbb{A}^{m+n}]$. We also have *projection morphisms*

$$\pi_X \colon X \times Y \to X \quad \text{and} \quad \pi_Y \colon X \times Y \to Y$$

defined by the tuples $(\bar{x}_1, \ldots, \bar{x}_m)$ and $(\bar{y}_1, \ldots, \bar{y}_n)$, where $\bar{x}_i$ and $\bar{y}_j$ are the images of x_i and y_j, respectively, under the quotient map $k[\mathbb{A}^{m+n}] \to k[\mathbb{A}^{m+n}]/I(X \times Y) = k[X \times Y]$.

The coordinate ring of $X \times Y$ is isomorphic to the *tensor product* of the coordinate rings of X and Y, that is

$$k[X \times Y] \simeq k[X] \otimes k[Y].$$

While the tensor product can be defined quite generally in categorical terms, in the case of k-algebras there is a very simple concrete definition. Recall that a k-algebra is, in particular, a k-vector space. If R and S are two k-algebras with bases $\{r_i\}_{i\in I}$ and $\{s_j\}_{j\in J}$, then the set of formal symbols $\{r_i \otimes s_j : i \in I, j \in J\}$ forms a basis for the tensor product $R \otimes S$. Products of vectors in $R \otimes S$ are computed via the distributive law and the rule

$$(r_{i_1} \otimes s_{j_1})(r_{i_2} \otimes s_{j_2}) = r_{i_1}r_{i_2} \otimes s_{j_1}s_{j_2}.$$

In the case of polynomial rings one naturally chooses a monomial basis, in which case this rule just amounts to multiplying monomials and keeping the variables in the monomials separated according to which polynomial ring they originally came from.

It is standard to generalize the $\otimes$ notation and write $r \otimes s$ for any $r \in R$ and $s \in S$, not just basis elements, with the understanding that $r \otimes s$ represents a linear combination of basis elements $\sum_{i,j} \alpha_{ij}(r_i \otimes s_j)$ that can be computed by applying the identities

$$\begin{aligned}(a+b)\otimes c &= a\otimes c + b\otimes c\\ a\otimes(b+c) &= a\otimes b + a\otimes c\\ (\gamma a)\otimes(\delta b) &= (\gamma\delta)(a\otimes b)\end{aligned}$$

where γ and δ denote elements of the field k. We should note that most elements of $R \otimes S$ are *not* of the form $r \otimes s$, but they can all be written as finite sums of elements of this form.

When R and S are commutative rings, so is $R \otimes S$. There are then natural embeddings of R and S into $R \otimes S$ given by the maps $r \to r \otimes 1_S$ and $s \to 1_R \otimes s$, and $1_R \otimes 1_S$ is the multiplicative identity in $R \otimes S$. The one additional fact that we need is that if R and S are affine algebras (finitely generated k-algebras that are integral domains), so is $R \otimes S$. In order to prove this we first note a basic fact that we will use repeatedly:

Lemma 16.3. *Let V be an affine variety with coordinate ring $k[V]$. There is a one-to-one correspondence between the maximal ideals of $k[V]$ and the points of V.*

Proof. Let $P = (a_1, \ldots, a_n)$ be a point on $V \subseteq \mathbb{A}^n$, and let m_P be the corresponding maximal ideal $(x_1 - a_1, \ldots, x_n - a_n)$ of $k[x_1, \ldots, x_n]$. Then $I(V) \subseteq m_P$, and the image of m_P in the quotient $k[V] = k[x_1, \ldots, x_n]/I(V)$ is a maximal ideal of $k[V]$. Conversely, every maximal ideal of $k[V]$ corresponds to a maximal ideal of $k[x_1, \ldots, x_n]$ that contains $I(V)$, which is necessarily of the form m_P for some $P \in V$, by Hilbert's Nullstellensatz. □

Lemma 16.4. *If R and S are both affine algebras, then so is $R \otimes S$.*

Proof. We need to show that $R \otimes S$ has no zero divisors. So suppose $uv = 0$ for some $u, v \in R \otimes S$. We will show that either $u = 0$ or $v = 0$.

We can write u and v as finite sums $u = \sum_{i \in I} r_i \otimes s_i$ and $v = \sum_{j \in J} r_j \otimes s_j$, with $r_i, r_j \in R$ and $s_i, s_j \in S$ all nonzero, and we can assume the sets $\{s_i\}_{i \in I}$ and $\{s_j\}_{j \in J}$ are each linearly independent over k by choosing the s_i and s_j to be basis vectors. Without loss of generality, we may assume $R = k[X]$, for some affine variety X. Let X_u be the zero locus of the r_i in X and and let X_v be the zero locus of the r_j in X. For any point $P \in X$ we have the evaluation map $\phi_P \colon k[X] \to k$ defined by $\phi_P(f) = f(P)$, which is a ring homomorphism from R to k that fixes k. We now extend ϕ_P to a k-algebra homomorphism $R \otimes S \to S$ by defining $\phi_P(r \otimes s) = \phi_P(r)s$. We then have

$$\phi_P(uv) = \phi_P(u)\phi_P(v) = \left(\sum_{i \in I} \phi_P(r_i)s_i\right)\left(\sum_{j \in J} \phi_P(r_j)s_j\right) = 0$$

Since S is an integral domain, one of the two sums must be zero, and since the s_i are linearly independent over k, either $\phi_P(r_i) = 0$ for all the r_i, in which case $P \in X_u$, or $\phi_P(r_j) = 0$ for all the r_j, in which case $P \in X_v$. Thus $X = X_u \cup X_v$. But X is irreducible, so either $X = X_u$, in which cace $u = 0$, or $X = X_v$, in which case $v = 0$. □

Corollary 16.5. *If X and Y are affine varieties, then so is $X \times Y$.*

Remark 16.6. This proof is a nice example of the interaction between algebra and geometry. We want to prove a geometric fact (a product of varieties is a variety), but it is easier to prove an algebraic fact (a tensor product of affine algebras is an affine algebra). But in order to prove the algebraic fact, we use a geometric fact (a variety is not the union of two proper algebraic subsets). Of course we could translate everything into purely algebraic or purely geometric terms, but the proofs are easier to construct (and easier to understand!) when we can move back and forth freely.

A product of projective varieties is defined similarly, but there is a new wrinkle; we now need two distinct sets of homogeneous coordinates. Points in $\mathbb{P}^m \times \mathbb{P}^n$ can be represented in the form $(a_0 : \ldots : a_m; b_0 : \ldots : b_n)$, where

$$(a_0 : \ldots : a_m; b_0 : \ldots : b_n) = (\lambda a_0 : \ldots : \lambda a_m; \mu b_0 : \ldots : \mu b_n)$$

for all $\lambda, \mu \in k^\times$. We are now interested in polynomials in $k[x_0, \ldots, x_m, y_0, \ldots, y_n]$ that are homogeneous in the x_i, and in the y_j, but not necessarily both. Another way of saying this is that we are interested in polynomials that are homogeneous as elements of $(k[x_0, \ldots, x_m])[y_0, \ldots, y_n]$, and as elements of $(k[y_0, \ldots, y_n])[x_0, \ldots, x_m]$. Let us call such polynomials (m, n)-*homogeneous.* We can then meaningfully define the zero locus of an (m, n)-homogeneous polynomial in $\mathbb{P}^m \times \mathbb{P}^n$ and give $\mathbb{P}^m \times \mathbb{P}^n$ the Zariski topology by taking algebraic sets to be closed.

Remark 16.7. The Zariski topology on $\mathbb{P}^m \times \mathbb{P}^n$ we have just defined is *not* the product of the Zariski topologies on $\mathbb{P}^m$ and $\mathbb{P}^n$. This will be explored on the problem set.

Definition 16.8. Let $X \subseteq \mathbb{P}^m$ and $Y \subseteq \mathbb{P}^n$ be algebraic sets with homogeneous ideals $I(X) \subseteq k[x_0, \ldots, x_m]$ and $I(Y) \subseteq k[y_0, \ldots, y_n]$. The *product* $X \times Y$ is the zero locus of the (m, n)-homogeneous polynomials in the ideal

$$I(X \times Y) := I(X)k[y_0, \ldots, y_n] + I(Y)k[x_0, \ldots, x_m]$$

of $k[x_0, \ldots, x_m, y_0, \ldots, y_n]$. We say that $X \times Y$ is a *variety* if the ideal $I(X \times Y)$ is prime.

As in the affine case, we again have $k[X \times Y] = k[X] \otimes k[Y]$, which implies that the product of two projective varieties is again a variety.

Remark 16.9. One can identify $\mathbb{P}^m \times \mathbb{P}^n$ with a subvariety of a larger projective space $\mathbb{P}^N$ (but N is definitely not $m + n$). Thus the product of two projective varieties is indeed a projective variety. This will be explored on the next problem set.

We may also consider products of affine and projective varieties. In this case we are interested in subsets of $\mathbb{P}^m \otimes \mathbb{A}^n$ that are the zero locus of polynomials in $k[x_0, \ldots, x_m, y_1, \ldots, y_n]$ that are homogeneous in x_i but may be inhomogeneous in the y_j. Per the remark above, we can smoothly embed a product of projective varieties in a single projective variety, and as we have already seen we can smoothly embed a product of affine varieties in a single affine variety. Thus any finite product of affine and projective varieties is isomorphic to one of (1) an affine variety, (2) a projective variety, (3) the product of a affine variety and a projective variety.

16.2 Complete varieties

We can now say what it means for a variety to be complete.

Definition 16.10. A variety X is *complete* if for every variety Y the projection $X \times Y \to Y$ is a *closed map*; this means that the projection of a closed set in $X \times Y$ is a closed in Y.

Remark 16.11. We get the same definition if we restrict to affine varieties Y. Any variety Y can be covered by a finite number of affine parts $\{U_i\}$, and if the projection $X \times U_i \to U_i$ is a closed map for each U_i, then the projection $X \times Y \to Y$ is also a closed map, since the union of a finite number of closed sets is a closed set.

Lemma 16.12. *If X is a complete variety then any morphism $\phi\colon X \to Y$ is a closed map whose image is a complete variety.*

Proof. Let us consider the set

$$\Gamma_\phi := \{(P, \phi(P)) : P \in X\} \subseteq X \times Y,$$

which is the *graph* of ϕ. It is a closed set, the zero locus of $y = \phi(x)$ (here the variables x and y represent points in X and Y that may have many coordinates; the exact equation can be explicitly spelled out in the ambient space containing $X \times Y$ using generators for $I(X)$, $I(Y)$, and the coordinate maps of ϕ but there is no need to do so). The projection map $X \times Y \to Y$ is a closed map, since X is complete, so $\operatorname{im}(\phi)$ is a closed subset of Y, and it must be irreducible, since it is the image of a variety. Similarly, if Z is any closed set in X, by considering the graph of the restriction of ϕ to Z and applying the fact that X is complete we can show that $\phi(Z)$ is closed. Thus ϕ is a closed map.

We now show that $\phi(X)$ is complete. So let Z be any variety and consider the projection $\phi(X) \times Z \to Z$. Let us define the morphism $\Phi\colon X \times Z \to Y \times Z$ by $\Phi(P, Q) = (\phi(P), Q)$. If V is a closed set in $\phi(X) \times Z \subseteq Y \times Z$, then its inverse image $\Phi^{-1}(V)$ is closed in $X \times Z$, since Φ is continuous. Since X is complete, the projection of $\Phi^{-1}(V)$ to Z is closed, but this is precisely the projection of V to Z, since the Z-component of Φ is the identity map. □

Lemma 16.13. *If X is complete then so is every subvariety of X.*

Proof. Let $V \subseteq X$ be a variety. For any variety Z the projection $V \times Z \to Z$ is the composition

$$V \times Z \to X \times Z \to Z,$$

where the first map is an inclusion and the second map is a projection, both of which are closed maps. Thus the projection $V \times Z \to Z$ is a closed map and V is complete. □

Theorem 16.14. *Every complete affine variety consists of a single point.*

Proof. We first consider $\mathbb{A}^1$ and the closed set $\{(x, y) : xy = 1\}$ in $\mathbb{A}^1 \times \mathbb{A}^1$. The projection to the second $\mathbb{A}^1$ is $\mathbb{A}^1 - \{0\}$, not a closed set, so the first $\mathbb{A}^1$ is not complete.

Now suppose X is an affine variety of positive dimension and let f be a function in $k[X]$ that does not lie in k; such an f exists since $k(X)$ has positive transcendence degree. The morphism $f\colon X \to \mathbb{A}^1$ that sends P to $f(P)$ most then be dominant, because the dual morphism of affine algebras $k[\mathbb{A}^1] \to k[X]$ is injective; it corresponds to the inclusion $k[f] \subseteq k[X]$ with $k \subsetneq k[f]$. But X is complete, so by Lemma 16.12 the image of $f\colon X \to \mathbb{A}^1$ is a complete variety, and f is dominant, so $\mathbb{A}^1$ is complete, a contradiction.

Thus every complete affine variety has dimension 0 and is therefore a point. □

With one trivial exception, affine varieties are not complete. In contrast, we will prove that every projective variety is complete.

In order to prove this we will apply a theorem of Chevalley that gives a criterion for the completeness of a variety in terms of the *valuation rings* in the function fields of all its subvarieties; this is known as the *valuative criterion* for completeness. But we first take a brief interlude to discuss valuation rings.

16.3 Valuation rings

We have already seen many examples of valuation rings in this course, but let us now formally define the general term.

Definition 16.15. A proper subring R of a field K is a *valuation ring* of K if for every $x \in K^\times$, either $x \in R$ or $x^{-1} \in R$ (possibly both).

Note that a valuation ring R is an integral domain (since it is a subring of a field), and that K is its field of fractions. Given an arbitrary integral domain R that is not a field, we say that R is a valuation ring if it is a valuation ring of its fraction field. In Problem Set 2 you proved that if K is any field with an nonarchimedean absolute value $\| \; \|$, then the set

$$R = \{x \in K : \|x\| \le 1\}$$

is a valuation ring. You also proved that such an R is a *local ring*.

Definition 16.16. A *local ring* is a ring R with a unique maximal ideal $\mathfrak{m}$. The field $R/\mathfrak{m}$ is the *residue field* of R.

Note that fields are included in the definition of a local ring (the unique maximal ideal is the zero ideal), but specifically excluded from the definition of a valuation ring.

Lemma 16.17. *A ring R is a local ring if and only if the set $R - R^\times$ is an ideal.*

Proof. If $R - R^\times$ is an ideal, then it contains every proper ideal and is therefore the unique maximal ideal of R. Conversely, every element of $R - R^\times$ lies in a maximal ideal, and if there is only one such ideal it must equal $R - R^\times$. □

Theorem 16.18. *Every valuation ring is a local ring.*

Proof. Let R be a valuation ring and let $\mathfrak{m} = R - R^\times$. We must show that $\mathfrak{m}$ is an ideal. If $a \notin R^\times$ then $ar \notin R^\times$ for all $r \in R$. So $\mathfrak{m}R \subseteq \mathfrak{m}$. If $a, b \in \mathfrak{m}$ then a/b or b/a lies in R. So $(a/b + 1)b = a + b$ or $(b/a + 1)a = b + a$ lies in $\mathfrak{m}$, hence $\mathfrak{m}$ is an ideal. □

A key property of valuation rings is that their ideals are totally ordered.

Lemma 16.19. *If $\mathfrak{a}$ and $\mathfrak{b}$ are two ideals of a valuation ring R then either $\mathfrak{a} \subseteq \mathfrak{b}$ or $\mathfrak{b} \subseteq \mathfrak{a}$.*

Proof. Suppose not. Then there exist $a \in \mathfrak{a} - \mathfrak{b}$ and $b \in \mathfrak{b} - \mathfrak{a}$, both nonzero. Either a/b or b/a lies in R, so either $(a/b)b = a \in \mathfrak{b}$ or $(b/a)a = b \in \mathfrak{a}$, both of which are contradictions. □

The proof of Lemma 16.19 allows us to compare nonzero elements of R: we have $a/b \in R$ if and only if $(a) \subseteq (b)$. This leads to the following definition.

Definition 16.20. Let R be a valuation ring with fraction field K. The *value group* of R is $\Gamma = K^\times / R^\times$. The *valuation* defined by R is the quotient map $v \colon K^\times \to \Gamma$.

The abelian group Γ is typically written additively, and it follows from Lemma 16.19 that it is totally ordered (its elements are associate classes and their inverses). We have

1. $v(x) = 0$ if and only if $x \in R^\times$,
2. $v(xy) = v(x) + v(y)$,
3. $v(x + y) \geq \min(v(x), v(y))$.

The first two properties are immediate from the definition; the third will be proved on the problem set. For $x \in K^\times$ we then have $v(x) \geq 0$ if and only if x is a nonzero element of R. By convention we extend v to K by defining $v(0) = \infty$, where ∞ is greater than every element of the valuation group Γ. We then have $R = \{x \in K : v(x) \geq 0\}$.[1]

When a valuation ring R is a PID, it is then a UFD with a unique (up to associates) prime element p that generates its maximal ideal. In this case $\Gamma \simeq \mathbb{Z}$, since for nonzero $a \in R$ we can associate $v(a)$ to the largest integer n for which $p^n|a$; this also determines $v(1/a) = -v(a)$. In this situation we say that Γ is *discrete* and call R a *discrete valuation ring*. Recall that earlier we defined discrete valuation rings as local rings that are PIDs but not fields. We will see show that this definition is equivalent, and also precisely characterize the distinctions in the inclusions

$$\text{discrete valuation rings} \quad \subset \quad \text{valuation rings} \quad \subset \quad \text{local rings}$$

Lemma 16.21. *Every finitely generated ideal of a valuation ring is principal.*

Proof. Let $(a_1, \ldots, a_n)$ be a finitely generated ideal of a valuation ring R with n minimal and suppose $n > 1$. We must have $a_1/a_2 \notin R$, else the generator $a_1 = (a_1/a_2)a_2$ is redundant. But then $a_2/a_1 \in R$ and $a_2 = (a_2/a_1)a_2$ is redundant, a contradiction. □

Lemma 16.22. *A local ring is a valuation ring if and only if it is an integral domain that is not a field and all of its finitely generated ideals are principal.*

Proof. The "only if" part of the statement is clear, so let us assume that R is a local ring that satisfies the hypothesis on the right, and let a/b be any element of its fraction field. The ideal (a, b) is finitely generated, hence principal, say $(a, b) = (c)$. Thus for some $d, e, f, g \in R$ we have $a = cd$, $b = ce$, and $c = af + bg = cdf + ceg$, and therefore $df + eg = 1$. If neither d nor e is a unit, then they both lie in the maximal ideal of R and so does 1, a contradiction. So one of d or e is a unit, and therefore one of $a/b = d/e$ and $b/a = e/d$ lies in R. □

The second lemma implies, in particular, that our two definitions of discrete valuation ring are equivalent. Together the two lemmas give a third definition.

Corollary 16.23. *A valuation ring is discrete if and only if it is Noetherian.*

When the fraction field K of a valuation ring R is an extension of a smaller field k that is contained in R, we say that R is a valuation ring of the extension K/k.

[1] Note that for $\Gamma \subseteq \mathbb{R}$ we define $\|x\| = c^{-v(x)}$ for some $c > 0$, so this agrees with $R = \{x \in K : \|x\| \leq 1\}$.

16.4 Localization of a ring at a prime

One of the main ways in which local rings arise is by *localizing* an integral domain at one of its prime ideals.

Definition 16.24. Let R be an integral domain and let $\mathfrak{p}$ be a prime ideal in R. The subring of R's fraction field defined by

$$R_\mathfrak{p} := \{a/b : a, b \in R, b \notin \mathfrak{p}\}$$

is called the *localization* of R at $\mathfrak{p}$.[2]

Remark 16.25. As we saw in Lecture 15, caution is needed when interpreting expressions like a/b in fraction fields of rings that are not necessarily UFDs; $R_\mathfrak{p}$ is a set of equivalence classes, and a/b is just one representative of a particular class. It may happen that the equivalence class a/b lies in $R_\mathfrak{p}$ even though $b \in \mathfrak{p}$; this occurs if $a/b = c/d$ for some $d \notin \mathfrak{p}$. We have $ad = bc$, so if $b \in \mathfrak{p}$ then either a or d lies in $\mathfrak{p}$, but it could be a and not d.

We view R as a subring of the localization $R_\mathfrak{p}$ via the canonical embedding $r \to r/1$.

Lemma 16.26. *The ring $R_\mathfrak{p}$ is a local ring with maximal ideal $\mathfrak{p}R_\mathfrak{p}$*

Proof. This is obvious when $R_\mathfrak{p}$ is a UFD, but we can't assume this; however we can assume that we always pick representatives $a/b \in R_\mathfrak{p}$ so that $b \notin \mathfrak{p}$. If $a/b \in R_\mathfrak{p}$ is not in $\mathfrak{p}R_\mathfrak{p}$ then clearly $a \notin \mathfrak{p}$ and therefore $b/a \in R_\mathfrak{p}$, so a/b is a unit. Conversely, if $a/b \in R_\mathfrak{p}$ is a unit then $(a/b)(c/d) = 1$ for some $c, d \in R$ with $d \notin \mathfrak{p}$. We then have $ac = bd$, and if a is in $\mathfrak{p}$, then so is bd, but then either $b \in \mathfrak{p}$ or $d \in \mathfrak{p}$, since $\mathfrak{p}$ is prime, which is a contradiction. Thus $R_\mathfrak{p} = \mathfrak{p}R_\mathfrak{p} \sqcup R_\mathfrak{p}^\times$, therefore $R_\mathfrak{p}$ is a local ring with maximal ideal $\mathfrak{p}R_\mathfrak{p}$. □

In general, the localization $R_\mathfrak{p}$ need not be a valuation ring, but provided that $\mathfrak{p}$ is nonzero it is always contained in one, as you will prove on the problem set.

16.5 Valuative criterion for completeness

We now return to our goal of proving that every projective variety is complete. Let X be a variety with coordinate ring $k[X]$, and let P be a point in X. We then define the ideal

$$m_P := \{f \in k[X] : f(P) = 0\}.$$

Note that we have defined what $f(P) = 0$ means, and even how to evaluate f at P, for all the varieties we have considered, so this definition applies to any variety, not just affine varieties. Indeed, m_P is the kernel of the evaluation map $k[X] \to k$ defined by $f \to f(P)$. This makes it clear that m_P is a maximal ideal, since the quotient $k[X]/m_P \simeq k$ is a field.

Definition 16.27. Let X be a variety with coordinate ring $k[X]$ and let $P \in X$. The *local ring of P on X* is the ring

$$\mathcal{O}_P := \mathcal{O}_{P,X} := k[X]_{m_P} = \{g/h \in k(X) : h \notin m_P\}.$$

With Remark 16.25 in mind, it is clear that $\mathcal{O}_P$ is precisely the ring of functions in $k(X)$ that are regular at P.

[2] Be sure not to confuse $R_\mathfrak{p}$ with the quotient $R/\mathfrak{p}$.

We are now ready to state Chevalley's valuative criterion for completeness.

Theorem 16.28. *Let X be a variety such that for every subvariety $Z \subseteq X$ and valuation ring R of $k(Z)/k$ there exists a point $P \in Z$ such that $\mathcal{O}_{P,Z} \subseteq R$. Then X is complete.*

The proof below is adapted from [2, Prop. 7.17].

Proof. So let Y be an affine variety and let $V \subseteq X \times Y$ be a closed set. We may assume that V is irreducible, since we can always write V as a finite union of irreducible sets (the coordinate ring of $X \times Y$ is Noetherian) and then prove that the image of each is closed, and we may replace Y with the image of $V \subseteq X \times Y \to Y$, since whether the image is closed or not does not depend on anything outside of its closure. We now replace X with the image Z of $V \subseteq X \times Y \to X$, to which we will apply the hypothesis of the theorem.

We have the following commutative diagram with dominant morphisms ϕ and ψ.

$$\begin{tikzcd}
 & & Y \\
V \arrow[r, "\subseteq"] \arrow[urr, bend left, "\varphi"] \arrow[drr, bend right, "\psi"'] & Z \times Y \arrow[ur, "\pi_Y"] \arrow[dr, "\pi_Z"] & \\
 & & Z
\end{tikzcd}$$

We need to show that the morphism φ is actually a surjection. So let Q be any point in Y; we will construct a point P such that (P, Q) is in V, which will prove $Q \in \varphi(V)$.

Let $\phi\colon k[Y] \to k$ be the evaluation map $\phi(g) = g(Q)$, which we note fixes k (and is therefore surjective). The morphism of affine algebras $\varphi^*\colon k[Y] \to k[V]$ is injective, since φ is dominant, thus we may regard $k[Y]$ as a subring of $k[V]$, which is in turn embedded in the function field $k(V)$. By Lemma 16.29 below, there exists a valuation ring S of $k(V)/k$ that contains the image of $k[Y]$ in $k(V)$ such that the quotient map $\Phi\colon S \to k$ from S to its residue field k is an extension of ϕ.

Let us now consider the inverse image $R \subseteq k(Z)$ of S under $\psi^*\colon k(Z) \to k(V)$. The ring R is a valuation ring of $k(Z)/k$, because its image S is a valuation ring of $k(V)/k$. By the hypothesis of the theorem there is a point $P \in Z$ such that local ring $\mathcal{O}_{P,Z}$ of Z at P is contained in R. We then have

$$k[Z] \subseteq \mathcal{O}_{P,Z} \subseteq R \xrightarrow{\psi^*} S \longrightarrow k$$

By construction, S contains $k[Y] \subseteq k(V)$, and it contains the injective image of $k[Z]$ under the map above. It follows that S contains the surjective image of $k[Z \times Y] \simeq k[Z] \otimes k[Y]$ in $k[V]$ under the morphism dual to the inclusion $V \subseteq Z \times Y$, and therefore S contains $k[V] \subseteq k(V)$. The intersection of $\ker \Phi$ with $k[V]$ is a maximal ideal of $k[V]$ corresponding to a point in V. This point must be (P, Q); in fact it suffices to show the second coordinate is Q, and this is clear: the map $\Phi\colon S \to k$ is an extension of $\phi\colon k[Y] \to k$, and for any $Q' \neq Q$ we can find a function in $k[Y]$ that vanishes at Q but not at Q' (since $k = \bar{k}$). □

The lemma used in the proof above is a standard result in commutative algebra that we won't prove here.

Lemma 16.29. *Let A be an integral domain contained in a field K and let $\phi\colon A \to k$ be a homomorphism to an algebraically closed field k. Then there exists a valuation ring B of K containing A and a homomorphism $\Phi\colon B \to k$ that extends ϕ. The kernel of Φ is then the maximal ideal of B and k is its residue field.*

Proof. Apply Propositions 5.21 and 5.23 of [1]. □

It will follow easily from Theorem 16.28 that all projective varieties are complete once we prove two lemmas. The first is a technical result that allows us to restrict the residue field of the valuation ring R that appears in the hypothesis of the thoerem.

Lemma 16.30. *Let R be a valuation ring of an extension F/k of an algebraically closed field k. Then there is a valuation ring $R' \subseteq R$ of F/k with residue field isomorphic to k.*

Proof. Let $\mathfrak{m}$ be the maximal ideal of R and let $K = R/\mathfrak{m}$ be its residue field. We may view k as a subfield of K, since the map $k \subseteq R \to R/\mathfrak{m} = K$ is a ring homomorphism of fields. So k is an integral domain contained in K, and the identity map $\phi\colon k \to k$ is a homomorphism to an algebraically closed field. By Lemma 16.29, there is a valuation ring S of K/k whose residue field is k. The map $k \subseteq S \to k$ is then the identity map.

The preimage of $R' = \Psi^{-1}(S) \subseteq R$ under the quotient map $\Psi\colon R \to K$ is a subring of R, and the kernel of the map $R' \to S \to k$ is a maximal ideal $\mathfrak{m}'$ (since k is a field), and $\mathfrak{m}'$ contains $\mathfrak{m} = \Phi^{-1}(0)$. We claim that R' is a valuation ring of F/k. It is clear that R' contains k, we just need to show that it is a valuation ring of F.

So let $x \in F$. If $x \notin R$ then $1/x \in \mathfrak{m} \subseteq \mathfrak{m}' \subseteq R'$. If $x \notin R$ but $x \notin R'$, then $x \notin \mathfrak{m}'$ and therefore $x \notin \mathfrak{m}$, implying that $1/x \in R$, since R is a valuation ring. The image of x in K under the quotient map $R \to K$ does not lie in S, since $x \notin R'$, so the image of $1/x$ in K must lies in S, since S is a valuation ring of K/k. Therefore $1/x \in R'$. Thus for every $x \in F$ either x or $1/x$ lies in R'. So R' is a valuation ring of F, and $R'/\mathfrak{m}' \simeq k$. □

Corollary 16.31. *If X is a variety such that for every subvariety $Z \subseteq X$ and valuation ring R of $k(Z)/k$ with residue field k there is a point $P \in Z$ such that $\mathcal{O}_{P,Z} \subseteq R$, then X is complete.*

The next lemma is almost trivial, but it is the essential reason why projective varieties are complete (in contrast to affine varieties), so we consider it separately.

Lemma 16.32. *Let R be a valuation ring of F. For any $x_0, \ldots, x_n \in F^\times$ there exists $\lambda \in F^\times$ such that $\lambda x_0, \ldots, \lambda x_n \in R$ and at least one λx_i is a unit in R.*

Proof. We proceed by induction. For $n = 0$ we may take $\lambda = 1/x_0$ so that $\lambda x_0 = 1 \in R^\times$. We now assume $\lambda x_0, \ldots, \lambda x_{n-1} \in R$ with $\lambda x_i \in R^\times$ for some $i < n$. If $\lambda x_n \in R$ then we are done, and otherwise $1/(\lambda x_n) \in R$ and we let $\lambda' = 1/x_n$. Then $\lambda' x_j = x_i/x_n = \lambda x_j/(\lambda x_n)$ lies in R for $j < n$, and $\lambda' x_n = 1 \in R^\times$. □

Theorem 16.33. *All projective varieties are complete.*

Proof. By Lemma 16.13, it is enough to show that $\mathbb{P}^n$ is complete. To do this we apply Corollary 16.31. Let Z be a variety in $\mathbb{P}^n$ and let R be a valuation ring of $k(Z)/k$ with residue field k. We will construct a point $P \in Z$ for which $\mathcal{O}_P \subseteq R$.

Let $y_0, \ldots, y_n$ be homogeneous coordinates for $\mathbb{P}^n$ and let $z_0, \ldots, z_n$ denote their images in $k(Z)$. Recall that elements of $k(Z)$ can be represented as rational functions whose numerator and denominator are homogeneous polynomials of the same degree; these correspond to homegenizations of elements of $k(Z_i)$ with respect to y_i, where $Z_i = Z \cap Z_i$ is a nonempty affine part of Z.

By Lemma 16.32 there exists $\lambda \in k(Z)^\times$ such that $\lambda z_0, \dots, \lambda z_n \in R$ with at least one $\lambda z_i \in R^\times$. Let $\phi \colon R \to k$ be the quotient map from R to its residue field, and let P be the projective point $(\phi(\lambda z_0) : \phi(\lambda z_1) : \dots : \phi(\lambda z_n))$, where we note that at least one $\phi(\lambda z_i)$ is nonzero. The point P lies in Z, since for any homogeneous $f \in I(Z)$ of degree d we have

$$f(\lambda z_0, \dots, \lambda z_n) = \lambda^d f(z_0, \dots, z_n) = 0$$

as an element of $k(Z)$, and therefore

$$0 = \phi(0) = \phi(f(\lambda z_0, \dots, \lambda z_n)) = f(\phi(\lambda z_0), \dots, \phi(\lambda z_n)) = f(P).$$

Any element of the local ring $\mathcal{O}_P$ can be written as g/h with $h(P) \neq 0$, and we can write g and h as homogeneous polynomials in $\lambda z_0, \dots, \lambda z_n$ that lie in R (since the λz_i generate $k(Z)$ as a k-algebra). We then have

$$\phi(h(\lambda z_0, \dots, \lambda z_n)) = h(\phi(\lambda z_0), \dots, \phi(\lambda z_n)) = h(P) \neq 0,$$

so $h \notin \ker \phi$, and therefore $h \in R^\times$, so $g/h \in R$. Thus $\mathcal{O}_P \subseteq R$, as desired. □

References

[1] M. F. Atiyah and I. G. MacDonald, *Introduction to commutative algebara*, Addison-Wesley, 1969.

[2] D. Bump, *Algebraic geometry*, World Scientific, 1998.

18.782 Introduction to Arithmetic Geometry **Fall 2013**
Lecture #17 **11/05/2013**

Throughout this lecture k denotes an algebraically closed field.

17.1 Tangent spaces and hypersurfaces

For any polynomial $f \in k[x_1,\ldots,x_n]$ and point $P=(a_1,\ldots,a_n)\in\mathbb{A}^n$ we define the affine linear form

$$f_P(x_1,\ldots,x_n) \;:=\; \sum_{i=1}^n \frac{\partial f}{\partial x_i}(P)(x_i-a_i).$$

The zero locus of f_P in $\mathbb{A}^n$ is an *affine hyperplane* in $\mathbb{A}^n$, a subvariety isomorphic to $\mathbb{A}^{n-1}$. Note that $f_P(P)=0$, so the zero locus contains P.

Definition 17.1. Let P be a point on an affine variety V. The *tangent space* of V at P is the variety $T_P(V)$ defined by the ideal $\{f_p : f\in I(V)\}$.

It is clear that $T_p(V)$ is a variety; indeed, it is the nonempty intersection of a set of affine hyperplanes in $\mathbb{A}^n$ and therefore an affine subspace of $\mathbb{A}^n$ isomorphic to $\mathbb{A}^d$, where $d=\dim T_P(V)$. Note that the definition of $T_P(V)$ does not require us to choose a set of generators for $I(V)$, but for practical applications we want to be able to compute $T_P(V)$ in terms of a finite set of generators for $I(V)$. The following lemma shows that we can do this, and, most importantly, it does not matter which set of generators we pick.

Lemma 17.2. *Let P be a point on an affine variety V. If $f_1,\ldots,f_m$ generate $I(V)$, then the corresponding affine linear forms $f_{1,P},\ldots,f_{m,P}$ generate $I(T_P(V))$.*

Proof. Let $g=\sum h_if_i$ be an element of $I(V)$. Applying the product rule and the fact that $f_{i,P}(P)=0$ yields

$$g_P=\sum_i\bigl(h_i(P)f_{i,P}+h_{i,P}f_i(P)\bigr)=\sum_i h_i(P)f_{i,P},\tag{1}$$

which is an element of the ideal $(f_{1,P},\ldots,f_{m,P})$. Thus $I(T_p(V))=(f_{1,P},\ldots,f_{m,P})$. □

When considering the tangent space of a variety at a particular point P, we may assume without loss of generality that $P=(0,\ldots,0)$, since we can always translate the ambient affine space $\mathbb{A}^n$; this is just a linear change of coordinates (indeed, this is the very definition of affine space, it is a vector space without a distinguished origin). We can then view the affine subspace $T_P(V)\subseteq\mathbb{A}^n$ as a linear subspace of the vector space k^n. The affine linear forms f_P are then linear forms on k^n, equivalently, elements of the dual space $(k^n)^\vee$.

Recall from linear algebra that the dual space $(k^n)^\vee$ is the space of linear functionals $\lambda\colon k^n\to k$. The orthogonal complement $S^\perp\subseteq(k^n)^\vee$ of a subspace $S\subseteq k^n$ is the set of linear functionals λ for which $\lambda(P)=0$ for all $P\in S$; it is a subspace of $(k^n)^\vee$, and since k^n has finite dimension n, we have $\dim S+\dim S^\perp=n$.

Theorem 17.3. *Let P be a point on an affine variety $V\subseteq\mathbb{A}^n$ with ideal $I(V)=(f_1,\ldots,f_m)$. If we identify $\mathbb{A}^n$ with the vector space k^n with origin at P, the subspace of $(k^n)^\vee$ spanned by the linear forms $f_{1,P},\ldots,f_{m,P}$ is $T_P(V)^\perp$, the orthogonal complement of $T_p(V)$.*

Andrew V. Sutherland

Proof. This follows immediately from Lemma 17.2 and its proof; the set of linear forms in $I(T_P(V))$ is precisely the set of linear forms that vanish at every point in $T_p(V)$, which, by definition, is the orthogonal complement $T_P^\vee$. Moreover, we see from (1) that every linear form in $I(T_P(V))$ is a k-linear combination of $f_{1,P}\ldots,f_{m,P}$. □

The vector space $T_P(V)^\perp$ is called the *cotangent space* of V at P. As noted above, as a variety, $T_P(V)$ is isomorphic to some $\mathbb{A}^d$, where $d = \dim T_P(V)$, and it follows that the dimenstion of $T_P(V)$ as a vector space is the same as its dimension as a variety, since $\dim \mathbb{A}^d = d = \dim_k k^d$. The dimension of $T_P(V)^\perp$ is then $n - d$.

Recall from Lecture 13 the Jacobian matrix

$$J_P = J_P(f_1,\ldots,f_m) := \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(P) & \cdots & \frac{\partial f_1}{x_n}(P) \\ \vdots & \cdots & \vdots \\ \frac{\partial f_m}{\partial x_1}(P) & \cdots & \frac{\partial f_n}{x_n}(P) \end{pmatrix}.$$

For a variety V with $I(V) = (f_1,\ldots,f_m)$, we defined a point $P \in V$ to be *smooth* (or nonsingular) precisely when $\operatorname{rank} J_P = n - \dim V$. Viewing J_P as the matrix of a linear transformation from $(k^n)^\vee$ to $(k^n)^\vee$ whose image is $T_P(V)^\perp$, we obtain the following corollary of Theorem 17.3.

Corollary 17.4. *Let P be a point on an affine variety $V \subseteq \mathbb{A}^n$ with $I(V) = (f_1,\ldots,f_m)$, and let $J_P = J_P(f_1,\ldots,f_m)$. Then $\dim T_P(V)^\perp = \operatorname{rank} J_P$ and $\dim T_P(V) = n - \operatorname{rank} J_P$. In particular, the rank of J_P does not depend on the choice of generators for $I(V)$ and P is a smooth point of V if and only if $\dim T_P = \dim V$.*

Remark 17.5. For projective varieties V we defined smooth points P as points that are smooth in all (equivalently, any) affine part containing P. One can also define tangent spaces and Jacobian matrices for projective varieties directly using generators for the homogeneous ideal of V. This is often more convenient for practical computations.

Corollary 17.13 makes it clear that, as claimed in Lecture 13, our notion of a smooth point $P \in V$ is well defined; it does not depend on which generators $f_1,\ldots,f_m$ of $I(V)$ we use to compute J_P, or even on the number of generators. Now we want to consider what can happen when $\dim T_P(V) \neq \dim V$. In this case $\dim T_P(V)$ must be strictly greater than $\dim V$; this is easy to see when V is defined by a single equation, since then $J_P(f)$ has just one row and its rank is either 0 or 1.

Definition 17.6. A variety V for which $I(V)$ is a nonzero principal ideal is a *hypersurface*.

Lemma 17.7. *Every hypersurface in $\mathbb{A}^n$ or $\mathbb{P}^n$ has dimension $n-1$.*

Proof. Let $V \subseteq \mathbb{A}^n$ be a hypersurface with $I(V) = (f)$ for some nonzero $f \in k[x_1,\ldots,x_n]$. We must have $\dim V \le n-1$, since $V \subsetneq \mathbb{A}^n$. Let $\phi\colon k[x_1,\ldots,x_n] \to k[x_1,\ldots,x_n]/(f)$ be the quotient map. We must have $f \notin k$, since $V \neq \emptyset$, so $\deg_{x_i} f > 0$ for some x_i, say x_1. If $\dim V < n-1$ then the transcendence degree of $k(V)$ is less than $n-1$, therefore $\phi(x_2),\ldots,\phi(x_n)$ must be algebraically dependent as elements of $k(V)$. Thus there exists $g \in k[x_2,\ldots,x_n]$ such that $g(\phi(x_2),\ldots,\phi(x_n)) = 0$. But then $\phi(g) = 0$, so $g \in \ker\phi = (f)$. But this is a contradiction, since $\deg_{x_1} g = 0$. So $\dim V = n-1$. If $V \subseteq \mathbb{P}^n$, then one of its affine parts V_i is a hypersurface in $\mathbb{A}^n$, and then $\dim V = \dim V_i = n-1$. □

The converse to Lemma 17.7 is true; every variety of codimension 1 is a hypersurface. This follows from the general fact that every variety is birationally equivalent to a hypersurface. Recall that a function field F/k if any finitely generated extension; the *dimension* of a function field is its transcendence degree.

Theorem 17.8. *Let F/k be a function field of dimension n. Then there exist algebraically independent elements $\alpha_1,\ldots,\alpha_n \in F$ and an element α_{n+1} algebraic over $k(\alpha_1,\ldots,\alpha_m)$ such that $F = k(\alpha_1,\ldots,\alpha_{n+1})$.*

The following proof is adapted from [1, App. 5, Thm. 1].

Proof. Let $\gamma_1,\ldots,\gamma_m$ be a set of generators for F/k of minimal cardinality m, ordered so that $\gamma_1,\ldots,\gamma_n$ is a transcendence basis (every set of generators contains a transcendence basis). If $m = n$ then we may take $\gamma_{n+1} = 0$ and we are done. Otherwise γ_{n+1} is algebraic over $k(\gamma_1,\ldots,\gamma_n)$, and we claim that in fact $m = n+1$ and we are also done.

Suppose $m > n+1$. Let $f \in k[x_1,\ldots,x_{n+1}]$ be irreducible with $f(\gamma_1,\ldots,\gamma_{n+1}) = 0$; such an f exists since $\gamma_1,\ldots,\gamma_{n+1}$ are algebraically dependent. We must have $\partial f/\partial x_i \neq 0$ for some x_i; if not than we must have $\operatorname{char}(k) = p > 0$ and $f = g(x_1^p,\ldots,x_{n+1}^p) = g^p(x_1,\ldots,x_{n+1})$ for some $g \in k[x_1,\ldots,x_{n+1}]$, but this is impossible since f is irreducible. It follows that γ_i is algebraic, and in fact separable, over $K = k(\gamma_1,\ldots,\gamma_{i-1},\gamma_{i+1},\ldots,\gamma_{n+1})$; the irreducible polynomial $f(\gamma_1,\ldots,\gamma_{i-1},x_i,\gamma_{i+1},\ldots,\gamma_{n+1|})$ has γ_i as a root, and its derivative is nonzero. Now γ_m is also algebraic over K, and it follows from the primitive element theorem [2, §6.10] that $K(\gamma_i,\gamma_m) = K(\delta)$ for some $\delta \in K$.[1] But this contradicts the minimality of m, so we must have $m = n+1$ as claimed. □

Remark 17.9. Theorem 17.8 holds for any perfect field k; it is not necessary for k to be algebraically closed.

Theorem 17.10. *Every affine (resp. projective) variety of dimension n is birationally equivalent to a hypersurface in $\mathbb{A}^{n+1}$ (resp. $\mathbb{P}^{n+1}$).*

Proof. Two projective varieties are birationally equivalent if and only if all their nonempty affine parts are, and the projective closure of a hypersurface is a hypersurface, so it suffices to consider affine varieties. Recall from Lecture 15 that varieties are birationally equivalent if and only if their function fields are isomorphic, and it follows from Theorem 17.8 that every function field arises as the function field of a hypersurface: if $k(V) = k(\gamma_1,\ldots,\gamma_{n+1})$ with $\gamma_1,\ldots,\gamma_n)$ algebraically independent, then there exists an irreducible polynomial f in $k[x_1,\ldots,x_{n+1}]$ for which $f(\gamma_1,\ldots,\gamma_{n+1}) = 0$, and then V is birationally equivalent to the zero locus of f in $\mathbb{A}^{n+1}$. □

Corollary 17.11. *The set of singular points of a variety is a closed subset; equivalently, the set of nonsingular points is a dense open subset.*

Proof. It suffices to prove this for affine varieties. So let $V \subseteq \mathbb{A}^n$ be an affine variety with ideal $(f_1,\ldots,f_m)$, and for any $P \in V$ let $J_P = J_P(f_1,\ldots,f_m)$ be the Jacobian matrix. Then

$$\mathrm{Sing(V) := \{P\colon\ \dim T_P(V) > \dim V\} = \{P\colon\ rank\, J_P < n - \dim V\}}$$

is the set of singular points on V. Let $r = n - \dim V$. We have $\operatorname{rank} J_P < r$ if and only if every $r \times r$ minor of J_P has determinant zero. If we now consider the matrix of polynomials

[1] As noted in [2], to prove $K(\alpha,\beta) = K(\delta)$ for some $\delta \in K(\alpha,\beta)$, we only need one of α,β to be separable.

$(\partial f_i/\partial x_j)$, the determinant of each of its $r \times r$ minors is a polynomial in $k[x_1,\ldots,x_n]$, and Sing(V) is the intersection of V with the zero locus of all these polynomials. Thus Sing(V) is an algebraic set, hence closed. □

Recall the one-to-one correspondence between points $P = (a_1,\ldots,a_n)$ in $\mathbb{A}^n$ and maximal ideals $M_P = (x_1 - a_1,\ldots,x_n - a_n)$ of $k[\mathbb{A}^n]$. If $V \subseteq \mathbb{A}^n$ is an affine variety, then the maximal ideals m_P of its coordinate ring $k[V] = k[\mathbb{A}^n]/I(V)$ are in one-to-one correspondence with the maximal ideals M_P of $k[\mathbb{A}^n]$ that contain $I(V)$; these are precisely the maximal ideals M_P for which $P \in V$.

If we choose coordinates so that $P = (0,\ldots,0)$, then M_P is a k-vector space that contains M_P^2 as a subspace, and the quotient space M_P/M_P^2 is then also a k-vector space. Indeed, its elements correspond to (cosets of) linear forms on k^n. We may similarly view m_P, m_P^2, and m_P/m_P^2 as k-vector spaces, and this leads to the following theorem.

Theorem 17.12. *Let P be a point on an affine variety V. Then $T_P(V)^\vee \simeq m_P/m_P^2$.*

Proof. As above we assume without loss of generality that $P = (0,\ldots,0)$. Then M_P consists of the polynomials in $k[x_1,\ldots,x_n]$ for which each term has degree at least 1 (equivalently, constant term 0). We now consider the linear transformation

$$D\colon M_P \to (k^n)^\vee$$

that sends $f \in M_P$ to the linear form $f_P \in (k^n)^\vee$. This map is surjective, and its kernel is M_P^2; we have $f_P = 0$ if and only if $\partial f/\partial x_i(0) = 0$ for $i = 1,\ldots,n$, and this occurs precisely when every term in f has degree at least 2, equivalently, $f \in M_P^2$. It follows that

$$M_P/M_P^2 \simeq (k^n)^\vee.$$

The restriction map $(k^n)^\vee \to (T_P)^\vee$ that restricts the domain of a linear form on k^n to $T_P(V)$ is surjective, and composing this with D yields a surjective linear transformation

$$d\colon M_P \to T_P(V)^\vee$$

whose kernel we claim is equal to $M_P^2 + I(V)$ (this is a sum of ideals in $k[x_1,\ldots,x_n]$ that is clearly a subset of M_P). A polynomial $f \in M_P$ lies in $\ker d$ if and only if the restriction of f_P to $T_P(V)$ is the zero function, which occurs if and only if $f_P = g_P$ for some $g \in I(V)$, since T_P the zero locus of g_P for $g \in I(V)$. But this happens if and only if $f - g$ lies $\ker D = M_P^2$, equivalently, $f \in M_P^2 + I(V)$.

We therefore have

$$T_P(V)^\vee \simeq \frac{M_P}{M_P^2 + I(V)} \simeq \frac{M_P/I(V)}{(M_P^2 + I(V)/I(V)} = \frac{M_P/I(V)}{M_P^2/I(V)} \simeq m_P/m_P^2. \qquad \square$$

Corollary 17.13. *The smooth points P on a variety V are precisely the points P for which*

$$\dim m_P/m_P^2 = \dim V = \dim k[V]$$

The three dimensions in the corollary above are, respectively, the dimension of m_P/m_P^2 as a k-vector space, the dimension of V as a variety, and the Krull dimension of the coordinate ring $k[V]$; as noted in Lecture 13, we always have $\dim V = \dim k[V]$. The key point is that we now have a completely algebraic notion of smooth points. If R is any affine algebra, the maximal ideals $\mathfrak{m}$ of R correspond to smooth points on a variety with coordinate ring R, and we can characterize the "smooth" maximal ideals as those for which $\dim_k \mathfrak{m}/\mathfrak{m}^2 = \dim R$, where $k = R_\mathfrak{m}/\mathfrak{m}$ is now the residue field of the localization of R at $\mathfrak{m}$. Smooth varieties then correspond to affine algebras R in which every maximal ideal is "smooth".

References

[1] I. R. Shafarevich, *Basic algebraic geometry*, 2nd edition, Springer-Verlag, 1994.

[2] B. L. van der Waerdan, *Algebra, Volume I*, 7th edition, Springer, 1991.

18.782 Introduction to Arithmetic Geometry **Fall 2013**
Lecture #18 **11/07/2013**

As usual, all the rings we consider are commutative rings with an identity element.

18.1 Regular local rings

Consider a local ring R with unique maximal ideal $\mathfrak{m}$. The ideal $\mathfrak{m}$ is, in particular, an abelian group, and it contains $\mathfrak{m}^2$ as a normal subgroup, so we can consider the quotient group $\mathfrak{m}/\mathfrak{m}^2$, where the group operation is addition of cosets:

$$(m_1 + \mathfrak{m}^2) + (m_2 + \mathfrak{m}^2) = (m_1 + m_2) + \mathfrak{m}^2.$$

But $\mathfrak{m}$ is also an ideal, so it is closed under multiplication by R, and it is a maximal ideal, so $R/\mathfrak{m}$ is a field (the residue field). The quotient group $\mathfrak{m}/\mathfrak{m}^2$ has a natural structure as an $(R/\mathfrak{m})$-vector space. Scalars are cosets $r + \mathfrak{m}$ in the field $R/\mathfrak{m}$, and scalar multiplication is defined by

$$(r + \mathfrak{m})(m + \mathfrak{m}^2) = rm + \mathfrak{m}^2.$$

In practice one often doesn't write out the cosets explicitly (especially for elements of the residue field), but it is important to keep the underlying definitions in mind; they are a valuable compass if you ever start to feel lost.

The motivation for this discussion is the case where R is the local ring $\mathcal{O}_P$ of regular functions at a point P on a variety V. In this setting $\mathfrak{m}/\mathfrak{m}^2$ is precisely the vector space m_P/m_P^2 that is isomorphic to the $T_P^\vee$, the dual of the tangent space at P; recall from the previous lecture that P is a smooth point of V if and only if $\dim m_P/m_P^2 = \dim V$. We now give an algebraic characterization of this situation that does not involve varieties. We write $\dim \mathfrak{m}/\mathfrak{m}^2$ to indicate the dimension of $\mathfrak{m}/\mathfrak{m}^2$ as an $(R/\mathfrak{m})$-vector space, and we write $\dim R$ to denote the (Krull) dimension of the ring R.

Definition 18.1. A Noetherian local ring R with maximal ideal $\mathfrak{m}$ is a *regular local ring* if $\dim \mathfrak{m}/\mathfrak{m}^2 = \dim R$ (note that Noetherian is included in the definition of regular).[1]

We are particularly interested in regular local rings of dimension 1, these correspond to rings $\mathcal{O}_P$ of regular functions at a smooth point P on a curve (a variety of dimension one).

Theorem 18.2. *A ring R is a regular local ring of dimension one if and only if it is a discrete valuation ring.*

Proof. We prove the easier direction first. Let R be a discrete valuation ring (DVR) with maximal ideal $\mathfrak{m} = (t)$. Then R is a local ring, and it is certainly Noetherian, since it is a principal ideal domain (PID). Its prime ideals are (0) and (t), so it has dimension 1, and $t + \mathfrak{m}^2$ generates $\mathfrak{m}/\mathfrak{m}^2$, so $\dim \mathfrak{m}/\mathfrak{m}^2 = 1$. Thus R is a regular local ring of dimension 1.

Let R be a regular local ring of dimension one. Its unique maximal ideal $\mathfrak{m}$ is not equal to $\mathfrak{m}^2$, since $\dim \mathfrak{m}/\mathfrak{m}^2 = 1 > 0$; in particular, $\mathfrak{m} \neq (0)$ and R is not a field. Let $t \in \mathfrak{m} - \mathfrak{m}^2$. Then $t + \mathfrak{m}^2$ generates $\mathfrak{m}/\mathfrak{m}^2$, since $\dim \mathfrak{m}/\mathfrak{m}^2 = 1$. By Corollary 18.4 of Nakayama's lemma (proved below), t generates $\mathfrak{m}$. So every $x \in R - (0)$ has the form $x = ut^n$, with $u \in R^\times$ and $n \in \mathbb{Z}_{\geq 0}$ (since R is a local ring with $\mathfrak{m} = (t)$), and every nonzero ideal is principal, of the form (t^n). It follows that the prime ideals in R are exactly (0) and (t), since R has dimension one. So $R = R/(0)$ is an integral domain, and therefore a PID, hence a DVR. □

[1] More generally, a Noetherian ring is *regular* if all of its localizations at prime ideals are regular.

To prove Corollary 18.4 used in the proof above we require a special case of what is known as Nakayama's lemma. The statement of the lemma may seem a bit strange at first, but it is surprisingly useful and has many applications.

Lemma 18.3 (Nakayama)**.** *Let R be a local ring with maximal ideal $\mathfrak{m}$ and suppose that M is a finitely generated R-module with the property $M = \mathfrak{m}M$. Then M is the zero module.*

Proof. Let $b_1, \ldots, b_n$ be generators for M. By hypothesis, every b_i can be written in the form $b_i = \sum_j a_{ij} b_j$ with $a_{ij} \in \mathfrak{m}$. In matrix form we have $B = AB$, where $B = (b_1, \ldots, b_n)^t$ is a column vector and $A = (a_{ij})$ is an $n \times n$ matrix with entries in $\mathfrak{m}$. Equivalently, $(I - A)B = 0$, where I is the $n \times n$ identity matrix. The diagonal entries $1 - a_{ii}$ of $I - A$ are units, because $1 - a_{ii}$ cannot lie in $\mathfrak{m}$ (otherwise $1 \in \mathfrak{m}$, which is not the case), and every element of $R - \mathfrak{m}$ is a unit (since R is a local ring). However the off-diagonal entries of $I - A$ all lie in $\mathfrak{m}$. Expressing the determinant d of $I - A$ as a sum over permutations, it is clear that $d = 1 + a$ for some $a \in \mathfrak{m}$, hence d is a unit and $I - A$ is invertible. But then $(I - A)^{-1}(I - A)B = B = 0$, which means that M is the zero module. □

Corollary 18.4. *Let R be a local Noetherian ring with maximal ideal $\mathfrak{m}$. Then $t_1, \ldots, t_n \in \mathfrak{m}$ generate $\mathfrak{m}$ if and only if their images generate $\mathfrak{m}/\mathfrak{m}^2$ as an $R/\mathfrak{m}$ vector space.*

Proof. The "only if" direction is clear. Let N be the ideal $(t_1, \ldots, t_n) \subseteq \mathfrak{m}$. If the images of $t_1, \ldots, t_n$ in $\mathfrak{m}/\mathfrak{m}^2$ generate $\mathfrak{m}/\mathfrak{m}^2$ as an $R/\mathfrak{m}$-vector space, then we have

$$\begin{aligned} N + \mathfrak{m}^2 &= \mathfrak{m} + \mathfrak{m}^2 \\ (N + \mathfrak{m}^2)/N &= (\mathfrak{m} + \mathfrak{m}^2)/N \\ \mathfrak{m}(\mathfrak{m}/N) &= \mathfrak{m}/N, \end{aligned}$$

where we have used $N/N = 0$ and $\mathfrak{m} + \mathfrak{m}^2 = \mathfrak{m}$ (since $\mathfrak{m}^2 \subseteq \mathfrak{m}$). By Nakayama's lemma, $M = \mathfrak{m}/N$ is the zero module, so $\mathfrak{m} = N$ and $t_1, \ldots, t_n$ generate $\mathfrak{m}$. □

18.2 Smooth projective curves

It follows from Theorem 18.2 that for a smooth curve C the local rings $\mathcal{O}_P = k[C]_{m_P}$ are all discrete valuation rings of $k(C)/k$. If C is a projective curve, then by Theorem 16.33 it is complete, and from the proof of Theorem 16.33 we know that it satisfies Chevalley's criterion: every valuation ring R of $k(C)/k$ contains a local ring $\mathcal{O}_P$. The fact that $\mathcal{O}_P$ is a discrete valuation ring actually forces $R = \mathcal{O}_P$; this is a consequence of the following theorem.

Theorem 18.5. *Let R_1 and R_2 be valuation rings with the same fraction field, let $\mathfrak{m}_1$ and $\mathfrak{m}_2$ be their respective maximal ideals, and suppose $R_1 \subsetneq R_2$. Then $\mathfrak{m}_2 \subsetneq \mathfrak{m}_1$ and* $\dim R_2 < \dim R_1$*. In particular, R_1 cannot be a discrete valuation ring.*

Proof. We first note $R_1 \subseteq R_2$ implies $R_1^\times \subseteq R_2^\times$. For $x \in R_2 - R_1$ we have $1/x \in R_1 \subseteq R_2$ and $x \in R_2^\times$, so $R_2 - R_1 \subseteq R_2^\times$. Thus $R_2 - R_2^\times \subseteq R_2 - R_1^\times = (R_2 - R_1) \sqcup (R_1 - R_1^\times)$, and this implies $\mathfrak{m}_2 = R_2 - R_2^\times \subseteq R_1 - R_1^\times = \mathfrak{m}_1$ since $R_2 - R_1$ and $R_2 - R_2^\times$ are disjoint. And for any $x \in R_2 - R_1$ we have $1/x \in R_2^\times = R_2 - \mathfrak{m}_2$ and $1/x \in R_1 - R_1^\times = \mathfrak{m}_1$, so $\mathfrak{m}_2 \subsetneq \mathfrak{m}_1$.

Every prime ideal of R_2 is contained in $\mathfrak{m}_2$, hence in $\mathfrak{m}_1$, and if $\mathfrak{p}$ is prime in R_2 then $\mathfrak{p} \cap R_1$ is clearly prime in R_1: if $ab \in \mathfrak{p}$ for some $a, b \in R_1 \subseteq R_2$ then one of a, b lies in $\mathfrak{p}$. Thus every chain of prime ideals in R_2 is also a chain of prime ideals in R_1, and in R_1 any such chain can be extended by adding the prime ideal $\mathfrak{m}_1$. Thus $\dim R_2 < \dim R_1$. If R_1 is a DVR then $\dim R_2 < \dim R_1 = 1$, but $\dim R_2 \geq 1$, since R_2 is a valuation ring (not a field), therefore R_1 is not a DVR. □

Thus we have a one-to-one correspondence between the points on a smooth projective curve C and the discrete valuation rings of $k(C)/k$.

Theorem 18.6. *Let C be a smooth projective curve. Every rational map $\phi\colon C \to V$ from C to a projective variety V is a morphism.*

Proof. Let $\phi = (\phi_0 : \cdots : \phi_n)$ and consider any point $P \in C$. Let us pick a uniformizer t for the discrete valuation ring $\mathcal{O}_P$ (a generator for the maximal ideal m_P), and let

$$n = \min\{\mathrm{ord}_P(\phi_1), \ldots, \mathrm{ord}_P(\phi_n)\},$$

where $\mathrm{ord}_P\colon k(C) \to k(C)^\times/\mathcal{O}_P^\times \simeq \mathbb{Z}$ is the discrete valuation of $\mathcal{O}_P$. If $n = 0$ then ϕ is regular at P, since then all the ϕ_i are defined at P and at least one is a unit in $\mathcal{O}_P^\times$, hence nonzero at P. But in any case we have

$$\mathrm{ord}_P(t^{-n}\phi_i) = \mathrm{ord}_P(\phi_i) - n \geq 0$$

for $i = 0, \ldots, n$, with equality for at least one value of i. It follows that

$$\big(t^{-n}\phi_0 : \cdots : t^{-n}\phi_n\big) = (\phi_0 : \cdots : \phi_n)$$

is regular at P. This holds for every $P \in C$, so ϕ is a regular rational map, hence a morphism. □

Corollary 18.7. *Every rational map $\phi\colon C_1 \to C_2$ between smooth projective curves is either constant or surjective.*

Proof. Projective varieties are complete, so $\mathrm{im}(\phi)$ is a subvariety of C_2, and since $\dim C_2 = 1$ this is either a point (in which case ϕ is constant) or all of C_2. □

Corollary 18.8. *Every birational map between smooth projective curves is an isomorphism.*

It follows from Corollary 18.8 that if a curve C_1 is birationally equivalent to any smooth projective curve C_2, then all such C_2 are isomorphic. We want to show that such a C_2 always exists. Recall that birationally equivalent curves have isomorphic function fields. Thus it suffices to show that every function field of dimension one actually arises as the function field of a smooth projective curve.

18.3 Function fields as abstract curves

Let F/k be a function field of dimension one, where k is an algebraically closed field. We know that if F is the function field of a smooth projective curve C, then there is a one-to-one correspondence between the points of C and the discrete valuation rings of F. Our strategy is to define an *abstract curve* C_F whose "points" correspond to the discrete valuation rings of F, and then show that it is actually isomorphic to a smooth projective curve.

So let $X = X_F$ be the set of all maximal ideals P of discrete valuation rings of F/k. The elements of $P \in X_F$ are called *points* (or *places*). Let $\mathcal{O}_{P,X} = \mathcal{O}_P$ denote the valuation ring with maximal ideal P, and let ord_P denote its associated valuation. For any $U \subset X$ the *ring of regular functions* on U is the ring

$$\mathcal{O}_X(U) = \mathcal{O}(U) := \cap_{P \in U} \mathcal{O}_P = \{f \in F : \operatorname{ord}_P(f) \geq 0 \text{ for all } P \in U\} \subseteq F,$$

and we call $\mathcal{O}(X)$ the *ring of regular functions* (or *coordinate ring*) of X. Note that $\mathcal{O}(X)$ is precisely the intersection of all the valuation rings of F/k.

For $f \in \mathcal{O}_P$ we define $f(P)$ to be the image of f in the residue field $\mathcal{O}_P/P \simeq k$; thus

$$f(P) = 0 \iff f \in P \iff \operatorname{ord}_P(f) > 0.$$

For $f \in \mathcal{O}_X$ we have $f(P) = 0$ if and only if $\operatorname{ord}_P(f) > 0$. We then give X the Zariski topology by taking as closed sets the zero locus of any subset of $\mathcal{O}(X)$.[2] If F is actually the function field of a smooth projective curve, all the definitions above agree with our usual notation, as we will verify shortly.

Definition 18.9. An *abstract curve* is the topological space $X = X_F$ with rings of regular functions $\mathcal{O}_{X,U}$ determined by the function field F/k as above. A morphism $\phi\colon X \to Y$ between abstract curves or projective varieties is a continuous map such that for every open $U \subseteq Y$ and $f \in \mathcal{O}_Y(U)$ we have $f \circ \phi \in \mathcal{O}_X(\phi^{-1}(U))$.

As you will verify in the homework, if X and Y are both projective varieties this definition of a morphism is equivalent to our earlier definition of a morphism between projective varieties. The identity map $X \to X$ is obviously a morphism, and we can compose morphisms: if $\phi\colon X \to Y$ and $\varphi\colon Y \to Z$ are morphisms, then $\varphi \circ \phi$ is continuous, and for any open $U \subseteq Z$ and $f \in \mathcal{O}_Z(U)$ we have $f \circ \varphi \in \mathcal{O}_Y(\varphi^{-1}(U))$, and then

$$f \circ (\varphi \circ \phi) = (f \circ \varphi) \circ \phi \in \mathcal{O}_X(\phi^{-1}(\varphi^{-1}(U))) = \mathcal{O}_X((\varphi \circ \phi)^{-1}(U)).$$

Thus we have a category whose objects include both abstract curves and projective varieties.

Let us verify that we have set things up correctly by proving that every smooth projective curve is isomorphic to the abstract curve determined by its function field. This follows immediately from our definitions, but it is worth unravelling them once just to be sure.

Theorem 18.10. *Let C be a smooth projective curve and let $X = X_{k(C)}$ be the abstract curve associated to its function field. Then C and X are isomorphic.*

Proof. For the sake of clarity, let us identify the points (discrete valuation rings) of X as maximal ideals m_P corresponding to points $P \in C$. As noted above there is a one-to-one correspondence between $P \in C$ and $m_P \in X$, we just need to show that this induces an isomorphism of curves. So let $\phi\colon C \to X$ be the bijection that sends P to m_P.

For any $U \subseteq C$ we have, by definition, $\mathcal{O}_C(U) = \cap_{P \in U} \mathcal{O}_{P,C}$ and $\mathcal{O}_X(V) = \cap_{m_P \in V} \mathcal{O}_{m_P,X}$, so $\mathcal{O}_C(U) = \mathcal{O}_X(\phi(U))$ In particular,

$$\mathcal{O}(C) = \mathcal{O}(\phi(C)) = \mathcal{O}(X),$$

hence the rings of regular functions of C and X are actually identical (not just in bijection). Moreover, for any open $U \subseteq X$ and $f \in \mathcal{O}_X(U)$ we have $f \circ \phi = f \in \mathcal{O}_C(\phi^{-1}(U))$, and for any open $U \subseteq C$ and $f \in \mathcal{O}_{C,U}$ we have $f \circ \phi^{-1} = f \in \mathcal{O}_X(\phi(U))$.

[2] As we will prove in this next lecture, this is just the cofinite topology: the open sets are the empty set and complements of finite sets.

A set $U \subseteq C$ is closed if and only if it is the zero locus of some subset of $\mathcal{O}(C)$, and for any $P \in C$, equivalently, any $\phi(P) \in X$, we have

$$f(P) = 0 \iff \operatorname{ord}_P(f) > 0 \iff f(\phi(P)) = 0,$$

where we are using the definition of $f(\phi(P)) = f(m_P)$ for $m_P \in X$ on the right. It follows that ϕ is a topology isomorphism from C to X; in particular, both ϕ and ϕ^{-1} are continuous. Thus ϕ and ϕ^{-1} are both morphisms, and $\phi \circ \phi^{-1}$ and $\phi^{-1} \circ \phi$ are the identity maps. □

One last ingredient before our main result; we want to be able to construct smooth affine curves with a specified function field that contain a point whose local ring is equal to a specific discrete valuation ring.

Lemma 18.11. *Let R be a discrete valuation ring of a function field F/k of dimension one. There exists a smooth affine curve C with $k(C) = F$ such that $R = \mathcal{O}_P$ for some $P \in C$.*

Proof. The extension F/k is finitely generated, so let $\alpha_1, \ldots, \alpha_n$ be generators, and replace α_i with $1/\alpha_i$ as required so that $\alpha_1, \ldots, \alpha_n \in R$. Let S be the intersection of all discrete valuation rings of F/k that contain the subalgebra $k[\alpha_1, \ldots, \alpha_n] \subseteq F$. Then $S \subseteq R$ is an integral domain with fraction field F. The kernel of the map from the polynomial ring $k[x_1, \ldots, x_n]$ to S that sends each x_i to α_i is a prime ideal I for which $S = k[x_1, \ldots, x_n]/I$. The variety $C \subseteq \mathbb{A}^n$ defined by I has coordinate ring $k[C] = S \subseteq R$ and function field $k(C) = F$, so it has dimension one and is a curve

Moreover, the curve C is smooth; its coordinate ring S is integrally closed (it is an intersection of discrete valuation rings, each of which is integrally closed), and by Lemma 18.12 below, all its local rings $\mathcal{O}_P$ are discrete valuation rings, hence regular, and therefore every point $P \in C$ is smooth.

Let $\phi\colon R \to R/\mathfrak{m} = k$ be the quotient map and consider the point $P(\phi(x_1), \ldots, \phi(x_n))$. Every f in the maximal ideal m_P of $\mathcal{O}_P$ satisfies

$$\phi(f) = \phi(f(x_1, \ldots, x_n)) = f(\phi(x_1), \ldots, \phi(x_n)) = f(P) = 0$$

and therefore lies in $\mathfrak{m}$. By Theorem 18.5, $R = \mathcal{O}_P$ as desired. □

The following lemma is a standard result of commutative algebra (so feel free to skip the proof on a first reading), but it is an essential result that has a reasonably straight-forward proof (using Theorem 18.2), so we include it here.[3]

Lemma 18.12. *If A is an integrally closed Noetherian domain of dimension one then all of its localizations at nonzero prime ideals are discrete valuation rings.*[4]

Proof. Let F be the fraction field of A and let $\mathfrak{p}$ be a nonzero prime ideal. We first note that $A_\mathfrak{p}$ is integrally closed. Indeed, if $x^n + a_{n-1}x^{n-1} + \cdots + a_0 = 0$ is an equation with $a_i \in A_\mathfrak{p}$ and $x \in F$, then we may pick $s \in A - \mathfrak{p}$ so that all the sa_i lie in A (let s be the product of all the denominators $c_i \notin \mathfrak{p}$ of $a_i = b_i/c_i$). Multiplying through by s^n yields an equation $(sx)^n + sa_{n-1}(sx)^{n-1} + \cdots + s^n a_0 = 0$ in $y = sx$ with coefficients in A. Since A is integrally closed, $y \in A$, therefore $x = y/s \in A_\mathfrak{p}$ as desired.

[3] There are plenty of shorter proofs, but they tend to use facts that we have not proved.

[4] Such rings are called *Dedekind domains*. They play an important role in number theory where they appear as the ring of integers of a number field. The key property of a Dedekind domain is that ideals can be uniquely factored into prime ideals, although we don't use this here.

Let $\mathfrak{m} = \mathfrak{p}A_{\mathfrak{p}}$ be the maximal ideal of $A_{\mathfrak{p}}$. The ring $A_{\mathfrak{p}}$ has dimension one, since $(0) \subsetneq \mathfrak{m}$ are all the prime ideals in $A_{\mathfrak{p}}$ (otherwise we would have a nonzero prime $\mathfrak{q}$ properly contained in $\mathfrak{m}$, but then $\mathfrak{q} \cap A$ would be a nonzero prime properly contained in $\mathfrak{p}$, contradicting $\dim A = 1$). Thus $R = A_{\mathfrak{p}}$ is a local ring of dimension one. By Theorem 18.2, to show that R is a DVR it suffices to prove that R is regular; it is clear that R is Noetherian (since A is), we just need to show $\dim \mathfrak{m}/\mathfrak{m}^2 = \dim R = 1$. By Nakayama's lemma, $\mathfrak{m}^2 \neq \mathfrak{m}$, so $\dim \mathfrak{m}/\mathfrak{m}^2 \neq 0$. To show $\dim \mathfrak{m}/\mathfrak{m}^2 = 1$ it suffices to prove that $\mathfrak{m}$ is principal. To do this we adapt an argument of Serre from [1, §I.1].

Let $S = \{y \in F : y\mathfrak{m} \subseteq R\}$, and let $\mathfrak{m}S$ denote the R-ideal generated by all products xy with $x \in \mathfrak{m}$ and $y \in S$ (just like an ideal product). Then $\mathfrak{m} \subseteq \mathfrak{m}S \subseteq R$, so either $\mathfrak{m}S = \mathfrak{m}$ or $\mathfrak{m}S = R$. We claim that the latter holds. Assuming it does, then $1 = \sum x_i y_i$ for some $x_i \in \mathfrak{m}$ and $y_i \in S$. The products $x_i y_i$ all lie in R but not all can lie in $\mathfrak{m}$, so some $x_j y_j$ is invertible. Set $x = x_j/(x_j y_j)$ and $y = y_j$ so that $xy = 1$, with $x \in \mathfrak{m}$ and $y \in S$. We can then write any $z \in \mathfrak{m}$ as $z = 1 \cdot z = xy \cdot z = x \cdot yz$. But $yz \in R$, since $y \in S$, so every $z \in \mathfrak{m}$ actually lies in (x). Thus $\mathfrak{m} = (x)$ is principal as desired, assuming $\mathfrak{m}S = R$.

We now prove that $\mathfrak{m}S = R$ by supposing the contrary and deriving a contradiction. We will do this by proving that $\mathfrak{m}S = \mathfrak{m}$ implies both $S \subseteq R$ and $S \not\subseteq R$. So assume $\mathfrak{m}S = \mathfrak{m}$.

We first prove $S \subseteq R$. Since $\mathfrak{m}S = \mathfrak{m}$, for any $\lambda \in S$ we have $\lambda\mathfrak{m} \subseteq \mathfrak{m}$. The ring R is Noetherian, so let $m_1, \ldots, m_k$ be generators for $\mathfrak{m}$. We then have k equations of the form $\sum_{i,j} a_{ij} m_j = \lambda m_i$ with $a_{ij} \in R$. Thus λ is an eigenvalue of the matrix (a_{ij}) and therefore a root of its characteristic polynomial, which is monic, with coefficients in R. Since R is integrally closed, $\lambda \in R$, and therefore $S \subseteq R$ as claimed.

We now prove $S \not\subseteq R$, thereby obtaining a contradiction. Let $x \in \mathfrak{m} - \{0\}$, and consider the ring $T_x = \{y/x^n : y \in R, n \geq 0\}$. We claim $T_x = F$: if not, it contains a nonzero maximal ideal $\mathfrak{q}$ with $x \notin \mathfrak{q}$ (since x is a unit in T_x), so $\mathfrak{q} \cap R \neq \mathfrak{m}$, and clearly $\mathfrak{q} \cap R \neq (0)$, but then $\mathfrak{q} \cap R$ is a prime ideal of R strictly between (0) and $\mathfrak{m}$, which contradicts $\dim R = 1$. So every element of $T_x = F$ can be written in the form y/x^n, and this holds for any $x \in \mathfrak{m}$. Applying this to a fixed $1/z$ with $z \in \mathfrak{m} - \{0\}$, we see that every $x \in \mathfrak{m} - \{0\}$ satisfies $x^n = yz$ for some $y \in R$ and $n \geq 0$, thus $x^n \in (z)$ for all $x \in \mathfrak{m}$ and sufficiently large n. Applying this to our generators $m_1, \ldots, m_k$ for $\mathfrak{m}$, choose n so that $m_1^n, \ldots, m_k^n \in (z)$, and then let $N = kn$ so that $(\sum_i r_i m_i)^N \in (z)$ for all choices of $r_i \in R$. Thus $\mathfrak{m}^n \subseteq (z)$ for all $n \geq N$, and there is some minimal $n \geq 1$ for which $\mathfrak{m}^n \subseteq (z)$. If $n = 1$ then $\mathfrak{m} = (z)$ is principal and we are done. Otherwise, choose $y \in \mathfrak{m}^{n-1}$ so that$y \notin (z)$ but $y\mathfrak{m} \subseteq (z)$. Then $(y/z)\mathfrak{m} \in R$, so $y/z \in S$, but $y/z \notin R$ (since $z \in \mathfrak{m}$), so $S \not\subseteq R$ as claimed. □

We are now ready to prove our main theorem.

Theorem 18.13. *Every abstract curve is isomorphic to a smooth projective curve.*

Proof. Let $X = X_F$ be the abstract curve associated to the function field F/k. Then $\mathcal{O}(X)$ is an affine algebra, and there is a corresponding affine curve A. The curve A is smooth, since all its local rings $\mathcal{O}_P$ are discrete valuation rings, but it is not complete, so not every point on X (each corresponding to a discrete valuation rings of F/k) corresponds to a point on A. So let C be the projective closure of A; the curve C need not be smooth, but it is complete, and it satisfies Chevalley's criterion. Thus for each point $P \in X$, the associated discrete valuation ring $\mathcal{O}_{P,X}$ contains the local ring $\mathcal{O}_{Q,C}$ of a point $Q \in C_1$. The point Q is certainly unique; if $\mathcal{O}_{P,X}$ contained two distinct local rings it would contain the entire function field, which is not the case (to see this, note that for any distinct $P, Q \in C$ the zero locus of $m_P + m_Q$ is empty).

So let $\phi\colon X \to C$ map each $P \in X$ to the unique $Q \in C_1$ for which $\mathcal{O}_{Q,C} \subseteq \mathcal{O}_{P,X}$. It is easy to see that ϕ is continuous; indeed, since we are in dimension one it suffices to note that it is surjective, and this is so: every local ring $\mathcal{O}_{Q,C}$ is contained in a discrete valuation ring $\mathcal{O}_{P,X}$ (possibly more than one, this can happen if Q is singular).[5] To check that it is a morphism, if $U \subseteq C$ is open and $f \in \mathcal{O}_C(U) = \cap_{Q\in U}\mathcal{O}_{Q,C}$ then we have $\mathcal{O}_X(\phi^{-1}(U)) = \cap_{\phi(P)\in U}\mathcal{O}_{P,X} \supseteq \cap_{Q\in U}\mathcal{O}_{Q,C}$ and therefore $f \circ \phi \in \mathcal{O}_X(V)$ as required.

Now let $C_1 = C$ and $\phi_1 = \phi$. There are finitely many singular points $Q \in C$ (the singular locus has dimension 0), and for each such Q the inverse image $\phi^{-1}(Q) \subsetneq X$ is closed and not equal to X (since ϕ is surjective and C has more than one point), so finite. Let $P_2, \ldots, P_n \in X$ be the finite list of points whose images under ϕ_1 are singular in C.

For each P_i we now let C_i be the projective closure of the smooth affine curve with function field F/k and a local ring $\mathcal{O}_{P,C_i}$ equal to $\mathcal{O}_{P_i,X}$, given by Lemma 18.11. Then $k(C_i) = F$ and the point on C_i corresponding to P_i is smooth by construction, since its local ring is precisely the discrete valuation ring $\mathcal{O}_{P_i}$. Define a surjective morphism $\phi_i\colon X \to C_i$ exactly as we did for ϕ_1.

We now consider the product variety $Y = \prod_i C_i$ and define the morphism $\varphi\colon X \to Y$ by $\varphi(P) = (\phi_1(P), \ldots, \phi_n(P))$. The variety Y is a product of projective varieties and can be smoothly embedded in a single projective space.[6] The image of φ in Y is a projective curve C whose function field is isomorphic to F, and C is smooth because, by construction, every point $P \in C$ is smooth in one of its affine parts. By Theorem 18.10, the smooth projective curve C is isomorphic to the abstract curve associated to its function field, namely, X. □

Corollary 18.14. *Every curve C is birationally equivalent to a smooth projective curve that is unique up to isomorphism.*

Proof. By 18.10 there exists an abstract curve corresponding to the function field $k(C)$, and by Theorem 18.13 this abstract curve is isomorphic to a smooth projective curve. Uniqueness follows from Corollary 18.8. □

The smooth projective curve to which a given curve C is birationally equivalent is called the *desingularization* C. Henceforth, whenever we write down an equation for a curve (which may be affine and/or have singularities) we can always assume that we are referring to its desingularization.

Remark 18.15. In the proof of Theorem 18.13 we made no attempt to control the dimension of the projective space into which we embedded the smooth projective curve C isomorphic to our abstract curve X. Using more concrete methods, one can show that it is always possible to embed C in $\mathbb{P}^3$. In general, one can do no better than this; indeed we will see plenty of examples of smooth projective curves that cannot be embedded in $\mathbb{P}^2$.

References

[1] J. P. Serre, *Local fields*, Springer, 1979.

[5] This follows from Problem 2 part 5 on Problem Set 8.

[6] Using the Segre morphism, as proved in problem 1 part 4 on Problem Set 8.

18.782 Introduction to Arithmetic Geometry **Fall 2013**
Lecture #19 **11/12/2013**

In this lecture (and henceforth) k denotes a perfect but not necessarily algebraically closed field, and $\bar{k}$ denote a fixed algebraic closure of k.

19.1 Curves and function fields

Henceforth we adopt the following definitions.

Definition 19.1. A *curve* C/k is a smooth projective variety of dimension one defined over k. A *function field* F/k is a finitely generated extension of k with transcendence degree one, such that k algebraically closed in F.

Other authors distinguish the curves we have defined as *nice curves*: one-dimensional varieties that are smooth, projective, and geometrically irreducible (irreducible over $\bar{k}$); for us varieties are geometrically irreducible by definition, so this last requirement is automatic. But perhaps a more fundamental characterization is that nice curves are isomorphic to the abstract curve defined by their function field.

In our definition of a function field, the requirement that k be algebraically closed in F is not a serious restriction, it is automatically satisfied when F is the function field of a curve C/k, so it is necessary for us to obtain the following equivalence of categories.[1]

Theorem 19.2. *The category of curves C/k with nonconstant morphisms and the category of function fields F/k with field homomorphisms that fix k are contravariantly equivalent under the functor that sends a curve C to the function field $k(C)$ and a nonconstant morphism of curves $\phi\colon C_1 \to C_2$ defined over k to the field homomorphism $\phi^*\colon k(C_2) \to k(C_1)$ defined by $\phi^* f = f \circ \phi$.*

Proof. For $k = \bar{k}$ this follows from: (1) a nonconstant morphism of smooth projective curves is surjective (Corollary 18.7), (2) a smooth projective curve is isomorphic to the abstract curve defined by its function field (Theorem 18.10). The inverse functor sends F/k to the smooth projective curve isomorphic to the abstract curve defined by F/k (Theorem 18.13), which is unique up to isomorphism (Corollary 18.8).

For $k \neq \bar{k}$, recall from Lecture 15 that if C_1 is defined over k and the morphism $\phi\colon C_1 \to C_2$ is defined over k then the induced morphism of function fields $\bar{k}(C_2) \to \bar{k}(C_1)$ restricts to a morphism $k(C_2) \to k(C_1)$. Conversely, given a function field F/k with k a perfect field, by Theorem 17.8 and Remark 17.9, we can write $F = k(x, \alpha)$, with α algebraic over the rational function field $k(x)$. If we then consider the minimal polynomial of α as an element of $k(x)[y]$ and clear denominators in the coefficients, we obtain an irreducible polynomial $f \in k[x, y]$. Because k is algebraically closed in F, this polynomial remains irreducible as an element of $\bar{k}[x, y]$ (by [1, III.3.6.8]), and therefore defines an affine variety of dimension one in $\mathbb{A}^2$ whose ideal is generated by $f \in k[x, y]$. We may then take C/k to be the (projective) desingularization of this affine variety, which is still defined over k.[2] □

[1] This follows from [1, III.3.6.8] which implies that F/k is the function field of some curve C/k if and only if k is algebraically closed in F (we can always use F to construct an algebraic set, but if k is not algebraically closed in F this set will not be irreducible (over $\bar{k}$), hence not a variety).

[2] The fact that the desingularization is defined over k is not obvious from our proof of its existence, but it can be proved by other means (the assumption that k is perfect is necessary).

Andrew V. Sutherland

From Theorem 19.2 we see that the study of curves and the study of function fields are one and the same, a fact that we shall frequently exploit by freely moving between the two categories. It is worth noting that this categorical equivalence does not hold for varieties of dimension greater than one.

Definition 19.3. The *degree* of a morphism of curves $\phi\colon C_1 \to C_2$ is the degree of the corresponding extension of function fields $\deg\phi = [k(C_1) : \phi^*(k(C_2))]$.

Remark 19.4. A note of caution. Since the field homomorphism $\phi^*\colon k(C_2) \to k(C_1)$ is necessarily injective, it is standard practice to identify $k(C_2)$ with its image in $k(C_1)$. Under this convention, one may then write $\deg\phi = [k(C_1) : k(C_2)]$. But the notation $[L : K]$ for the degree of a field extension L/K is ambiguous if K is simply a field embedded in L, rather than an actual subfield. Without knowing the embedding, there is in general no way to know what $[L : K]$ actually is!

This does not cause a problem for number fields, but function fields are another story; there are many different ways to embed one function field into another, and different embeddings may have different degrees. As a simple example, consider the map $\varphi\colon k(x) \to k(x)$ that sends x to x^2 and fixes k. The image of φ is a proper subfield of $k(x)$ (namely, $k(x^2)$) which is isomorphic to $k(x)$ but not equal to $k(x)$ as a subfield. Indeed, as a $k(x^2)$-vector space, $k(x)$ has dimension 2, and we have $[k(x) : \varphi(k(x))] = \deg\varphi = 2$ as expected. But if we identify $k(x)$ with its image $\varphi(k(x))$ then we would write $[k(x) : k(x)] = 2$, which is confusing to say the least.

Corollary 19.5. *A morphism of curves is an isomorphism if and only if its degree is one.*

19.2 Divisors

Definition 19.6. Let C/k be a curve with $k = \bar{k}$. A *divisor* of C is a formal sum

$$D := \sum_{P\in C} n_P P$$

with $n_P \in \mathbb{Z}$ and all but finitely many $n_P = 0$. The set of points P for which $n_P \neq 0$ is called the *support* of D. The divisors of C form a free abelian group under addition, the *divisor group* of C, denoted $\operatorname{Div} C$.

Definition 19.7. Let F/k be a function field with $k = \bar{k}$. A *divisor* of F is a formal sum

$$D := \sum_{P\in X_F} n_P P$$

with $n_P \in \mathbb{Z}$ and all but finitely many $n_P = 0$. Here X_F denotes the abstract curve defined by F/k, whose points P are the maximal ideals of the discrete valuation rings of F/k. The divisors of F form a free abelian group $\operatorname{Div} F$ under addition; if C is the smooth projective curve with function field F, this group is isomorphic to $\operatorname{Div} C$ and we may use them interchangeably.

We now want to generalize to the case where k is not necessarily algebraically closed. Let $G_k = \operatorname{Gal}(\bar{k}/k)$ be the absolute Galois group of k (as usual $\bar{k}$ is a fixed algebraic closure).

Definition 19.8. A divisor $D = \sum n_P P \in \operatorname{Div} C$ is *defined over* k if for all $\sigma \in G_k$ we have $D^\sigma = D$, where

$$D^\sigma = \left(\sum n_P P\right) = \sum n_P P^\sigma.$$

The subset of $\operatorname{Div} C$ defined over k forms the subgroup of *k-rational divisors*.

Note that $D = D^\sigma$ does not necessarily imply $P = P^\sigma$ for all P in the support of D. But if $D = D^\sigma$ for all $\sigma \in G_k$, then it must be the case that $n_{P^\sigma} = n_P$ for all $\sigma \in G_k$. Thus we can group the terms of a k-rational divisor into G_k-orbits with a single coefficient n_P applied to all the points in the orbit. Equivalently, we can view a k-rational divisor as a sum over G_k-orbits of points $P \in C(\bar{k})$. This turns out to be a better way of defining the group of k-rational divisors on a curve that is defined over k.

Definition 19.9. Let C/k be a curve defined over k. The G_k-orbits of $C(\bar{k})$ are called *closed points*, which we also denote by P. A *rational divisor* of C/k is a formal sum

$$D := \sum n_P P$$

where P ranges over the closed points of C/k, with $n_P \in \mathbb{Z}$ and all but finitely many $n_P = 0$. The group of rational divisors on C is denoted $\operatorname{Div}_k C$.

For function fields F/k the divisor class group is defined exactly as when $k = \bar{k}$. As before, X_F is the set of maximal ideals of discrete valuation rings of F/k, which we shall henceforth call *places* of F/k in order to avoid confusion. The places of F/k are in one-to-one correspondence with the closed points of the corresponding curve C/k. In the case that $k = \bar{k}$ this follows from the *Nullstellensatz*, each point P on C is the zero locus of a place of F/k (the maximal ideal m_P), and vice versa. When k is not algebraically closed the same statement still holds, provided we replace "point" with "closed point". To see this, just apply the action of G_k to a point $P \in C(\bar{k})$ and its corresponding maximal ideal $M_P \in \bar{k}[x_1, \ldots, x_n]$; the G_k-orbit of P is a closed point of C/k and the intersection of $k[x_1, \ldots, x_n]$ with the union of the ideals in the G_k-orbit of M_P is a maximal ideal of $k[x_1, \ldots, x_n]$ whose reduction modulo $I(C)$ is a place of F/k.

Remark 19.10. Be sure not to confuse the closed points of C/k with the set of rational points $C(k)$. The points in $C(k)$ correspond to a proper subset of the set of closed points, the trivial G_k-orbits that consist of a single element. But every point in $C(\bar{k})$ is contained in a closed point of C/k. Indeed, this is the key advantage to working with closed points; they contain all the essential information about C/k (even in cases where $C(k)$ is empty), while allowing us to work over k rather than $\bar{k}$, which has both theoretical and practical advantages. But it is important to remember that the set of closed points depends on the ground field k, not just C. We will consistently write C/k to remind ourselves of this fact. In more advanced treatments one writes C_k and regards C_k and $C_{k'}$ as distinct objects for any extension k'/k, even when the equations defining C are exactly the same; switching from C_k to $C_{k'}$ is known as *base extension*.

Definition 19.11. Let f be a nonzero element of a function field F/k. The *divisor of* f is

$$\operatorname{div} f := \sum_{P \in X_F} \operatorname{ord}_P(f) P.$$

Such divisors are said to be *principal.*[3]

[3] The principal divisor $\operatorname{div} f$ is also often denoted by (f), but we will not use this notation.

In order for the definition above to make sense, we need to know that $\operatorname{ord}_P(f)$ is zero for all but finitely many P. Under our categorical equivalence, we can assume that $F = k(C)$ for some curve C/k, which makes this easy to prove. Note that for any closed point, a function $f \in k(C)$ vanishes at a point $P \in C(\bar{k})$ if and only if it vanishes on the entire G_k-orbit of P. Thus it makes sense to say whether a closed point P of C/k lies in the zero locus of f or not.

Theorem 19.12. *Let F/k be a function field. For any $f \in F^\times$ we have $\operatorname{ord}_P(f) = 0$ for all but finitely many places P of F.*

Proof. Let C be the smooth projective curve with function field $k(C) \simeq F$, and let us identify F with $k(C)$. Let f be a nonzero element of the coordinate ring $k[C]$. We then have $\operatorname{ord}_P(f) = 0$ unless the closed point P lies in the zero locus of f. But the zero locus of f is a closed set properly contained in the one-dimensional variety C (since $f \neq 0$), hence finite. The general case $f = g/h \in k(C)$ is similar, now $\operatorname{ord}_P(f) = 0$ unless P is in the zero locus of either g or h, both of which are finite. □

A sum of principal divisors is a principal divisor, since

$$\operatorname{div} f + \operatorname{div} g = \operatorname{div} fg$$

(this follows from the fact that each $\operatorname{ord}_P\colon k(C)^\times \to \mathbb{Z}$ is a homomorphism). We also have $\operatorname{div} 1 = 0$, thus the map $k(C)^\times \to \operatorname{Div}_k C$ defined by $f \mapsto \operatorname{div} f$ is a group homomorphism. Its image is $\operatorname{Princ}_k C$, the group of k-rational principal divisors of C.

Definition 19.13. Let C/k be a curve. The quotient group

$$\operatorname{Pic}_k C := \operatorname{Div}_k C / \operatorname{Princ}_k C$$

is the *Picard group* of C (also known as the *divisor class group* of C). Elements D_1 and D_2 of $\operatorname{Div}_k C$ that have the same image in $\operatorname{Pic}_k C$ are said to be *linearly equivalent*. We write $D_1 \sim D_2$ to indicate this equivalence.

Theorem 19.14. *We have an exact sequence*

$$1 \to k^\times \to k(C)^\times \xrightarrow{\operatorname{div}} \operatorname{Div}_k C \to \operatorname{Pic}_k C \to 0.$$

Proof. The only place where exactness is not immediate from the definitions is at $k(C)^\times$; we need to show that $\ker \operatorname{div} = k^\times$. It is clear that $k^\times$ lies in $\ker \operatorname{div}$; any $f \in k^\times$ lies in the unit group of every discrete valuation ring of $k(C)/k$, in which case $\operatorname{ord}_P(f) = 0$ for all P. Equality follows from the fact that k is algebraically closed in $k(C)$. This means that k is equal to the full field of constants of the function field $k(C)/k$, which is precisely the intersection of the unit groups of all the valuation rings of $k(C)/k$, equivalently, the set of functions $f \in k(C)$ for which $\operatorname{ord}_P(f) = 0$ for all P (another way to see this is to note that if f is nonzero on every closed point of C/k, then the zero locus of f is the empty set and therefore f is a unit in the coordinate ring). □

Definition 19.15. For a principal divisor $\operatorname{div} f = \sum n_P P$, the divisors

$$\operatorname{div}_0 f = \sum_{n_P > 0} n_P P \quad \text{and} \quad \operatorname{div}_\infty f = \sum_{n_P < 0} -n_P P$$

are called the *divisor of zeros* and the *divisor of poles* of f, respectively. We have

$$\operatorname{div} f = \operatorname{div}_0 f - \operatorname{div}_\infty f.$$

The quantities $\deg \operatorname{div}_0 f$ and $\deg \operatorname{div}_\infty f$ count the zeros and poles of f, with appropriate multiplicities. While is is intuitively clear that these two quantities should be equal (recall that we can represent f as the ratio of two homogeneous polynomials of the same degree), to prove this rigorously we will establish a more general result that tells us that the fibers (inverse images) of a morphism of curves ϕ all have cardinality equal to $\deg \phi$, provided that we count the points in each fiber with the correct multiplicities.

Remark 19.16. In what follows we work exclusively with morphisms $\phi\colon C_1 \to C_2$ defined over k, by which we mean that both the curves *and the morphism* are over k. There are situations where one does want to consider morphisms that are note defined over k (even though the curves are defined over k), but in order to keep things simple we will not consider this at this stage (we can always base extend to a field where everything is defined).

Lemma 19.17. *Let $\phi\colon C_1 \to C_2$ be a morphism of curves defined over k and let P be a closed point of C_1/k. Then $\phi(P)$ is a closed point of C_2/k.*

Proof. Let P be the G_k-orbit $\{P_1, \ldots, P_d\}$, where $d = \deg P$. We have $\phi(P_i)^\sigma = \phi(P_i^\sigma)$ for all $\sigma \in G_k$, since ϕ is defined over k, and it follows that the set $\phi(P) = \{\phi(P_1), \ldots, \phi(P_d)\}$ is fixed by G_k, hence a union of G_k-orbits. For each P_i we have $P_i = P_1^\sigma$ for some $\sigma \in G_k$, and it follows that $\phi(P_i) = \phi(P_1^\sigma) = \phi(P_1)^\sigma$, show every $\phi(P_i)$ is in the G_k-orbit of $\phi(P_1)$, so $\phi(P)$ consists of a single G_k-orbit and is a closed point. □

With Lemma 19.17 in hand, we can now sensibly speak of a morphism $\phi\colon C_1 \to C_2$ defined over k as a map of closed points.

Definition 19.18. Let $\phi\colon C_1 \to C_2$ be a morphism defined over k, and $\phi^*\colon k(C_2) \to k(C_1)$ the corresponding morphism of function fields. The *ramification index* (also called the *ramification degree*) of ϕ at a closed point P of C_1 (equivalently, a place P of $k(C_1)$) is

$$e_\phi(P) := \operatorname{ord}_P(\phi^* t_Q),$$

where $t_Q \in k(C_2)$ is a uniformizer at $Q = \phi(P)$, that is, a generator for the place Q of $k(C_2)$. If $e_\phi(P) = 1$, then ϕ is *unramified at* P, and if $e_\phi(P) = 1$ for all closed points P of C_1/k we say that ϕ is *unramified.*

Definition 19.19. Let $\phi\colon C_1 \to C_2$ be a morphism defined over k. The *pullback map* ϕ^* on divisors is the homomorphism $\phi^*\colon \operatorname{Div}_k C_2 \to \operatorname{Div}_k C_1$ defined by

$$\phi^*(Q) := \sum_{P \in \phi^{-1}(Q)} e_\phi(P) P,$$

where (Q) denotes the divisor in $\operatorname{Div}_k C_2$ with support $\{Q\}$ and $n_Q = 1$. We also define the *pushforward map* ϕ_* on divisors as the homomorphism $\phi_*\colon \operatorname{Div} C_1 \to \operatorname{Div} C_2$ defined by

$$\phi_*(P) = [k(P) : \phi^*(k(P))]\phi(P) = \frac{\deg P}{\deg \phi(P)} \phi(P).$$

When $k = \bar{k}$ is algebraically closed, the pushforward map just sends the divisor (P) to the divisor $(\phi(P))$, but in general we want to scale things so that $\deg \phi_*(P) = \deg P$.

It is clear that both ϕ^* and ϕ_* are group homomorphisms, and if ϕ is unramified then for all divisors D we have

$$\phi_*(\phi^*(D)) = \deg(\phi)D.$$

You will prove on the problem set that in fact this is true regardless; the composition $\phi_* \circ \phi^*$ corresponds to multiplication by $\deg(\phi)$ on $\operatorname{Div}_k C_2$.

Remark 19.20. Using ϕ^* to denote both the pullback map $\operatorname{Div}_k C_2 \to \operatorname{Div}_k C_1$ and the dual morphism $k(C_2) \to k(C_1)$ of function fields induced by $\phi\colon C_1 \to C_2$ might seem like an unfortunate collision of notation, but it is standard and intentional. Recall that the kernel of the divisor map $\operatorname{div}\colon C \to \operatorname{Div}_k C$ is just $k^\times$, so up to scalars we can identify a function $f \in k(C)$ with the corresponding divisor $\operatorname{div} f \in \operatorname{Div}_k C$. The pullback map ϕ^* maps principal divisors to prinicipal divisors, thus for any $f, g \in k(C_2)^\times$ we have

$$\phi^* \operatorname{div} f = \phi^* \operatorname{div} g \quad \Longleftrightarrow \quad \phi^* f = \lambda \phi^* g \text{ for some } \lambda \in k^\times.$$

Definition 19.21. Let C/k be a curve and let F/k be the corresponding function field. If P is a closed point of C/k, or a place of F/k, we define the *degree* of P to be the dimension of the *residue field* $k(P) = \mathcal{O}_P/m_P$ over k (where $m_P = P$ if P is a place of F/k), that is,

$$\deg P := [k(P) : k].$$

Equivalently, $\deg P$ is the cardinality of the closed point P as a G_k-orbit of points in $C(\bar{k})$ (see [1, Cor. 3.6.5] for a proof of this equivalence, which depends on the fact that k is a perfect field). The degree of a divisor $D = \sum n_P P$ in the group of k-rational divisors is

$$\deg D := \sum n_P \deg P.$$

Note that when $k = \bar{k}$, we have $\deg P = 1$ for all P, so in this case $\deg D = \sum n_P$.

Theorem 19.22. *Let $\phi\colon C_1 \to C_2$ be a morphism of curves defined over k. Then for each closed point Q of C_2/k,*

$$\deg \phi^*(Q) = \deg \phi \deg Q$$

Here ϕ^* is the pullback map on divisors. This theorem effectively says that the fibers (inverse images of points) of the morphism ϕ all have cardinality equal $\deg \phi$, provided that we count them correctly. Our definition of the degree of a divisor accounts for the size of the Galois orbit corresponding to a closed point (so we are effectively counting $\bar{k}$-points on both sides), and the ramification index e_ϕ incorporated in the definition of the pullback map ϕ^* correctly accounts for ramification.

We will prove Theorem 19.22 in the next lecture. Let us end this lecture by proving that any nonzero function on a curve has the same number of zeros and poles. The proof is essentially immediate from the definitions; in an advanced text it might be written in one line or simply left to the reader. But we will take the time to unravel all the definitions in gory detail, as this provides an excellent opportunity to check our understanding.

Corollary 19.23. *Let $f \in k(C)^\times$ for some curve C/k. Then f has the same number of zeros and poles (counted with multiplicity), that is,*

$$\deg \operatorname{div}_0 f = \deg \operatorname{div}_\infty f.$$

If $f \in k^\times$ then this number is 0, and otherwise it is equal to $[k(C) : k(f)]$.
In any case, we always have $\deg \operatorname{div} f = 0$.

Proof. For $f \in k^\times$ we have $\operatorname{div} f = 0$ and the corollary holds. Otherwise f is transcendental over k (because k is algebraically closed in $k(C)$), and it defines a morphism $f\colon C \to \mathbb{P}^1$ as follows: if $f = g/h$ with $g, h \in k[C]$ represented by homogeneous functions of the same degree, with h nonzero, then the morphism f is given by $(g : h)$.[4] Recall that this represents an equivalence class of tuples that we can scale by any $\lambda \in k(C)^\times$.

Let $(x : y)$ be homogeneous coordinates for $\mathbb{P}^1$, and define $0 = (0 : 1)$ and $\infty = (1 : 0)$. Note that 0 and ∞ are both rational points on $\mathbb{P}^1$, hence we may identify them with the corresponding closed points (they are each the unique element of their G_k-orbit).

The place of $k(\mathbb{P}^1)$ corresponding to the closed point 0 is (the maximal ideal of) the discrete valuation that measures the order of vanishing of a homogeneous rational function $r(x, y)$ at $(0 : 1)$, equivalently, it measures the order of vanishing of $r(x/y, 1)$ at $0/1$. Similarly, the place corresponding to ∞ measures the order of vanishing of $r(x/y, 1)$ at $1/0$, equivalently, the order of vanishing of $r(1, y/x)$ at $0/1$.

The obvious choice of uniformizers for the places 0 and ∞ are the functions $t_0 = x/y$ and $t_\infty = y/x$. The images of these uniformizers under the field embedding $f^*\colon k(\mathbb{P}^1) \to k(C)$ induced by f are, by definition,

$$\begin{aligned} f^* t_0 &= t_0 \circ f = g/h = f, \\ f^* t_\infty &= t_\infty \circ f = h/g = 1/f. \end{aligned}$$

Now let us consider a closed point P of C for which $f(P) = 0$ (so $f(P') = 0$ for any/all points P' in the G_k-orbit P). The ramification index of f at P is, by definition,

$$e_f(P) = \operatorname{ord}_P(f^* t_0) = \operatorname{ord}_P(f).$$

If we instead consider a closed point P of C for which $f(P) = \infty$, we then have

$$e_f(P) = \operatorname{ord}_P(f^* t_\infty) = \operatorname{ord}_P(1/f) = -\operatorname{ord}_P(f).$$

Applying the pullback map $f^*\colon \operatorname{Div}_k \mathbb{P}^1 \to \operatorname{Div}_k C$ to the divisor (0) yields[5]

$$f^*(0) = \sum_{f(P)=0} e_f(P) P = \sum_{f(P)=0} \operatorname{ord}_P(f) P.$$

But notice that the places P of $k(C)$ where f has positive valuation correspond exactly to the closed points P of C/k where $f(P) = 0$ (hence we use the same symbol P in both cases). Thus, by definition,

$$f^*(0) = \sum_{f(P)=0} \operatorname{ord}_P(f) P = \operatorname{div}_0 f$$

Similarly, the places where f has negative valuation are those where $f(P) = \infty = (1 : 0)$, equivalently, $(1/f)(P) = 0 = (0 : 1)$. Thus

$$f^*(\infty) = \sum_{f(P)=\infty} e_f(P) P = \sum_{f(P)=\infty} -\operatorname{ord}_P(f) P = \operatorname{div}_\infty f.$$

[4] How do we know f is a morphism? Because every rational map from a (smooth projective) curve to a projective variety is a morphism; see Corollary 18.7.

[5] Note that this is not the zero element of $\operatorname{Div}_k C$, which is the divisor whose support is the empty set. The divisor (0) has support $\{0\}$.

Applying Theorem 19.22 to $f\colon C \to \mathbb{P}^1$ with $Q = 0$ and $Q = \infty$ (both of which have degree one, since they are rational points), we have

$$\deg \operatorname{div}_0 f = \deg f^*(0) = \deg f \deg 0 = \deg f = \deg f \deg \infty = \deg f^*(\infty) = \deg \operatorname{div}_\infty f,$$

where $\deg f = [k(C) : f^*(k(\mathbb{P}^1))]$ is the degree of f as a morphism.[6] We know that $k(\mathbb{P}^1)$ is isomorphic to the field of rational functions $k(t)$, thus the image of $f^*\colon k(\mathbb{P}^1) \to k(C)$ is completely determined by the image of t (since f must fix k), and we have $f^*(t) = t \circ f = f$, so $\deg f = [k(C) : k(f)]$. Finally, we note that

$$\deg \operatorname{div} f = \deg(\operatorname{div}_0 f - \operatorname{div}_\infty f) = \deg \operatorname{div}_0 f - \deg \operatorname{div}_\infty f = 0$$

as claimed. □

References

[1] H. Stichtenoth, *Algebraic function fields and codes*, Springer, 2009.

[6] This is almost certainly *not* the degree of the homogeneous polynomials g and h we chose to represent f. These were chosen arbitrarily and could have any degree; they are only defined modulo the equivalence relation on rational maps and modulo the ideal $I(C)$ in any case.

20.1 Degree theorem for morphisms of curves

Let us restate the theorem given at the end of the last lecture, which we will now prove.

Theorem 20.1. *Let $\phi\colon C_1 \to C_2$ be a morphism of curves defined over k. Then for each closed point Q of C_2/k,*

$$\deg \phi^*(Q) = \deg \phi \deg Q$$

Before beginning the proof, let us first show that we can assume without loss of generality that k is algebraically closed. If the closed point Q is the G_k-orbit $\{Q_1, \ldots, Q_d\}$, with $d = \deg Q$, after base extension to $\bar{k}$ we have

$$\deg \phi^*(Q) = \deg \phi^*(Q_1 + \cdots + Q_d) = \deg \phi^*(Q_1) + \cdots + \deg \phi^*(Q_d),$$

since both the degree map and the pullback map $\phi^*\colon \mathrm{Div}_{\bar{k}}(C_2) \to \mathrm{Div}_{\bar{k}}(C_1)$ are group homomorphisms. If we assume the theorem holds over $\bar{k}$, then every term on the right is equal to $\deg \phi$ and the sum is $d \deg \phi = \deg \phi \deg Q$.

We now prove the theorem assuming $k = \bar{k}$, following the approach of [1, III.2].

Proof of Theorem 20.1. Fix $Q \in C_2$, and let $\mathcal{O}_Q$ be its local ring of regular functions. The set $\phi^{-1}(Q)$ is finite because ϕ is not constant and C_1 is an irreducible algebraic set of dimension one (so all its proper closed subsets are finite). Let $P_1, \ldots, P_n \in C_1$ be the elements of $\phi^{-1}(Q)$, let $\mathcal{O}_1, \ldots, \mathcal{O}_n$ be the corresponding local rings of regular functions, and define

$$\mathcal{O} = \bigcap_{i=1}^{n} \mathcal{O}_i.$$

By Lemma 20.4 below, there exist uniformizers $t_1, \ldots, t_n$ for $\mathcal{O}_1, \ldots \mathcal{O}_n$ such that

$$\mathrm{ord}_{P_i}(t_j) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

The maximal ideals of $\mathcal{O}$ are $(t_1), \ldots, (t_n)$ and each nonzero $f \in \mathcal{O}$ factors uniquely as

$$f = u t_1^{e_1} \cdots t_n^{e_n},$$

with $u \in \mathcal{O}^\times$ and $e_i = \mathrm{ord}_{P_i}(f)$.

Under the map $\phi^*\colon k(C_2) \to k(C_1)$, for any $f \in \mathcal{O}_Q$ we have

$$\mathrm{ord}_{P_i}(\phi^* f) = \mathrm{ord}_{P_i}(f \circ \phi) = \mathrm{ord}_Q(f) \geq 0,$$

thus $\phi^*(\mathcal{O}_Q)$ is a subring of $\mathcal{O}$. If we now let t_Q be a uniformizer for $\mathcal{O}_Q$, and put $t = \phi^* t_Q$ we have

$$t = \phi^* t_Q = u t_1^{e_1} \cdots t_n^{e_n}$$

where $e_i = \mathrm{ord}_{P_i}(\phi^* t_Q) = e_\phi(P_i)$. Since $t_1, \ldots, t_n$ are pairwise relatively prime (meaning that $(t_i) + (t_j) = \mathcal{O}$ for all $i \neq j$), by the Chinese remainder theorem we have

$$\mathcal{O}/(t) \simeq \bigoplus_{i=1}^{n} \mathcal{O}/(t_i^{e_i}) \tag{1}$$

as a direct sum of rings that are also k-vectors spaces (hence k-algebras). To prove the theorem we will compute the dimension of $\mathcal{O}(t)$ in two different ways, corresponding to the two sides of the equality $\deg \phi^*(Q) = \deg \phi$ that we are trying to prove.

On the LHS of the equality we wish to prove, the degree of the divisor $\phi^*(Q)$ is

$$\deg \phi^*(Q) = \sum e_\phi(P_i) \deg P_i = \sum e_i, \tag{2}$$

since we have $\deg P_i = 1$ for $k = \bar{k}$. We claim that this is precisely the dimension of $\mathcal{O}(t) \simeq \bigoplus_i \mathcal{O}/(t_i^{e_i})$ as a k-vector space, which we will prove below.

On the RHS of the equality, after identifying $k(C_2)$ with its image $\phi^*(k(C_2))$ we have

$$\deg \phi = [k(C_1) : k(C_2)],$$

which we claim is equal to the rank of $\mathcal{O}$ as an $\mathcal{O}_Q$-module ($\mathcal{O}_Q$ is embedded in $\mathcal{O}$ via ϕ^*). The ring $\mathcal{O}$ is an integral domain that is finitely generated as a module over the principal ideal domain $\mathcal{O}_Q$, so it is torsion free and isomorphic to $\mathcal{O}_Q^{\oplus r}$ for some integer r (by the structure theorem for modules over PIDs), hence it makes sense to speak of its rank r.

The fields $k(C_1)$ and $\phi^*(k(C_2))$ are the fraction fields of the rings $\mathcal{O}$ and $\mathcal{O}_Q$, respectively, and it follows that the maximal number of elements of $\mathcal{O}$ that are linearly independent over $\mathcal{O}_Q$ is exactly the same as the maximal number of elements of $k(C_1)$ that are linearly independent over $k(C_2)$, which is precisely $[k(C_1) : k(C_2)] = \deg \phi = d$. If we choose a basis $\alpha_1, \ldots, \alpha_d$ for $k(C_1)$ over $k(C_2)$ and let $e = \min\{\operatorname{ord}_{P_i}(\alpha_j) : 0 \le i \le n, 0 \le j \le d\}$, then the functions $\alpha_1/t^e, \ldots, \alpha_d/t^e$ are regular at all the P_i and therefore lie in $\mathcal{O}$. They are linearly independent over $\mathcal{O}_Q$, thus $r \ge d$, and clearly $r \le d$, since any r elements of $\mathcal{O} \subseteq k(C_1)$ that are linearly independent over $\mathcal{O}$ are also linearly independent over its fraction field $k(C_2)$. We have $\mathcal{O}_Q/(t) \simeq k$, since (t) is a maximal ideal, so $\dim_k \mathcal{O}/(t) = r = d = \deg \phi$.

To prove $\dim_k \mathcal{O}(t) = \deg \phi^*(Q)$, by (1) and (2) it suffices to show that $\dim_k \mathcal{O}/(t_i^{e_i}) = e_i$. We claim that for any positive integer n, each function $f \in \mathcal{O}$ can be written uniquely as

$$f \equiv a_0 + a_1 t_i + \cdots + a_{n-1} t_i^{n-1} \bmod t_i^n,$$

with each $a_i \in k$. Applying this with $n = e_i$ will yield the desired result.

For $n = 1$ we let $a_0 = f(P_i) \in k$. We then have $\operatorname{ord}_{P_i}(f - a_0) = \operatorname{ord}_{P_i}(f - f(P_i)) \ge 1$, so $f \equiv a_0 \bmod t_i$ as desired, and clearly a_0 is uniquely determined. We now proceed by induction on n, assuming that $f \equiv g = a_0 + a_1 t_i + \cdots a_{n-1} t_i^{n-1} \bmod t_i^n$. The $\operatorname{ord}_{P_i}(f - g) \ge n$, so $h = t_i^{-n}(f - g)$ is regular at P_i and therefore lies in $\mathcal{O}$ (since $\operatorname{ord}_{P_j}(t_i) = 0$ for $j \ne i$). Now let $a_n = h(P_i) \in k$. Then $\operatorname{ord}_{P_i}(t_i^n(h - a_n)) \ge n + 1$ and we have $f \equiv g + a_n t^n \bmod t^{n+1}$ as desired. □

The key to the proof of Theorem 20.1 is Lemma 20.3, which gave us the independent uniformizers $t_1, \ldots, t_n$ we needed. In order to prove the lemma we need a tight form of the (nonarchimedean) triangle inequality for valuations.

Lemma 20.2 (Triangle equality)**.** *Let $v\colon F^\times \to \Gamma$ be a valuation on a field F. For any $x, y \in F^\times$ such that $v(x) \ne v(y)$ we have $v(x + y) = \min((v(x), v(y))$.*

Proof. Assume $v(x) < v(y)$. By the triangle inequality, $v(x + y) \ge \min(v(x), v(y))$. If this is not tight, $v(x + y) > v(x)$, but then $v(x) = v((x + y) - y) \ge \min(v(x + y), v(y)) > v(x)$, a contradiction. □

We now prove the main lemma we need, which is more generally known as the theorem of *independence of valuations* for function fields.

Lemma 20.3 (Independence of valuations)**.** *Let $P_1, \ldots, P_n$ be distinct places of a function field F. Then there exist $t_1, \ldots, t_n$ so that $v_i(t_j) = \delta_{ij}$ (Kronecker delta), where v_i denotes the valuation for P_i.*

Proof. If $n = 1$, we can take t_1 to be any uniformizer for P_1. We now proceed by induction, assuming that $t_1, \ldots, t_{n-1}$ satisfy $v_i(t_j) = \delta_{ij}$. It suffices to find t_n with $v_n(t_n) = 1$ and $v_i(t_n) = 0$ for $0 \le i < n$. With such a t_n, we can then replace each t_i with t_i/t_n^e, where $e = v_n(t_i)$, so that $v_n(v_i) = 0$ and $v_i(t_j) = \delta_{ij}$ as required.

If $v_n(t_i) = 0$ for $0 \le i < n$, we can simply pick a uniformizer for P_n and multiply it by suitable powers of the t_i so that this is achieved, so let us assume otherwise. We now pick $s_1, \ldots, s_{n-1}$ in $\mathcal{O}_{P_n}$ with $s_i \notin \mathcal{O}_{P_i}$; this is possible because none of the $\mathcal{O}_{P_i}$ contain $\mathcal{O}_{P_n}$, by Theorem 18.5. Then $v_n(s_i) \ge 0$ and $v_i(s_i) < 0$ for $0 \le i < n$. By replacing each s_i with $s_i^{e_i}$ for some suitably large $e_i > 0$ we can arrange it so that at each valuation v_j, for $0 \le j < n$, the value $\min\{v_j(s_i^{e_i}) : 0 \le i < n\}$ is achieved by a unique $s_i^{e_i}$ (possibly the same $s_i^{e_i}$ for different v_j's). For $s = \sum s_i^{e_i}$ we then have $v_j(s) < 0$ for $0 \le j < n$, by the triangle equality, and $v_n(s) \ge 0$.

Now let t be a uniformizer for $\mathcal{O}_{P_n}$, so $v_n(t) = 1$. If $v_n(s) = 0$ then we can replace t by $s^e t$ for some suitable e so that $v_i(t) < 0$ for $0 \le i < n$ and $v_n(t) = 1$, and if $v_n(s) > 0$ we can achieve the same goal by replacing t with $s^e + t$ (again by the triangle equality).

Now let w be the product of t with suitable powers of $t_1, \ldots, t_{n-1}$ so that $v_i(w) = 0$ for $0 \le i < n$. If $v_n(w) = 0$ then apply the same procedure to $t + t^e$ for some suitably chosen $e > 0$ so that this is not the case (we have $v_n(t_i) \ne 0$ for some t_i, so this is always possible). Finally, if $v_n(w) < 0$ then replace w with $1/w$ so $v_n(w) > 0$. We than have $v_i(w) = 0$ for $0 \le i < n$ and $v_n(w) > 0$.

Now let $z = w + 1/t$. We have $v_i(1/t) > 0$ for $0 \le i < n$ and $v_n(1/t) = -1$, so by the triangle equality, $v_i(z) = 0$ for $0 \le i < n$ and $v_n(z) = -1$. For $t_n = 1/z$ we then have $v_i(t_n) = 0$ for $0 \le i < n$ and $v_n(t_n) = 1$ as desired, and we are done. □

Corollary 20.4. *Let $\mathcal{O}_1, \ldots, \mathcal{O}_n$ be distinct discrete valuation rings of a function field F/k. The ring $\mathcal{O} = \cap_i \mathcal{O}_i$ has exactly n nonzero prime ideals $(t_1), \ldots, (t_n)$, each principal and generated by a uniformizer for $\mathcal{O}_i$. Every nonzero $f \in \mathcal{O}$ can be uniquely factored as $f = ut_1^{e_1} \cdots t_n^{e_n}$ with $u \in \mathcal{O}^\times$ and $e_i = \operatorname{ord}_{P_i}(f) \ge 0$.*

Proof. The elements $t_1, \ldots, t_n$ given by Lemma 20.3 are uniformizers for $\mathcal{O}_1, \ldots, \mathcal{O}_n$, and it follows that every $f \in F^\times$ can then be written uniquely in the form $x = ut_1^{e_1} \cdots t_n^{e_n}$ with $u \in \mathcal{O}^\times$ and $e_i = \operatorname{ord}_{P_i}(x)$. The nonzero elements of $\mathcal{O}$ are precisely those for which the e_i are all nonnegative, and the lemma is then clear. □

We now note a further corollary of the lemma, which is an analog of the weak approximation theorem we proved in Lecture 11.

Corollary 20.5 (Weak approximation for function fields)**.** *Let $P_1, \ldots, P_n$ be distinct places of a function field F/k, and let $f_1, \ldots, f_n \in F$ be given. For every positive integer N there exists $f \in F$ such that $\operatorname{ord}_{P_i}(f - f_i) > N$ for $0 \le i < n$.*

Proof. Let $t_1, \ldots, t_n$ be as in Lemma 20.3. As in the proof of Theorem 20.1, we can construct Laurent polynomials $g_i \in k((t_i))$ such that $g_i \equiv f_i \bmod t_i^N$, where the first nonzero term of

g_i is $a_s t_i^s$ where $a_s = \operatorname{ord}_{P_i}(f)$. We then have $\operatorname{ord}_{P_i}(g_i - f_i) \geq N$, and $\operatorname{ord}_{P_j}(g_i) \geq 0$ for $j \neq i$ since $\operatorname{ord}_{P_j}(t_i) = 0$ for $j \neq i$, this follows from the triangle inequality. Multiplying each g_i by $(t_1 \cdots t_{i-1} t_{i+1} \cdots t_n)^N$ and summing the results yields the desired function f. □

Note that in terms of absolute values, making the valuation $\operatorname{ord}_{P_i}(f - f_i)$ large corresponds to making the corresponding absolute value $|f - f_i|_{P_i}$ small. To make the analogy with Theorem 11.7 more precise, we could construct the completions of F_{P_i} at each place P_i and then the f_i given in the theorem would lie in F_{P_i} but f would still lie in F. The relationship between F and its completions F_{P_i} is then exactly analogous to the relationship between $\mathbb{Q}$ and its completions $\mathbb{Q}_{p_i}$.

20.2 Divisors of degree zero

It follows from Theorem 20.1 that the group of principal divisors $\operatorname{Princ}_k C$ is a subgroup of the group of degree zero divisors $\operatorname{Div}_k^0 C$, the quotient $\operatorname{Div}_k^0 C / \operatorname{Princ}_k C$ is denoted $\operatorname{Pic}_k^0 C$. Equivalently, $\operatorname{Pic}_k^0 C$ is the kernel of the degree map $\operatorname{Pic} C \to \mathbb{Z}$. We then have the exact sequence

$$1 \to k^\times \to k(C)^\times \to \operatorname{Div}_k^0 C \to \operatorname{Pic}_k^0 C \to 0.$$

Up to now all the groups of divisors and divisor classes we have considered have been infinite, but this is not true of Pic_k^0. The case where Pic_k^0 is trivial is already an interesting result.

Theorem 20.6. *Assume $k = \bar{k}$. Then $C \simeq \mathbb{P}^1$ if and only if $\operatorname{Pic}_k^0 C = \{0\}$.*

Proof. The forward implication is easy. Each point $P = (a_0, a_1) \in \mathbb{P}^1$ is the zero locus of the polynomial $f_P(x_0, x_1) = a_1 x_0 - a_0 x_1$, and if we have a divisor $D = \sum n_P P$ we can construct a corresponding homogeneous rational function $f = \prod f_P^{n_P}$. If D has degree zero then the numerator and denominator of f have the same degree and f is an element of $k(\mathbb{P}^1) \simeq k(C)$, so $D = \operatorname{div} f$. Thus $\operatorname{Div}_k C = \operatorname{Princ}_k C$ and $\operatorname{Pic}_k^0 C = 0$.

Now let P and Q be distinct points in $C(k)$; such P and Q exist because k is algebraically closed. Then $f = f_P/f_Q$ is a non-constant function in $C(k)$ that defines a morphism $(f_P : f_Q)$ from C to $\mathbb{P}^1$. The polynomials f_P and f_Q have degree one, and this implies that the morphism f has degree one and is an isomorphism. To check this, we can use Theorem 20.1 with $Q = 0$ and $t_0 = x/y$ to compute

$$\deg f = \deg f^*(0) = e_f(P) = \operatorname{ord}_P(f^* t_0) = \operatorname{ord}_P(t_0 \circ f) = \operatorname{ord}_P(f_P/f_Q) = 1. \qquad \square$$

Now let us consider the general case, where k is not necessarily algebraically closed. We then need to work with closed points, but the forward implication still holds: if C/k is isomorphic to $\mathbb{P}^1/k$ then $\operatorname{Pic}_k^0 C$ is trivial; the polynomials f_P in the proof are now irreducible polynomials that may have degree greater than one, but that doesn't change the argument.

But the converse is more interesting. We can always find closed points P and Q on C/k, but for the above proof to work we need them to have degree one, otherwise the function f_P/f_Q will not be an isomorphism. Equivalently, we need C/k to have two distinct rational points P and Q; these are closed points of degree one. We already know from earlier in the course that if C/k has genus 0 and even one rational point then it is isomorphic to $\mathbb{P}^1/k$ (and then it has more than two rational points). But if C/k has positive genus it can happen that C/k has one rational point and $\operatorname{Pic}_k^0 C = \{0\}$, but C cannot be isomorphic to $\mathbb{P}^1$, because $\mathbb{P}^1$ has genus zero. Indeed, this is exactly what happens for the elliptic curve

$y^2 = x^3 + 7$ over $\mathbb{Q}$, whose only rational point is ∞. So we need to add the hypothesis that C/k have two distinct rational points in order to get a theorem that works for general k.

Corollary 20.7. *Let C/k be a curve with at least two distinct rational points. Then C/k is isomorphic to $\mathbb{P}^1/k$ (with the isomorphism defined over k) if and only if* $\operatorname{Pic}_k^0 C = \{0\}$.

As an interesting consequence, if C has genus greater than zero and at least two rational points, then $\operatorname{Pic}_k^0 C$ cannot be trivial. The elliptic curve $C\colon y^2 = x^3 - 1$ over $k = \mathbb{Q}$ is such an example, with $\operatorname{Pic}_k^0 C$ of order 2.

References

[1] I. R. Shafarevich, *Basic algebraic geometry*, 2nd edition, Springer-Verlag, 1994.

[2] H. Stichtenoth, *Algebraic function fields and codes*, Springer, 2009.

18.782 Introduction to Arithmetic Geometry **Fall 2013**
Lecture #21 **11/19/2013**

As usual, k is a perfect field, but not necessarily algebraically closed. Throughout this lecture C/k denotes a curve (smooth projective variety of dimension one) and F/k the corresponding function field. To simplify the notation, for any place P of F/k and divisor $D = \sum n_P P$, we define $\operatorname{ord}_P(D) = n_P$.

21.1 Riemann-Roch spaces

We have seen that the degree of a divisor is a key numerical invariant that is preserved under linear equivalence; recall that two divisors are linearly equivalent if their difference is a principal divisor, equivalently, they correspond to the same element of the Picard group. We now want to introduce a second numerical invariant associated to each divisor. In order to do this we first partially order divisors by defining the relation $\le$ on $\operatorname{Div}_k C$ by

$$A \le B \quad \Longleftrightarrow \quad \operatorname{ord}_P(A) \le \operatorname{ord}_P(B) \text{ for all } P.$$

As usual, P ranges over all closed points of C/k, equivalently, all places of $k(C)$, but of course the inequality on the right is automatically satisfied for all but finitely many P. This partial ordering is compatible with divisor addition, since

$$A \le B \quad \Longrightarrow \quad A + C \le B + C,$$

for any divisor C. We also note that

$$A \le B \text{ and } C \le D \quad \Longrightarrow \quad A + C \le B + D.$$

It is important to remember that $\le$ is not a total ordering on $\operatorname{Div}_k C$; most pairs of divisors are incomparable.

Definition 21.1. A divisor $D \ge 0$ is said to be *effective*. As with principal divisors $\operatorname{div} f = \operatorname{div}_0 f - \operatorname{div}_\infty f$, every divisor can be written uniquely as the difference of two effective divisors, as $D = D_0 - D_\infty$, where

$$D_0 := \sum_{\operatorname{ord}_P(D) > 0} \operatorname{ord}_P(D)P \qquad \text{and} \qquad D_\infty := \sum_{\operatorname{ord}_P(D) < 0} -\operatorname{ord}_P(D)P.$$

We now define the Riemann-Roch space of a divisor.

Definition 21.2. The *Riemann-Roch* space of a divisor D is the k-vector space

$$\mathcal{L}(D) := \{f \in k(C)^\times : \operatorname{div} f \ge -D\} \cup \{0\}.$$

That $\mathcal{L}(D)$ is a vector space follows immediately from:

1. $\operatorname{div} \lambda f = \operatorname{div} f + \operatorname{div} \lambda = \operatorname{div} f$ for all $\lambda \in k^\times$;
2. $\operatorname{ord}_P(f + g) \ge \min(\operatorname{ord}_P(f), \operatorname{ord}_P(g)$ for all $f, g \in F^\times$.

Example 21.3. If $D = 3P - 2Q$ then $\mathcal{L}(D)$ is the set of functions in $k(C)$ that have at most a triple pole at P, and at least a double zero at Q, and poles nowhere else (but they may have have zeros of any order at points other than Q).

Example 21.4. If $D = -P$ then $\mathcal{L}(D)$ is the set of functions that have a zero at P and no poles at all. The only such function is the zero function (which lies in $\mathcal{L}(D)$ by definition). More generally, for any $D < 0$ we have $\mathcal{L}(D) = \{0\}$.

Example 21.5. If $D = 0$ then $\mathcal{L}(D)$ is the set of functions that have no poles at all. By Corollary 19.23, for $f \in \mathcal{L}(0)$ we have $\deg \operatorname{div}_\infty f = 0$ if and only if $f \in k^\times$, so $\mathcal{L}(0) = k$.

We now show that $\mathcal{L}(D)$ is preserved (up to isomorphism) by linear equivalence.

Lemma 21.6. *For any linearly equivalent divisors $A \sim B$ we have $\mathcal{L}(A) \simeq \mathcal{L}(B)$.*

Proof. We have $A - B = \operatorname{div} f$ for some $f \in k(C)^\times$, and we claim that the maps $g \mapsto fg$ and $g \mapsto g/f$ are inverse k-linear maps from $\mathcal{L}(A)$ to $\mathcal{L}(B)$ and from $\mathcal{L}(B)$ to $\mathcal{L}(A)$, respectively. Linearity is clear, and if $\operatorname{div} g \geq -A$ then

$$\operatorname{div} fg = \operatorname{div} f + \operatorname{div} g \geq \operatorname{div} f - A = -B.$$

Similarly, if $\operatorname{div} g \geq -B$ then

$$\operatorname{div} g/f = \operatorname{div} g - \operatorname{div} f \geq -B - \operatorname{div} f = -A.$$

Thus we have defined linear transformations from $\mathcal{L}(A)$ to $\mathcal{L}(B)$, hence $\mathcal{L}(A) \simeq \mathcal{L}(B)$. □

The following lemma shows that non-trivial Riemman-Roch spaces arise only (and always) for divisors that are linearly equivalent to an effective divisor.

Lemma 21.7. *We have $\mathcal{L}(D) \neq \{0\}$ if and only if $D \sim D'$ for some $D' \geq 0$.*

Proof. If $f \in \mathcal{L}(D)$ is nonzero, then $\operatorname{div} f \geq -D$, and $D \sim D' = D + \operatorname{div} f \geq 0$. Conversely, if $D \sim D' \geq 0$ then $-D \leq D' - D = \operatorname{div} f$ for some $f \in k(C)^\times$, hence $\mathcal{L}(D) \neq \{0\}$. □

Lemma 21.8. *For any two divisors $A \leq B$ we have $\mathcal{L}(A) \subseteq \mathcal{L}(B)$ and*

$$\dim(\mathcal{L}(B)/\mathcal{L}(A)) \leq \deg B - \deg A.$$

Proof. It is clear that $\mathcal{L}(A) \subseteq \mathcal{L}(B)$, and that the inequality holds if $A = B$. We now prove the inequality in the case $B = A + P$, for some place P. Let t be a uniformizer at P, let $k(P) = \mathcal{O}_P/P$ be the residue field of P, and let $n = \operatorname{ord}_P(B)$. Now define the linear transformation $\phi\colon \mathcal{L}(B) \to k(P)$ by $\phi(f) = (t^n f)(P) = t^n f \bmod P$; we have

$$\operatorname{ord}_P(t^n f) = n + \operatorname{ord}_P(f) \geq 0$$

for $f \in \mathcal{L}(B)$, so $t^n f \in \mathcal{O}_P$ and ϕ is well-defined. The image of ϕ lies in $k(P) = k^{\deg P}$, and its kernel consists of subspace of functions $f \in \mathcal{L}(B)$ for which $\operatorname{ord}_P(t^n f) \geq 1$, equivalently, $\operatorname{ord}_P(f) \geq 1 - n = -\operatorname{ord}_P(A)$, which is precisely $\mathcal{L}(A)$. We have $\mathcal{L}(B)/\ker \phi \simeq \operatorname{im} \phi$, so

$$\dim(\mathcal{L}(B)/\mathcal{L}(A)) = \dim \operatorname{im} \phi \leq \dim k(P) = \deg P = \deg B - \deg A. \tag{1}$$

The general case follows from repeated application of the same result. If

$$A = B_0 < B_1 < B_2 < \cdots < B_m = B,$$

where $B = \sum n_P P$ and $m = \sum n_P$, then each difference $B_{i+1} - B_i$ is a single place P_i. Applying (1) gives $\dim(\mathcal{L}(B_{i+1})/\mathcal{L}(B_i)) = \deg B_{i+1} - \deg B_i = \deg P_i$. Summing yields the desired result $\dim(\mathcal{L}(B)/\mathcal{L}(A)) \leq \deg B - \deg A$. □

We now prove that the dimension of a Riemann-Roch space is finite.

Theorem 21.9. *For any divisor D we have $\dim \mathcal{L}(D) \le \deg D_0 + 1$.*

Proof. Applying Lemma 21.8 with $B = D$ and $A = 0$ yields

$$\dim(\mathcal{L}(D_0)/\mathcal{L}(0)) \le \deg D_0 - \deg 0 = \deg D_0.$$

As noted in Example 21.5, we have $\mathcal{L}(0) = k$, and therefore

$$\dim \mathcal{L}(D_0) = \dim(\mathcal{L}(D_0)/\mathcal{L}(0)) + 1 \le \deg D_0 + 1.$$

We also have $D \le D_0$, so by Lemma 21.8, $\mathcal{L}(D) \subseteq L(D_0)$, and we have

$$\dim \mathcal{L}(D) \le \dim \mathcal{L}(D_0) \le \deg D_0 + 1$$

as claimed. □

Definition 21.10. The *dimension* $\ell(D)$ of a divisor is the dimension of $\mathcal{L}(D)$.

The following corollary summarizes what we know about $\ell(D)$ so far.

Corollary 21.11. *The following hold:*

(a) *$\ell(0) = 1$.*

(d) *If $A \sim B$ then $\ell(A) = \ell(B)$ and $\deg(A) = \deg(B)$.*

(c) *For any $A \le B$ we have $\ell(B) - \ell(A) \le \deg B - \deg A$.*

(d) *For all $D \ge 0$ we have $\ell(D) \le \deg D + 1$.*

(e) *If $\deg D < 0$ then $\ell(D) = 0$.*

Proof. (a) follows from Example 21.5, (b) is Lemma 21.6 and Corollary 19.23, (c) is Lemma 21.6, (d) is Theorem 21.9, and (e) follows from Lemma 21.7. □

An equivalent form of (c) that we will often use is

$$A \le B \qquad \Longrightarrow \qquad \deg A - \ell(A) \le \deg B - \ell(B).$$

Lemma 21.12. *If $\deg D = 0$ then $\ell(D) = 1$ if D is principal and $\ell(D) = 0$ otherwise.*

Proof. If $D = \operatorname{div} f$ is principal, then $f \in \mathcal{L}(D)$, so $\ell(D) \ge 1$ and by Lemma 21.7 we must have $D \sim D' \ge 0$. But $\deg D' = \deg D = 0$, so $D' = 0$ and $\ell(D) = \ell(0) = 1$. Now suppose $\ell(D) \ge 1$. As just argued, we must have $\ell(D) = 1$, so there is a nonzero $f \in \mathcal{L}(D)$, and since $\operatorname{div} f \ge -D$, we have $D + \operatorname{div} f \ge 0$. But $\deg(D + \operatorname{div} f) = 0$, so $D + \operatorname{div} f = 0$ and therefore $D = -\operatorname{div} f = \operatorname{div} 1/f$ is principal. Taking the contrapositive, if D is not principal then we must have $\ell(D) = 0$. □

It follows from Lemma 21.12 that the inequality in Theorem 21.9 is not tight for curves for which $\operatorname{Pic}_k^0 C$ is not trivial, since this implies the existence of non-principal divisors of degree 0. On the other hand, for $C = \mathbb{P}^1$, the inequality is tight for all effective divisors (as noted at the end of Lecture 20, there is a gap between these two cases, one can have $\operatorname{Pic}_k^0 C = \{0\}$ and $C \not\simeq \mathbb{P}^1$; we will address this gap in the next lecture).

Lemma 21.13. *If C is isomorphic to $\mathbb{P}^1$ then $\ell(D) = \deg D + 1$ for all $D \ge 0$.*

Proof. If $C \simeq \mathbb{P}^1$ then $k(C)$ is the field of all rational functions over k. We claim that given any effective divisor $A = \sum n_P P$ we can construct a function $f_A \in k(C)$ with $\operatorname{div}_\infty f = A$.

Proof of claim: We just need to show that we can construct $\operatorname{div}_0 f$ with support disjoint from $\operatorname{div}_\infty f$. If k is infinite this is easy: pick a degree one place $P \notin \operatorname{Supp}(A)$ and let $\operatorname{div}_0 f = (\deg A)P$. If k is finite, then, as noted in Lecture 3, there exist monic irreducible polynomials of every degree in $k[t]$, and each corresponds to a place of $k(C)$. If A consists of more than a single place, no place of degree $\deg A$ can lie in the support of A, so pick one such place P and let $\operatorname{div}_0 f = P$. Otherwise A consists of a single place and we can pick a degree one place P not in the support of A and let $\operatorname{div}_0 f = (\deg A)P$ as above.

Now let $0 = A_0 < A_1 < \cdots < A_m = D$ be a maximal chain of divisors, let $P_i = A_i - A_{i-1}$ for $1 \le i \le m$, and let $f_1, \ldots, f_m \in k(C)$ satisfy $\operatorname{div}_\infty f_i = A_i$ (note that the list $P_1, \ldots, P_m$ may contain repetitions). These functions are linearly independent over k, since for any nonempty subset the f_i with maximal index i has a pole at P_i of order greater than that of any f_j with $j < i$, and the triangle equality then precludes any non-trivial relations. Finally, for each point P_i there is a subspace $V_i \subseteq \mathcal{L}(D)$ corresponding to functions f for which $\operatorname{div} f = \operatorname{div} f_i$, and $\mathcal{L}(D)$ contains the direct sum of these subspaces, since no pair intersects non-trivially. If we consider the linear transformation $\phi\colon V_i \to k(P_i)$ defined by $f \mapsto (t_i^{n_i} f)(P_i)$, where t_i is a uniformizer for P_i and $n_i = -\operatorname{ord}_{P_i}(f)$, it is clear that $\ker \phi$ is trivial, and ϕ is surjective becuase $k(C)$ is the rational function field. So $\dim V_i = \deg P_i$.

We then have

$$\ell(D) = \dim \mathcal{L}(D) \ge \dim \mathcal{L}(0) + \sum \dim V_i = 1 + \sum \deg P_i = 1 + \deg D,$$

as claimed. □

Remark 21.14. If you think the proof of Lemma 21.13 is a lot of effort to prove something that should be obvious, your are right. Once we prove the Riemann-Roch theorem it will follow trivially (as will many other results). Our purpose in proving it now is to help motivate the definition of genus.

We know that the inequality in Theorem 21.9 is tight when C is *rational* (isomorphic to $\mathbb{P}^1$), but not in general. As we will show, for suitable divisors D (which will turn out to be almost all of them), the quantity $\deg D + 1 - \ell(D)$ tells us something intrinsic to the function field $k(C)$; roughly speaking, it measure how far C is from being rational.[1] One way to think about this metric is as a measure of the functions that are "missing" from $k(C)$.

We now show that the $\deg D + 1 - \ell(D)$ is bounded, independent of D.

Theorem 21.15. *There is a non-negative integer g such that*

$$\deg(D) + 1 - \ell(D) \le g$$

holds for all $D \in \operatorname{Div}_k C$.

The proof below is adapted from [1, Prop. 1.4.14].

Proof. Let $f \in k(C)$ be transcendental over k, and let $A = \operatorname{div}_\infty f \ge 0$. Let $v_1, \ldots, v_d$ be a basis for $k(C)/k(f)$, where $d = \deg A = [k(C) : k(f)]$ (by Corollary 19.24). Choose a divisor $B \ge 0$ so that $\operatorname{div} v_i \ge -B$ for each v_i (this is clearly possible).

[1] Modulo annoying special cases like genus 0 curves that are not rational (and genus 1 curves that are not elliptic curves). Such annoyances can be eliminated by insisting on at least one rational point.

For any integer $n \geq 0$, the set of functions $S = \{v_i f^j : 1 \leq i \leq d, 0 \leq j \leq n\}$ is clearly linearly independent over k, since $v_1, \ldots, v_d$ are linearly independent over $k(f)$ and f is transcendental over k. And $S \subseteq \mathcal{L}(nA + B)$, since $\operatorname{div}(v_i f^j) \geq -nA - B$ for all $v_i f^j \in S$. Therefore

$$\ell(nA + B) \geq d(n+1) = (n+1)\deg A \tag{2}$$

for all $n \geq 0$. But we also have $nA \leq nA + B$, since $B \geq 0$, and Corollary 21.11.c implies

$$\ell(nA + B) - \ell(nA) \leq \deg(nA + B) - \deg(nA) = \deg B. \tag{3}$$

Combining (2) and (3) yields

$$\ell(nA) \geq \ell(nA + B) - \deg B \geq (n+1)\deg A - \deg B = \deg(nA) + (\deg A - \deg B).$$

It follows that

$$\deg(nA) + 1 - \ell(nA) \leq \deg A - \deg B + 1$$

for all $n \geq 0$. Let $g = \deg A - \deg B + 1$ so that

$$\deg(nA) + 1 - \ell(nA) \leq g, \tag{4}$$

where we note that $g \geq 0$, by Corollary 21.11.d.

Now let D be any divisor in $\operatorname{Div}_k C$ and write $D = D_0 - D_\infty$ as the difference of two effective divisors. We claim that D_0 is equivalent to an effective divisor $D' \leq nA$, for some n. By Corollary 21.11.c, we have

$$\ell(nA) - \ell(nA - D_0) \leq \deg(nA) - \deg(nA - D_0) = \deg D_0,$$

and applying (4) yields

$$\ell(nA - D_0) \;\geq\; \ell(nA) - \deg D_0 \;\geq\; \deg(nA) + 1 - g + \deg D_0.$$

The RHS is clearly positive for sufficiently large n, so pick n so that $\ell(nA - D_0) > 0$ and let $f \in \mathcal{L}(nA - D_0)$ be nonzero. Now define $D' := D_0 - \operatorname{div} f$ so that

$$D' = D_0 - \operatorname{div} f \leq D_0 - (D_0 - nA) = nA,$$

as claimed. We have $D \leq D_0$, so $\ell(D_0) - \ell(D) \leq \deg D_0 - \deg D$, by Corollary 21.11.c, and

$$\begin{aligned}\deg D + 1 - \ell(D) &\leq \deg D_0 + 1 - \ell(D_0)\\ &= \deg D' + 1 - \ell(D')\\ &\leq \deg(nA) + 1 - \ell(nA)\\ &\leq g,\end{aligned}$$

where we used $D' \sim D_0$ equality, $D' \leq nA$ to get the second inequality, and then (4). □

References

[1] H. Stichtenoth, *Algebraic function fields and codes*, Springer, 2009.

18.782 Introduction to Arithmetic Geometry **Fall 2013**
Lecture #22 **11/21/2013**

Throughout this lecture C/k is a curve over a perfect but not necessarily algebraically closed field k and F/k denotes the corresponding function field.[1]

In the last lecture we defined $\ell(D)$ as the dimension of the Riemann-Roch space $\mathcal{L}(D)$ of the divisor D, and we proved that $\ell(D)$ is invariant under linear equivalence. It is immediate from the definitions that for any divisors $A \le B$ we have $\ell(A) \le \ell(B)$ and $\deg(A) \le \deg(B)$. Less obvious is the fact that

$$\deg(A) - \ell(A) \le \deg(B) - \ell(B),$$

but we proved that this holds for all $A \le B$; see Lemma 21.8. As we are particularly interested in the quantities on the two sides of the above inequality, let us define

$$r(D) := \deg(D) - \ell(D).$$

Then $r(D)$ is preserved under linear equivalence and $A \le B \implies r(A) \le r(B)$. At the end of Lecture 21 we proved that for every curve C/k there is an integer $g \ge 0$ such that

$$r(D) \le g - 1 \tag{1}$$

for all $D \in \operatorname{Div}_k C$. We also showed that for $C \simeq \mathbb{P}^1$ we can take $g = 0$, and that $r(D) = -1$ for all $D \ge 0$. We always have $r(0) = 0 - 1 = -1$, so $r(D) \ge -1$ for all $D \ge 0$.

22.1 The genus of a curve

We now define the genus of a curve.

Definition 22.1. The *genus* of the curve C/k is defined by

$$g := \max\{r(D) + 1 : D \in \operatorname{Div}_k(C)\}.$$

In other words, g is the least integer for which (1) holds.

Remark 22.2. This definition of the genus of a curve is sometimes called the *geometric genus* to distinguish it from other notions of genus that we won't consider in this course. For (smooth projective) curves the different definitions all agree.

We now give the complete statement of Riemann's Theorem, most of which was proved in Theorem 21.15.

Theorem 22.3 (Riemann's Theorem)**.** *Let C/k be a curve of genus g. Then $r(D) \le g - 1$ for all $D \in \operatorname{Div}_k C$, and equality holds for all divisors of sufficiently large degree.*

Proof. We have already proved the inequality. Let us pick a divisor A for which $r(A) = g-1$; some such A exists, by the definition of g. We will show that $r(D) = g - 1$ whenever $\deg D \ge \deg A + g = c$.

[1] Recall that our curves are smooth projective varieties of dimension one, and that our varieties are geometrically irreducible.

So assume $\deg D \ge c$. We have $r(D-A) = \deg(D-A) - \ell(D-A) \le g-1$, so

$$\ell(D-A) \;\ge\; \deg(D-A) + 1 - g \;\ge\; c - \deg A + 1 - g \;=\; 1.$$

There is a nonzero $f \in \mathcal{L}(D-A)$, so let $D' = D + \operatorname{div} f \ge D + A - D = A$. Then

$$r(D) = r(D') \ge r(A) = g-1,$$

and we already know that $r(D) \le g-1$, so $r(D) = g-1$. □

We now want to refine Riemann's Theorem to obtain a more precise statement that will tell us exactly what "sufficiently large" means and give us a measure of how far the inequality $r(D) \le g-1$ is from being an equality for any particular divisor D; this is the Riemann-Roch theorem.

Definition 22.4. Let C/k be a curve of genus g. For $D \in \operatorname{Div}_k C$, the non-negative integer

$$i(D) := g - 1 - r(D)$$

is the *index of speciality* of D. Divisors for which $i(D) > 0$ are said to be *special.*

We know from Riemann's Theorem that $i(D) = 0$ for all D of sufficiently large degree, and we also know that $i(0) = g$, since

$$i(0) = g - 1 - r(0) = g - 1 - \deg 0 + \ell(0) = g - 1 - 0 + 1 = g.$$

22.2 The ring of adeles

To compute the index of speciality we introduce the adele ring. Our presentation roughly follows that in [2, §1.5].

Definition 22.5. The *adele ring* of the function field F/k is the subring $\mathcal{A} = \mathcal{A}_F$ of the direct product $\prod_P F$ consisting of those elements $\alpha = (\alpha_P)$ for which $\alpha_P \in \mathcal{O}_P$ for all but finitely many P. The elements of $\mathcal{A}$ are called *adeles.*

Remark 22.6. The adele ring $\mathcal{A}$ is also called the *ring of repartitions.* It is often defined in terms of the P-adic completions of F, but we don't need to take completions to prove the Riemann-Roch theorem so we won't (some authors refer to $\mathcal{A}$ as the ring of *pre-adeles*).

The function field F/k is canonically embedded in $\mathcal{A}$ via the diagonal embedding

$$f \mapsto (f, f, f, \ldots).$$

Adeles of this form are called *principal adeles*, terminology that is consistent with our notion of a principal divisor; those divisors that correspond to elements of the function field. Like F, the adele ring $\mathcal{A}$ is a k-vector space. We extend each valuation ord_P of F/k to a valuation on $\mathcal{A}$ by defining $\operatorname{ord}_P(\alpha) = \operatorname{ord}_P(\alpha_P)$ for $\alpha_P \ne 0$ and setting $\operatorname{ord}_P(0) = \infty$, where ∞ is greater than any element of $\mathbb{Z}$.

Definition 22.7. For a divisor D the *adele space* of D is the k-vector space

$$\mathcal{A}(D) := \{\alpha \in \mathcal{A} : \operatorname{ord}_P(\alpha) \ge -\operatorname{ord}_P(D) \text{ for all places } P\}.$$

It contains the Riemann-Roch space $\mathcal{L}(D) = \mathcal{A}(D) \cap F$ as a subspace, and it is in turn a subspace of the adele ring $\mathcal{A}$.

The adele space of a divisor gives us additional information beyond what we get from the Riemann-Roch space that will allow us to characterize the index of speciality in a canonical way. We first prove three lemmas.

Lemma 22.8. *For any two divisors $A \le B$ we have $\mathcal{A}(A) \subseteq \mathcal{A}(B)$ and*

$$\dim \mathcal{A}(B)/\mathcal{A}(A) = \deg B - \deg(A),$$

as k-vector spaces.

Proof. The inclusion $\mathcal{A}(A) \subseteq \mathcal{A}(B)$ is clear. As in the proof of Lemma 21.8, it suffices to consider the case $B = A + P$ for some place P, and the proof is exactly the same. We pick a uniformizer t for P and define the linear map $\phi\colon \mathcal{A}(B) \to k(P)$ by $\phi(f) = (t^n f)(P)$, where $n = \operatorname{ord}_P(B)$. The map ϕ is surjective and its kernel is $\mathcal{A}(A)$, hence

$$\dim(\mathcal{A}(B)/\mathcal{A}(A)) = \dim k(P) = \deg P = \deg B - \deg A. \qquad \square$$

Lemma 22.9. *For any two divisors $A \le B$ we have $\mathcal{A}(A) + F \subseteq \mathcal{A}(B) + F$ and*

$$\dim \frac{\mathcal{A}(B) + F}{\mathcal{A}(A) + F} = r(B) - r(A),$$

as k-vector spaces, where F is embedded diagonally in $\mathcal{A}$.

Proof. The inclusion is clear, and the map

$$\mathcal{A}(B) \to \mathcal{A}(B) + F \to (\mathcal{A}(B) + F)/(\mathcal{A}(A) + F)$$

is surjective, with kernel $\mathcal{A}(B) \cap (\mathcal{A}(A) + F)$. We therefore have

$$\frac{\mathcal{A}(B) + F}{\mathcal{A}(A) + F} \simeq \frac{\mathcal{A}(B)}{\mathcal{A}(B) \cap (\mathcal{A}(A) + F)} = \frac{\mathcal{A}(B)}{\mathcal{A}(A) + \mathcal{L}(B)} \simeq \frac{\mathcal{A}(B)/\mathcal{A}(A)}{(\mathcal{A}(A) + \mathcal{L}(B))/\mathcal{A}(A)}.$$

Applying Lemma 22.8 and taking dimensions gives

$$\dim \frac{\mathcal{A}(B) + F}{\mathcal{A}(A) + F} = \deg B - \deg A - \dim \frac{\mathcal{A}(A) + \mathcal{L}(B)}{\mathcal{A}(A)}.$$

Finally we note that

$$\dim \frac{\mathcal{A}(A) + \mathcal{L}(B)}{\mathcal{A}(A)} = \dim \frac{\mathcal{L}(B)}{\mathcal{A}(A) \cap \mathcal{L}(B)} = \dim \frac{\mathcal{L}(B)}{\mathcal{L}(A)} = \ell(B) - \ell(A),$$

thus

$$\dim \frac{\mathcal{A}(B) + F}{\mathcal{A}(A) + F} = \deg B - \deg A - (\ell(B) - \ell(A)) = r(B) - r(A). \qquad \square$$

Lemma 22.10. *For any divisor D for which $r(D) = g - 1$ we have*

$$\mathcal{A} = \mathcal{A}(D) + F.$$

Proof. Let $\alpha \in \mathcal{A}$. We will show $\alpha \in \mathcal{A}(D) + F$. Let us pick a divisor $D' \ge D$ such that $\alpha \in \mathcal{A}(D') + F$; this is clearly possible. We have $g - 1 = r(D) \le r(D') \le g - 1$, by Riemann's Theorem, so $r(D') = g - 1$. By Lemma 22.9 we have

$$\dim \frac{\mathcal{A}(D') + F}{\mathcal{A}(D) + F} = r(D') - r(D) = (g-1) - (g-1) = 0,$$

so $\mathcal{A}(D') + F = \mathcal{A}(D) + F$ and therefore $\alpha \in \mathcal{A}(D) + F$. $\square$

We can now determine the index of speciality of a divisor in terms of its adele space.

Theorem 22.11. *Let F/k be a function field. For any divisor $D \in \operatorname{Div}_k F$ we have*

$$i(D) = \dim \mathcal{A}/(\mathcal{A}(D) + F).$$

Proof. By Riemann's Theorem there exists a divisor $D' \geq D$ for which $r(D') = g - 1$; we just need to make the degree of $D' \geq D$ large enough, and this is clearly possible; if $D \neq 0$ we can take a multiple of $D_0 + D_\infty$. By Lemma 22.10 we then have $\mathcal{A} = \mathcal{A}(D') + F$, thus

$$\dim \frac{\mathcal{A}}{\mathcal{A}(D) + F} = \dim \frac{\mathcal{A}(D') + F}{\mathcal{A}(D) + F} = r(D') - r(D) = g - 1 - r(D) = i(D),$$

where we use Lemma 22.9 to get the second equality. □

We now have an equality that holds for all divisors

$$g = r(D) + 1 + i(D),$$

and we know that $i(D) = 0$ for all divisors of sufficiently large degree. But we would like to characterize $i(D)$ in a canonical way that does not involve the adele ring; this will yield the Riemann-Roch theorem.

22.3 Canonical divisors

Definition 22.12. Let F/k be a function field and let $\mathcal{A}$ be its adele ring. For a divisor $D \in \operatorname{Div}_k F$ the space of *Weil differentials* $\Omega(D)$ is the orthogonal complement of $\mathcal{A}(D) + F$ (its annihilator in the dual space $\mathcal{A}^\vee$). Explicitly, this is the set of all linear functionals $\omega\colon \mathcal{A} \to k$ that vanish on $\mathcal{A}(D) + F$. The k-vector space

$$\Omega = \Omega_F := \bigcup_{D \in \operatorname{Div}_k F} \Omega(D)$$

is the space of *Weil differentials* for F/k.

It is clear that Ω is a k-vector space: if $\omega_1 \in \Omega(D_1)$ and $\omega_2 \in \Omega(D_2)$ then $\omega_1 + \omega_2$ lies in $\Omega(D)$, where $D = D_1 \wedge D_2$ is defined by $\operatorname{ord}_P(D) = \min(\operatorname{ord}_P(D_1), \operatorname{ord}_P(D_2))$.

Lemma 22.13. *For any divisor $D \in \operatorname{Div}_k F$ we have* $\dim \Omega(D) = i(D)$.

Proof. The quotient space $\mathcal{A}/(\mathcal{A}(D) + F)$ has finite dimension $i(A)$, by Theorem 22.11, thus it has the same dimension as its dual, which is canonically isomorphic to the orthogonal complement of $\mathcal{A}(D) + F$, which is precisely $\Omega_F(D)$.[2] Thus $\dim \Omega(D) = i(D)$. □

We have seen that the space of Weil differentials Ω is a k-vector space; we now make Ω an F-vector space by defining $f\omega \in \Omega$ for $f \in F$ and $\omega \in \Omega$ as the linear functional $\mathcal{A} \to k$ that sends α to $\omega(f\alpha)$, in other words

$$(f\omega)(\alpha) = \omega(f\alpha),$$

for all $\alpha \in \mathcal{A}$.

[2] If V/W is any quotient, the map $\Phi\colon (V/W)^\vee \to W^\perp$ defined by $\Phi(\lambda)(v) = \lambda(v + W)$ is an isomorphism.

Theorem 22.14. *Let F/k be a function field and let Ω be its space of Weil differentials. Then $\dim_F \Omega = 1$.*

Proof. Clearly $\Omega \neq 0$, so let $\omega_1, \omega_2 \in \Omega$ be nonzero. We will show that $\omega_1/\omega_2 \in F$.

For $i = 1, 2$, let D_i be such that $\omega_i \in \Omega(D_i)$ and define the k-linear map

$$\begin{aligned} \phi_i \colon \mathcal{L}(D_i + D) &\to \Omega(-D), \\ f &\mapsto f\omega_i, \end{aligned}$$

where D is a fixed divisor to be determined. For any $\alpha + g$ in $\mathcal{A}(-D) + F$ we have

$$(f\omega_i)(\alpha + g) = \omega_i(f\alpha) + \omega_i(fg) = 0 + 0 = 0,$$

since ω_i vanishes on $fg \in F$ and

$$\mathrm{ord}_P(f\alpha) = \mathrm{ord}_P(f) + \mathrm{ord}_P(\alpha) \geq \mathrm{ord}_P(-D_i - D) + \mathrm{ord}_P(D) = \mathrm{ord}_P(-D_i)$$

for all P, so ω_i vanishes on $f\alpha$. Thus ϕ_i is well defined, and it is clearly injective.

We claim that for an appropriate choice of D we have

$$\phi_1(\mathcal{L}(D_1 + D)) \cap \phi_2(\mathcal{L}(D_2 + D)) \neq \{0\}. \tag{2}$$

Assuming the claim, we may pick nonzero $f_1 \in \mathcal{L}(D_1 + D)$ and $f_2 \in \mathcal{L}(D_2 + D)$ such that $\phi_1(f_1) = \phi_2(f_2)$. Then $f_1\omega_1 = f_2\omega_2$ and $\omega_1/\omega_2 = f_2/f_1 \in F$ as desired.

We now prove that there is a divisor D for which (2) holds. By Riemann's Theorem, we can pick $D > 0$ of sufficiently large degree so that $r(D_i + D) = g - 1$ for $i = 1, 2$. Let U_i be the image of $\mathcal{L}(D_i + D)$ in $\Omega(-D)$ under ϕ_i. We want to show $\dim(U_1 \cap U_2) > 0$. We have

$$\dim \Omega(-D) = i(-D) = g - 1 - r(-D) = g - 1 - \deg(-D) + \ell(-D) = g - 1 + \deg(D),$$

since $\ell(-D) = 0$ for $D > 0$. We have $U_1 + U_2 \subseteq \Omega(-D)$, and therefore

$$\dim \Omega(-D) \geq \dim(U_1 + U_2) = \dim U_1 + \dim U_2 - \dim(U_1 \cap U_2),$$

where all the dimensions are as k-vector spaces. Thus

$$\begin{aligned} \dim(U_1 \cap U_2) &\geq \dim U_1 + \dim U_2 - \dim \Omega(-D) \\ &= \ell(D_1 + D) + \ell(D_2 + D) - g + 1 - \deg D \\ &= \deg(D_1 + D) - r(D_1 + D) + \deg(D_2 + D) - r(D_2 + D) - g + 1 - \deg D \\ &= \deg(D_1 + D) - g + 1 + \deg(D_2 + D) - g + 1 - g + 1 - \deg D \\ &= \deg D + \deg D_1 + \deg D_2 - 3g + 3. \end{aligned}$$

By choosing D of sufficiently large degree, we can make the RHS positive. □

Lemma 22.15. *For any nonzero $\omega \in \Omega$ there is a unique divisor D_ω such that $D \leq D_\omega$ for all divisors D for which $\omega \in \Omega(D)$.*

Proof. By Lemma 22.13, we have $\dim \Omega(D) = i(D)$, so $i(D) > 0$ for all divisors D such that $\omega \in \Omega(D)$. At least one such D exists, since $\omega \in \Omega$, so let us choose D_ω maximal subject to the constraint $\omega \in \Omega(D_\omega)$; a maximal D_ω exists because $i(D) = 0$ for all D of sufficiently large degree, by Riemann's Theorem. We now prove that D_ω is unique.

Suppose not. Then there are two distinct divisors D_1 and D_2 that are maximal subject to the constraints $\omega \in \Omega(D_1)$ and $\omega \in \Omega(D_2)$. Since D_1 and D_2 are incomparable, there exist distinct places P_1 and P_2 such that

$$\operatorname{ord}_{P_1}(D_1) > \operatorname{ord}_{P_1}(D_2) \quad \text{and} \quad \operatorname{ord}_{P_2}(D_2) > \operatorname{ord}_{P_2}(D_1).$$

We claim that $\omega \in \Omega(D_1 + P_2)$, contradicting the maximality of D_1. Write $\alpha \in \mathcal{A}(D_1 + P_2)$ as $\alpha = \alpha_1 + \alpha_2$, where α_1 is zero at P_2 and equal to α otherwise, while α_2 is equal to α at P_2 and zero otherwise. Then $\alpha_1 \in \mathcal{A}(D_1)$ and $\alpha_2 \in \mathcal{A}(D_2)$ and $\omega(\alpha) = \omega(\alpha_1) + \omega(\alpha_2) = 0$, since $\omega \in \Omega(D_1)$ and $\omega \in \Omega(D_2)$. But then $\omega \in \Omega(D_1 + P_2)$ as claimed. □

Definition 22.16. For a nonzero Weil differential $\omega \in \Omega$ we define the *divisor of* ω to be the unique divisor

$$\operatorname{div} \omega := D_\omega$$

given by Lemma 22.15. A divisor D is said to be *canonical* if $D = \operatorname{div} \omega$ for some $\omega \in \Omega$. We also define $\operatorname{ord}_P(\omega) := \operatorname{ord}_P(\operatorname{div} \omega)$.

Lemma 22.17. *For any nonzero $f \in F$ and nonzero $\omega \in \Omega$ we have*

$$\operatorname{div}(f\omega) = \operatorname{div} f + \operatorname{div} \omega.$$

Proof. We have $f\omega \in \Omega(\operatorname{div} f + \operatorname{div} \omega)$, since for any $g + \alpha$ in $\mathcal{A}(\operatorname{div} f + \operatorname{div} \omega) + F$:

$$(f\omega)(g + \alpha) = \omega(fg + f\alpha) = \omega(fg) + \omega(f\alpha) = 0 + 0 = 0,$$

because ω vanishes on F and

$$\operatorname{ord}_P(f\alpha) = \operatorname{ord}_P(f) + \operatorname{ord}_P(\alpha) \geq \operatorname{ord}_P(\operatorname{div} f) + \operatorname{ord}_P(-D - \operatorname{div} f) = \operatorname{ord}_P(-D)$$

for all places P of F, so $\omega(f\alpha) = 0$. It follows that $\operatorname{div} f\omega \geq \operatorname{div} f + \operatorname{div} \omega$.

The same argument shows that $\operatorname{div} \omega = \operatorname{div} f^{-1} f\omega \geq \operatorname{div} f^{-1} + \operatorname{div} f\omega = \operatorname{div} f\omega - \operatorname{div} f$, and therefore $\operatorname{div} f\omega \leq \operatorname{div} f + \operatorname{div} \omega$, so the claimed equality holds. □

Corollary 22.18. *The canonical divisors form a single linear equivalence class.*

Proof. Let $D_1 = \operatorname{div} \omega_1$ and $D_2 = \operatorname{div} \omega_2$ be two canonical divisors for F/k. Then ω_1 and ω_2 are both nonzero, and by Theorem 22.14, we have $\omega_2 = f\omega_1$ for some $f \in F^\times$. But then $D_2 = \operatorname{div} \omega_2 = \operatorname{div} f\omega_1 = \operatorname{div} f + \operatorname{div} \omega_1 = \operatorname{div} f + D_1$, so $D_1 \sim D_2$.

Now suppose $D_1 = \operatorname{div} \omega_1$ is a canonical divisor and $D_2 = D_1 + \operatorname{div} f$ for some $f \in F^\times$. Then $D_2 = \operatorname{div} \omega_1 + \operatorname{div} f = \operatorname{div} f\omega_1$ is canonical. □

Thus their is a unique element of the Picard group $\operatorname{Pic}_k C$ corresponding to the class of canonical divisors. This is a truly remarkable fact; given the rather abstract definition of the Picard group, there is no *a priori* reason to expect that it should have a uniquely distinguished element other than zero. As we shall see in the next lecture, the canonical divisor class is typically not the zero divisor, and the case where it is is actually quite interesting.

We now show that, like elements of the function field, Weil differentials are determined up to a scalar factor in $k^\times$ by their divisors.

Corollary 22.19. *Two nonzero Weil differentials $\omega_1, \omega_2 \in \Omega$ have the same divisor if and only if one $\omega_2 = c\omega_1$ for some $c \in k^\times$.*

Proof. Since $\omega_1 \neq 0$ and $\dim_F \Omega = 1$, we can write $\omega_2 = f\omega_1$ for some $f \in F^\times$. Then $\operatorname{div} \omega_2 = \operatorname{div} f\omega_1 = \operatorname{div} f + \operatorname{div} \omega_1$, so if $\operatorname{div} \omega_1 = \operatorname{div} \omega_2$ then $\operatorname{div} f = 0$ and $f \in k^\times$. Conversely, $\operatorname{div} \omega_2 = \operatorname{div} c\omega_1 = \operatorname{div} c + \operatorname{div} \omega_1 = \operatorname{div} \omega_1$, for any $c \in k^\times$. □

22.4 The Riemann-Roch Theorem

We now have almost everything we need to prove the Riemann-Roch Theorem. The last ingredient is the Duality Theorem, which gives us an isomorphism between Riemann-Roch spaces and spaces of Weil differentials.

Theorem 22.20 (Duality)**.** *For any divisor D and canonical divisor $W = \operatorname{div}\omega$, the linear map $\phi\colon \mathcal{L}(W-D) \to \Omega(D)$ defined by $\phi(f) = f\omega$ is an isomorphism of k-vector spaces. In particular, we have $i(D) = \ell(W-D)$ for all divisors D.*

Proof. For any nonzero $f \in \mathcal{L}(W-D)$ and $\omega \in \Omega(D)$ we have

$$\operatorname{div} f\omega = \operatorname{div} f + \operatorname{div}\omega \geq -(W-D) + W = D,$$

thus $f\omega \in \Omega(D)$, and $\operatorname{im}\phi \subseteq \Omega(D)$. It is clear that ϕ is linear, and its kernel is obviously trivial, so it is injective. To show that ϕ is surjective, let ω' be any nonzero element of $\Omega(D)$. By Theorem 22.14 we can write $\omega' = f\omega$ for some $f \in F^\times$, and since

$$\operatorname{div} f + W = \operatorname{div} f + \operatorname{div}\omega = \operatorname{div} f\omega = \operatorname{div}\omega_1 \geq D,$$

we have $\operatorname{div} f \geq -(W-D)$ and therefore $f \in \mathcal{L}(W-D)$, so $\omega' = \phi(f)$. Thus ϕ is surjective, hence an isomorphism, and $i(D) = \dim\Omega(D) = \ell(W-D)$, by Lemma 22.13. □

Theorem 22.21 (Riemann-Roch Theorem)**.** *Let W be a canonical divisor of the genus g curve C/k. For every divisor D we have*

$$\ell(D) = \deg(D) + 1 - g + \ell(W-D).$$

Proof. Immediate from Definition 22.4 and Theorem 22.20. □

Corollary 22.22. *For any canonical divisor W of a genus g curve we have*

$$\ell(W) = g, \qquad \deg W = 2g-2, \qquad i(W) = 1.$$

Proof. We apply the Riemann-Roch Theorem twice, first with $D = 0$, which gives

$$\ell(0) = \deg 0 + 1 - g + \ell(W),$$

and since $\deg 0 = 0$ and $\ell(0) = 1$, we have $\ell(W) = g$. Taking $D = W$ gives

$$\ell(W) = \deg W + 1 - g + \ell(0),$$

which implies $\deg W = 2g-2$ and $i(W) = \ell(W-W) = 1$. □

We can now give an exact value for the constant c in Riemann's Theorem.

Corollary 22.23. *For all divisors D of a genus g curve C/k with $\deg D > 2g-2$ we have*

$$\ell(D) = \deg D + 1 - g,$$

equivalently, $i(D) = 0$.

Proof. By the Riemann-Roch Theorem,

$$\ell(D) = \deg(D) + 1 - g + \ell(W - D)$$

where W is a canonical divisor. We have

$$\deg(W - D) = \deg W - \deg D < 2g - 2 - (2g - 2) = 0,$$

so $\ell(W - D) = 0$ and the corollary follows. □

We can also give some more down-to-earth characterization of a canonical divisor.

Corollary 22.24. *For a divisor D of a genus g curve, the following are equivalent:*

(a) *D is a canonical divisor.*

(b) *$\ell(D) = g$ and $\deg D = 2g - 2$.*

(c) *$i(D) = 1$ and $\deg D$ is maximal among divisors with $i(D) = 1$.*

Proof. That (a) implies (b) is immediate from Corollary 22.22, and the implications (b)⇒(c) and (c)⇒(a) both follow from the combination of Corollaries 22.22 and 22.23. □

Finally we note a very useful fact.

Theorem 22.25. *The genus of a curve C/k over a perfect field k is preserved under base extension.*[3]

Proof. Let k'/k be an extension of the perfect field k (hence a separable extension). It suffices to show that if $D \in \operatorname{Div}_k C$ is a canonical divisor for C/k then it is also a canonical divisor for C/k'. Clearly $\deg D$ is not changed under base extension (some closed points may split, but the total degree does not change), so it suffices to show that the dimension $\ell(D)$ of the Riemann-Roch space $\mathcal{L}(D) \subseteq k(C)$ does not change under base extension. The key point here is that any finite-dimensional k'-vector subspace of $k'(C)$ has a basis that lies in $k(C)$; this follows from a general algebraic result that we will not prove here; see Proposition 1 in §3 of the appendix to [1]. Thus $\ell(D)$ does not change under base extension. □

References

[1] I. R. Shafarevich, *Basic algebraic geometry*, 2nd edition, Springer-Verlag, 1994.

[2] H. Stichtenoth, *Algebraic function fields and codes*, Springer, 2009.

[3] This theorem is not necessarily true when k is not perfect. It is possible for the genus to decrease under an inseparable base extension.

18.782 Introduction to Arithmetic Geometry **Fall 2013**
Lecture #23 **11/26/2013**

As usual, a curve is a smooth projective (geometrically irreducible) variety of dimension one and k is a perfect field.

23.1 Genus zero curves with a rational point

Earlier in the course we showed that all plane conics with a rational point are isomorphic to $\mathbb{P}^1$. We now show that this applies to any genus zero curve with a rational point.

Theorem 23.1. *Let C/k be a curve with a rational point. Then C has genus zero if and only if it is isomorphic to $\mathbb{P}^1$ (over k).*

Proof. Every curve that is isomorphic to $\mathbb{P}^1$ has genus zero; this follows from Lemma 21.13 and the Riemann-Roch theorem. Conversely, for a curve of genus $g = 0$ with a rational point P, the Riemann-Roch theorem implies

$$\ell(P) = \deg(P) + 1 - g = 1 + 1 - 0 = 2,$$

since $\deg P = 1 > 2g - 2 = -2$. Thus there exists a non-constant function $f \in \mathcal{L}(P)$, and such an f has a simple pole at P and no other poles. It follows that $\operatorname{div}_\infty f = \deg P = 1$, hence f gives a degree-one morphism from C to $\mathbb{P}^1$ that is defined over k, since $f \in k(C)$, and this is an isomorphism (here we use Corollaries 19.3 and 19.5). □

Remark 23.2. If C/k does not have a rational point, we might instead let P be any closed point (these always exist). Everything in the above proof works except that now we have $\operatorname{div}_\infty f = \deg P > 1$. The function f still defines a morphism to $\mathbb{P}^1$, but it is not an isomorphism because its degree is greater than one. But if we base extend C/k to a finite extension k'/k over which the place P splits into degree one places, then we can show that C/k' is isomorphic to $\mathbb{P}^1$. So every curve of genus zero is isomorphic to $\mathbb{P}^1$ over a finite extension of its ground field.

23.2 Genus one curves with a rational point

Theorem 23.3. *Let C/k be a curve with a rational point. Then C has genus one if and only if it is isomorphic to a plane curve of the form*

$$y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6, \tag{1}$$

with $a_1, a_2, a_3, a_4, a_6 \in k$.

Proof. Let C/k be a curve of genus one with a rational point P. For any positive integer n we have $\deg nP = n > 2g - 2 = 0$, so by the Riemann-Roch theorem,

$$\ell(nP) = \deg(nP) + 1 - g = n + 1 - 1 = n.$$

In particular, $\mathcal{L}(2P)$ has dimension 2. Clearly $k \in \mathcal{L}(2P)$, since $0 \geq -2P$, so $\mathcal{L}(2P)$ has a k-basis of the form $\{1, x\}$ for some $x \in k(C) - k$. The space $\mathcal{L}(3P)$ contains $\mathcal{L}(2P)$ and has dimension 3, so it has a basis of the from $\{1, x, y\}$ for some $y \in k(C)^\times$. The functions $1, x, y, x^2$ all belong to $\mathcal{L}(4P)$ and have poles of distinct orders $0, 2, 3, 4$ at P, respectively,

Andrew V. Sutherland

thus they are linearly independent and form a basis for $\mathcal{L}(4P)$. By the same argument, $(1, x, y, x^2, xy)$ is a basis for $\mathcal{L}(5P)$.

But $\mathcal{L}(6P)$ contains both x^3 and y^2, as well as $1, x, y, x^2, xy$. Thus we have 7 elements in a k-vector space of dimension 6, and these must satisfy a linear equation. This equation must contain terms ax^3 and by^2 with $a, b \neq 0$ (otherwise we are left with a linearly independent set of terms), and if we replace x by ax/b and y by by/a, after multiplying through by b^3/a^4 and homogenizing we obtain an equation in the desired form (1).

Now suppose we have a curve C/k defined by an equation of the form (1). If we homogenize (1) and use projective coordinates $(x : y : z)$, then $P = (0 : 1 : 0)$ is a rational point, and it is clearly the only point on C (rational or otherwise) with $z = 0$, since $z = 0$ forces $x = 0$ and all points $(0 : y : 0)$ are projectively equivalent.

The function x (projectively represented as x/z) defines a morphism $(x : z)$ from C to $\mathbb{P}^1$ of degree $[k(C) : k(x)] = 2$, since $k(C) = k(x, y)$ and the minimal equation of y over $k(x)$ has degree 2 (note that C is a curve, and in particular an irreducible algebraic set, so equation (1) must be irreducible). It follows that $\operatorname{div}_\infty x = 2$ (by Corollary 19.23), and since the function x has a pole only at points with z-coordinate 0, it must have a double pole at P. By the same argument, the function y has a pole of order 3 at P. The set of functions $\{x^i y^j\}$ contains elements with poles of order $n = 2i + 3j$ at P for $n = 0$ and all $n \geq 2$, and none of these functions has any other poles. Thus we can construct a set of n linearly independent functions with poles of order $0, 2, 3, \ldots, n$, all of which lie $\mathcal{L}(nP)$. Applying the Riemann-Roch theorem with n sufficiently large yields

$$n \leq \ell(nP) = \deg(nP) + 1 - g = n + 1 - g,$$

so the genus g of C is at most 1.

To show that $g \neq 0$, consider the rational map ι defined by $(x : -y - a_1x - a_3z : z)$. The map ι leaves the RHS of (1) unchanged, and on the LHS we have

$$y(y + a_1x + a_3z) \mapsto (-y - a_1x - a_3z)(-y - a_1x - a_3z + a_1x + a_3z) = (y + a_1x + a_3z)y,$$

which is also unchanged, so ι is a morphism from C to itself. The morphism ι is clearly invertible (it is its own inverse), so it is an automorphism. Let us now determine the points fixed by ι. Clearly $(0 : 1 : 0)$ is fixed, and a point with $z \neq 0$ is fixed if and only if $y = -y - a_1x - a_3z$. Assuming $\operatorname{char}(k) \neq 2$, this is equivalent to $y = -(a_1x + a_3z)/2$. There are then three possibilities for x, corresponding to the roots of the cubic

$$x^3 + a_2x^2 + a_4x + a_6z + (a_1x + a_3z)^2/4.$$

These roots are distinct, since a repeated root would correspond to a singularity on the smooth curve C. Thus ι fixes exactly 4 points in $\bar{k}(C)$. If $g = 0$, then C is isomorphic to $\mathbb{P}^1$, by Theorem 23.1, and the only automorphism of $\mathbb{P}^1$ that fixes four points in $\mathbb{P}^1$ is the identity map, by Lemma 23.7 below. But ι is clearly not the identity map on $\bar{k}(C)$, indeed, it fixes only the 4 points already mentioned, thus $g \neq 0$.

If $\operatorname{char}(k) = 2$ one needs a different argument to show $g \neq 0$; see [2]. □

Corollary 23.4. *Every genus one curve C/k with a rational point is isomorphic to a plane cubic curve.*

Remark 23.5. It is also true that every (smooth) plane cubic has genus one, but we won't prove this here. The fact that genus one curves with a rational point can always be embedded in $\mathbb{P}^2$ is noteworthy; the corresponding statement is already false in genus 2.

Remark 23.6. The automorphism ι used in the proof of Theorem 23.3 is an example of an *involution*, an automorphism whose composition with itself is the identity map.

We now prove the lemma used in the proof of Theorem 23.3.

Lemma 23.7. *Suppose ϕ is an automorphism of $\mathbb{P}^1$ that fixes more than 2 points in $\mathbb{P}^1(\bar{k})$. Then ϕ is the identity map.*

Proof. Without loss of generality, we assume ϕ fixes the point $\infty = (1 : 0)$; if not we can apply a linear transformation to $\mathbb{P}^1$ that moves a point fixed by ϕ to ∞. The restriction ϕ_a of ϕ to $\mathbb{A}^1(\bar{k}) = \mathbb{P}^1(\bar{k}) - \{\infty\}$ is then a bijection, and also a morphism of affine varieties. As a morphism from $\mathbb{A}^1 \to \mathbb{A}^1$ the map ϕ_a is a polynomial map, say $\phi_a = (f)$, and f must have degree one since ϕ_a is a bijection. If the equation $f(x) = x$ has more than one solution, then both sides must be equal as polynomials of degree one (two points uniquely determine a line), but then ϕ_a is the identity map. □

Remark 23.8. One can extend the argument above to show that every automorphism of $\mathbb{P}^1$ is a rational function of the form $(ax + by)/(cx + dy)$ with $ad - bc \neq 0$, also known as a *Möbius transformation*. It is easy to see that every non-trivial Möbius transformation fixes exactly 2 points (over $\bar{k}$); they correspond to rotations of the Riemann sphere.

Definition 23.9. Equation (1) in Theorem 23.3 is called a *Weierstrass equation*.

Remark 23.10. There is no a_5 in a Weierstrass equation. As can be seen from the proof of Theorem 23.3, each coefficient a_i appears in front of a function with a pole of order $6 - i$ at the given rational point (and no other poles). There are no functions with only a single pole of order $6 - 5 = 1$ on a curve of genus one (indeed, such a function would give an isomorphism to $\mathbb{P}^1$).

Lemma 23.11. *Let C/k be a curve defined by a Weierstrass equation*

$$y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6.$$

If the characteristic of k is not 2 (resp. not 2 or 3) then a_1 and a_3 (resp. $a_1, a_2,$ and a_3) can be made zero via a linear change of coordinates.

Proof. If $\operatorname{char}(k) \neq 2$ then we can complete the square on the LHS, writing it as

$$(y + (a_1x + a_3)/2)^2 - (a_1x + a_3)^2/4.$$

Setting $u = y + (a_1x + a_3)/2)$ and moving the remaining terms to the RHS yields

$$u^2 = x^3 + a_2'x^2 + a_4'x + a_6',$$

for some $a_2', a_4', a_6' \in k$. If we also have $\operatorname{char}(k) \neq 3$, we can depress the cubic on the RHS by setting $v = x + a_2'/3$, yielding

$$u^2 = v^3 + a_4''v + a_6''$$

with $a_4'', a_6^{''} \in k$. □

Definition 23.12. A Weierstrass equation with $a_1a_2a_3 = 0$ is a *short Weierstrass equation*.

Lemma 23.13. *The short Weierstrass equation $y^2 = x^3 - a_4x - a_6$ defines a genus one curve if and only if $4a_4^3 + 27a_6^2 \neq 0$.*

Proof. The partial derivatives of $f(x, y, z) = y^2z - x^3 - a_4xz^2 - a_6z^3$ are

$$\begin{aligned}\partial f/\partial x &= -3x^2 - a_4z^2,\\ \partial f/\partial y &= 2yz,\\ \partial f/\partial z &= y^2 - 2a_4xz - 3a_6z^2.\end{aligned}$$

Let X be the zero locus of f in $\mathbb{P}^2$. If $P = (x_0 : y_0 : z_0)$ is a singular point of X, then $z_0 \neq 0$ (otherwise we must have $x_0 = y_0 = 0$, but this is not a valid projective point). We then have $y_0 = 0$, (a) $3x_0^2 + a_4z_0^2 = 0$, and (b) $2a_4x_0z_0 + 3a_6z_0^2 = 0$. Writing $x_0 = -3a_6z_0/(2a_4)$ via (b) and plugging into (a) gives

$$\begin{aligned}3(-3a_6z_0)^2/(2a_4)^2 + a_4z_0^2 &= 0\\ 27a_6^2 + 4a_4^3 &= 0.\end{aligned}$$

The calculations above are reversible, so X has a singular point if and only if $4a_4^3+27a_6^2 = 0$. If X has no singular points then it must be irreducible; it is a hypersurface in $\mathbb{P}^2$ and it were the union of two or more curves, then the intersection points would be singular. Thus X is a curve defined by a Weierstrass equation with the rational point $(0 : 1 : 0)$, so by Theorem 23.3 it has genus one. □

In what follows we may assume for the sake of simplicity, that the characteristic of k is not 2 or 3 so that we can work with short Weierstrass equations; everything we do can be extended to the general case, the equations are just more complicated to write down.

23.3 Elliptic curves

Definition 23.14. An *elliptic curve* E/k is a genus one curve with a distinguished rational point O. Equivalently, an elliptic curve is a curve E/k defined by a Weierstrass equation with the distinguished rational point $O = (0 : 1 : 0)$.

Notice that an elliptic curve E/k is, strictly speaking, more than just the curve E, it is the pair (E, O). If E has two rational points, say O_1 and O_2, then (E, O_1) and (E, O_2) are two different elliptic curves. In practice one typically works with elliptic curves given by Weierstrass equations, in which case the point O is always taken to be the point $(0 : 1 : 0)$ at infinity; thus we may refer to E/k as an elliptic curve without explicitly mentioning O.

Remark 23.15. Elliptic curves are obviously *not ellipses* (ellipses are curves of genus zero), but there is a connection. If one attempts to compute the circumference of an ellipse with semi-major axis a and eccentricity e by applying the arc-length formula, one finds that the circumference is given by

$$4a\int_0^1 \sqrt{\frac{1-e^2t}{1-t^2}}dt.$$

This is known as an *elliptic integral* (incomplete, and of the second kind), and it does not have a simple closed form. However, the integrand $u(t)$ satisfies the equation

$$u^2(1-t^2) = 1-e^2t^2,$$

and this defines a genus one curve with a rational point, an elliptic curve.[1] The theory of elliptic curves originated in the study of solutions to integrals like the one above, leading to the notion of *elliptic functions* that arise in complex analysis as solutions to non-linear differential equations that correspond to Weierstrass equations.

Theorem 23.16. *Let E/k be an elliptic curve with distinguished point O. The map ϕ that sends the point $P \in E(k)$ to the class of the divisor $P - O$ in $\operatorname{Pic}_k^0(E)$ is a bijection. This induces a commutative group operation on $E(k)$ defined by*

$$P_1 + P_2 := \phi^{-1}(\phi(P_1) + \phi(P_2)),$$

in which O acts as the identity element.

Proof. We first show that ϕ is injective. If $P - O \sim Q - O$ then $P \sim Q$. We then have $\operatorname{div} f = P - Q$ for some $f \in k(E)$. If f is nonzero then it gives an isomorphism $E \to \mathbb{P}^1$, which is impossible, since E has genus one. So $P = Q$ and ϕ is injective.

Now suppose D is any divisor of degree 0. Then $D + O$ has degree $1 \geq 2g - 1 = 1$, and by the Riemann-Roch theorem

$$\ell(D + O) = \deg(D + O) + 1 - g = 1 + 1 - 1 = 1,$$

so there is a nonzero $f \in \mathcal{L}(D+O)$ such that $\operatorname{div} f + D + O \geq 0$. But $\deg(\operatorname{div} f + D + O) = 1$, so we must have $\operatorname{div} f + D + O = P$ for some $P \in E(k)$. Thus $D \sim P - O$. □

Thus the set of rational points $E(k)$ form an abelian group. The same applies to every base extension of E, so we the set $E(k')$ is also an abelian group (also with O as the identity), for any extension k'/k; this follows from the fact that the genus of a curve is preserved under base extension (of a perfect field), so E/k' is also an elliptic curve.[2]

We now want to describe the group operation more explicitly. For this purpose we use the following construction. Let us assume our elliptic curve E/k is given by a Weierstrass equation, hence embedded in $\mathbb{P}^2$. If L is a line in $\mathbb{P}^2$ defined by a linear form (a homogeneous polynomial of degree one) with coefficients in k, then the intersection $(L \cap E)$ corresponds to a divisor in $D_L = \operatorname{Div}_k E$ of degree 3. This follows from Bezout's Theorem, and the fact that $(L \cap E)$ is fixed by the action of $G_k = \operatorname{Gal}(\bar{k}/k)$; the set $(L \cap E)$ is a union of Galois orbits, each a closed point of E/k, and each occurs in D_L with multiplicity corresponding to the intersection number of E and L at each $\bar{k}$-point in the orbit (these all must coincide).

Lemma 23.17. *Let E/k be an elliptic curve in $\mathbb{P}^2$, and let L_1, L_2 be lines in $\mathbb{P}^2$ defined by linear homogeneous polynomials $\ell_1, \ell_2 \in k[x, y, z]$. Let $f \in k(E)$ be image of ℓ_1/ℓ_2 under the map $k(\mathbb{P}^2) \to k(E)$ induced by the inclusion $E \subseteq \mathbb{P}^2$. Then*

$$\operatorname{div} f = (L_1 \cap E) - (L_2 \cap E).$$

Proof. This follows from Bezout's theorem and the discussion above. □

We now give an explicit description of the group operation on an elliptic curve $E(k)$ defined by a Weierstrass equation. Any two points P and Q in $E(k)$ uniquely determine a line L_1 that is defined over k (if $P = Q$ then take the line tangent to E at $P = Q$). By Bezout's Theorem, $L \cap E$ contains a third point R, and this point must lie in $E(k)$ because

[1] The given equation is singular at $u = 0$ and $t = \pm 1$, but its desingularization is an elliptic curve.

[2] This is not always true when k is not perfect.

P, Q, and $L_1 \cap E$ are all fixed by G_k. If we now let L_2 be the line $z = 0$ at infinity, and let ℓ_1 and ℓ_2 be the linear forms defining L_1 and L_2, then

$$\operatorname{div} \ell_1/\ell_2 = (L_1 \cap E) - (L_2 \cap E) = P + Q + R - 3O = (P - O) + (Q - O) + (R - O)$$

since O is the only point on E with $z = 0$ (so by Bezout's Thoerem, the intersection number $I_P(L_2 \cap E)$ must be 3). This divisor is principal, hence equivalent to the zero divisor, and in terms of the group operation on $E(K)$ this implies

$$P \oplus Q \oplus R = O,$$

where the symbol $\oplus$ denotes the group operation on $E(k)$.[3] and we recall that O is the identity element of the group operation. This is summed up in the following corollary.

Corollary 23.18. *Let E/k be an elliptic curve defined by a Weierstrass equation. The sum of any three points in $E(k)$ that lie on a line is zero under the group law on $E(k)$.*

Corollary 23.18 completely determines the group operation on $E(k)$. To avoid ambiguity, we will temporarily use $\oplus$ to denote the group operation on $E(k)$, in order to distinguish it from addition in $\operatorname{Div}_k E$. Given any two points P and Q we compute their sum $R = P \oplus Q$ by noting that $P \oplus Q = R$ holds if and only if $P \oplus Q \ominus R = O$, so we may compute the negation of R as the third point on the line determined by P and Q. To get R itself, we use the fact that $R \ominus R = O$ if and only if $O \oplus R \ominus R = O$, so we obtain R as the third point on the line determined O and the negation of R. To sum up, the group law on $E(k)$ can be defined as follows.

Corollary 23.19 (Geometric group law)**.** *Let P and Q be rational points on elliptic curve embedded in $\mathbb{P}^2$. Then $P \oplus Q$ is the negation of the third point in the intersection of E and the line uniquely determined by P and Q.*

For explicit computations, we can use the Weierstrass equation for E to compute the coordinates of the point $P \oplus Q$ as rational functions of the coordinates of P and Q. The case where either P or Q is equal to O is obvious, so we assume otherwise, in which case neither P nor Q lies on the line $z = 0$ at infinity and we can work in the affine patch $z \neq 0$.

In order to simplify the formulas, let us assume that $\operatorname{char}(k) \neq 2, 3$ so that E/k can be defined by a short Weierstrass equation

$$y^2 = x^3 + a_4 x + a_6. \tag{2}$$

The additive inverse of any affine point $P = (x_0 : y_0 : 1)$ is $(x_0 : -y_0 : 1)$, since the third point on the line $x - x_0 z$ determined by P and O (and of course O is its own inverse).

We now consider how to compute the sum of two affine points $P_1 = (x_1 : y_1 : z_1)$ and $P_2 = (x_2 : y_2 : z_2)$. Let us first dispose of some easy cases. If $x_1 = x_2$ then $y_1 = \pm y_2$, and if $y_1 = -y_2$ then P_2 is the negation of P_1 and their sum is O, so we assume this is not the case. We then have two possibilities, either $x_1 \neq x_2$, or $P_1 = P_2$. In the latter case, if $y_1 = 0$ then $P_1 = P_2$ is its own negation (a point of order two) and $P_1 \oplus P_2 = O$.

In every other case the slope λ of the line L determined by P_1 and P_2 is given by

$$\lambda = \frac{y_2 - y_2}{x_2 - x_1} \quad (x_1 \neq x_2), \qquad \lambda = \frac{3x_1^2 + a_4}{2y_1} \quad (P_1 = P_2 \text{ and } y_1 \neq 0),$$

[3] We use $\oplus$ here to avoid confusion with the symbol $+$ used to denote addition of divisors (and to write divisors as formal sums of closed points); later, when there is no risk of confusion we will simply use $+$ to denote the group operation on $E(k)$.

and $(y - y_1) = \lambda(x - x_1)$ is the equation for L.

Substituting this into (2) gives the cubic equation

$$\begin{aligned}(\lambda(x - x_1) + y_1)^2 &= x^3 + a_4x + a_6\\ 0 &= x^3 - \lambda^2x^2 + \cdots,\end{aligned}$$

whose solutions are precisely x_1, x_2, and x_3. We now observe that the sum of the roots of any cubic polynomial are equal to the negation of its quadratic coefficient, so $x_1 + x_2 + x_3 = \lambda^2$. This determines x_3; plugging x_3 into the equation for L and negating the result gives y_3.

Theorem 23.20 (Algebraic group law)**.** *Let E/k be an elliptic curve given by the short Weierstrass equation $y^2 = x^3 + a_4x + a_6$. If $P_1 = (x_1 : y_1 : 1)$ and $P_2 = (x_2 : y_2 : 1)$ are affine points whose sum is an affine point $P_3 = (x_3 : y_3 : z_3)$, then*

$$\begin{aligned}x_3 &= \lambda^2 - x_1 - x_2,\\ y_3 &= \lambda(x_1 - x_3) - y_1,\end{aligned}$$

where $\lambda = (y_2 - y_1)/(x_2 - x_1)$ if $P_1 \neq P_2$ and $\lambda = (3x_1^2 + a_4)/(2y_1)$ if $P_1 = P_2$.

Remark 23.21. One can define the group operation on $E(k)$ directly via either Corollary 23.19 or Theorem 23.20 (and one can extend Theorem 23.20 to general Weierstrass equations). But in order to show that this actually makes $E(k)$ a group, one must verify that the group operation is associative, and this is a surprisingly non-trivial exercise. The advantage of using the bijection between $E(k)$ and $\operatorname{Pic}_k^0 E$ given by Theorem 23.16 to define the group operation is that it is clear that this defines a group!

It follows from Theorem 23.20 that for any fixed point $P \in E(\bar{k})$, the *translation-by-P map* τ_P that sends Q to $P \oplus Q$ can be defined as a rational map (hence a morphism) from E to itself; it clearly has an inverse (replace P with its negation), so τ_P is an automorphism. The same is true of the *negation map* that sends P to its additive inverse. See [2, III.3.6] for details.

23.4 Abelian varieties

A variety whose points form a group such that the group operations are defined by morphisms, is called an *algebraic group*. More generally, this terminology is also applied to any open subset of a variety (a *quasi-variety*). Examples include $\mathbb{A}^1$, with the group operation given by addition of coordinates, and the general linear group GL_n, which can be viewed as an open subset of $\mathbb{A}^{n^2}$ corresponding to matrices with nonzero determinant.

It follows from Theorem 23.20 that an elliptic curve is an algebraic group, and in fact an *abelian variety*.

Definition 23.22. An *abelian variety* is a projective algrebraic group.

It follows from Theorem 23.20 that an elliptic curve is an abelian variety of dimension one. In fact, one can show that every abelian variety of dimension one is isomorphic to an elliptic curve. It might seem strange that the definition of an abelian variety does not include the requirement that group actually be abelian. This turns out to be a necessary consequence of requiring the algebraic group to be a *projective* variety. We will only prove this for abelian varieties of dimension one, but it is true in general.

Theorem 23.23. *An abelian variety is an abelian group.*

Proof in dimension one. Let G be an abelian variety of dimension one, and for any $h \in G$ consider the morphism $\phi_h \colon G \to G$ defined by $\phi_h(g) = g^{-1}hg$. Since G is a projective variety (hence complete), the image of ϕ_h is also a projective variety, which must be either a point or all of G. Let e be the identity element of G. For $h = e$ image of ϕ_h is clearly just the point $h = e$. For $h \neq e$ the image of ϕ_h cannot contain e, because $g^{-1}hg = e$ implies $hg = g$ and $h = e$. So the image of ϕ_h is a always a point, and it must be the point h, since $\phi_h(e) = h$. Thus for all $g, h \in G$ we have $\phi_h(g) = g^{-1}hg = h$, equivalently, $hg = gh$, so G is abelian.

The proof of Theorem 23.23 in the general case is essentially the same; one first shows that the dimension of $\operatorname{im}\phi_h$ must be the same for every $h \in G$, and since $\dim\operatorname{im}\phi_e = 0$, the image ϕ_h is a point for every $h \in G$ and the proof then proceeds as above; see [1, §4.3]. □

Now that we know abelian varieties are in fact abelian, we will write the group operation additively. When working with morphisms of abelian varieties it is natural to distinguish morphisms ϕ that preserve the group structure, that is, we would like $\phi(g+h) = \phi(g)+\phi(h)$. An obvious necessary condition is $\phi(0) = 0$. This turns out to be sufficient.

Theorem 23.24. *Let $\phi \colon G \to H$ be a morphism of abelian varieties for which $\phi(0) = 0$. Then ϕ is a group homomorphism.*

Proof. For each $h \in G$ let $\phi_h \colon G \to H$ be the morphism $\phi_h(g) = \phi(g) + \phi(h) - \phi(g + h)$. Then $\phi_0(g) = \phi(0) + \phi(g) - \phi(g + 0) = 0$ for all $g \in G$. As in the proof of Theorem 23.23, the image of ϕ_h is a single point for all $h \in G$, and since $\phi_h(0) = \phi(0)+\phi(h)-\phi(0+h) = 0$, that point must be 0. It follows that $\phi_h(g) = 0$ for all $g, h \in G$, therefore we always have $\phi(g) + \phi(h) = \phi(g + h)$ and ϕ is a group homomorphism. □

References

[1] I. R. Shafarevich, *Basic algebraic geometry*, 2nd edition, Springer-Verlag, 1994.

[2] J.H. Silverman, *The arithmetic of elliptic curves*, 2nd edition, Springer, 2009.

18.782 Introduction to Arithmetic Geometry **Fall 2013**
Lecture #24 **12/03/2013**

24.1 Isogenies of elliptic curves

Definition 24.1. Let E_1/k and E_2/k be elliptic curves with distinguished rational points O_1 and O_2, respectively. An *isogeny* $\varphi\colon E_1 \to E_2$ of elliptic curves is a surjective morphism that maps O_1 to O_2.

As an example, the negation map that send $P \in E(\bar{k})$ to its additive inverse is an isogeny from E to itself; as noted in Lecture 23, it is an automorphism, hence a surjective morphism, and it clearly fixes the identity element (the distinguished rational point O).

Recall that a morphism of projective curves is either constant or surjective, so any nonconstant morphism that maps O_1 to O_2 is automatically an isogeny. The composition of two isogenies is an isogeny, and the set of elliptic curves over a field k and the isogenies between them form a category; the identity morphism in this category is simply the identity map from an elliptic curve to itself, which is is clearly an isogeny. Given that the set of rational points on an elliptic curve form a group, it would seem natural to insist that, as morphisms in the category of elliptic curves, isogenies should preserve this group structure. But there is no need to put this requirement into the definition, it is necessarily satisfied.

Theorem 24.2. *Let E_1/k and E_2/k be elliptic curves and let $\varphi\colon E_1 \to E_2$ be an isogeny defined over k. Then φ is a group homomorphism from $E_1(L)$ to $E_2(L)$, for any algebraic extension L/k.*

Proof. This is essentially immediate (just consider the pushforward map on divisors), but let us spell out the details.

By base extension to L, it suffices to consider the case $L = k$. For $i = 1, 2$, let O_i be the distinguished rational point of E_i and let $\phi_i\colon E_i(k) \to \operatorname{Pic}^0_k E_i$ be the group isomorphism that sends $P \in E_i(k)$ to the divisor class $[P - O_i]$. Let $\varphi_*\colon \operatorname{Pic}^0_k E_1 \to \operatorname{Pic}^0_k E_2$ be the pushforward map on divisor classes of degree zero. For any $P \in E_1(k)$ we have $\varphi_*([P]) = [\varphi(P)]$, since P and $\varphi(P)$ both have degree one, and

$$\varphi_*(\phi_1(P)) = \varphi_*([P - O_1]) = [\varphi_*(P - O_1)] = [\varphi(P) - \varphi(O_1)] = [\varphi(P) - O_2] = \phi_2(\varphi(P)).$$

For any $P, Q \in E_1(k)$ with $P \oplus Q = R$ we have

$$\begin{aligned}
P \oplus Q &= R\\
\phi_1(P) + \phi_1(Q) &= \phi_1(R)\\
\varphi_*(\phi_1(P) + \phi(Q)) &= \varphi_*(\phi_!(R))\\
\varphi_*(\phi_1(P)) + \varphi_*(\phi_1(Q)) &= \varphi_*(\phi_1(R))\\
\phi_2(\varphi(P)) + \phi_2(\varphi(Q)) &= \phi_2(\varphi(R))\\
\varphi(P) \oplus \varphi(Q) &= \varphi(R),
\end{aligned}$$

where $\oplus$ denotes the group operation on both $E_1(k)$ and $E_2(k)$. □

Now that we know that an isogeny $\varphi\colon E_1 \to E_2$ is a group homomorphism, we can speak of its kernel $\varphi^{-1}(O_2)$. One can view $\ker \varphi$ as a set of closed points of E_1/k, but it is more useful to view it as a subgroup of $E_1(\bar{k})$.

Andrew V. Sutherland

Definition 24.3. Let $\varphi\colon E_1 \to E_2$ be an isogeny of elliptic curves over k. The *kernel* of φ, denoted $\ker\varphi$ is the kernel of the group homomorphism $\varphi\colon E_1(\bar{k}) \to E_2(\bar{k})$.

Recall the translation-by-Q automorphism $\tau_Q\colon E \to E$ that sends P to $P \oplus Q$. The induced map $\tau_Q^*\colon \bar{k}(E) \to \bar{k}(E)$ is an automorphism of the function field $\bar{k}(E)$.

Lemma 24.4. *Let $\varphi\colon E_1 \to E_2$ be an isogeny of elliptic curves. For each $P \in \ker\varphi$, the automorphism τ_P^* fixes $\varphi^*(\bar{k}(E_2))$, and the map $\ker\varphi \to \operatorname{Aut}(\bar{k}(E_1)/\varphi^*(\bar{k}(E_2)))$ defined by $P \mapsto \tau_P^*$ is an injective group homomorphism.*

Proof. Let $P \in \ker\varphi$ and let $f \in \bar{k}(E_2)$. Then

$$\tau_P^*(\varphi^*(f))(Q) = (f \circ \varphi \circ \tau_P)(Q) = f(\varphi(P \oplus Q)) = f(\varphi(Q)) = (f \circ \varphi)(Q) = \varphi^*(f)(Q),$$

since φ is a group homomorphism and P lies in its kernel. Thus τ_P fixes $\varphi^*(\bar{k}(E_2))$.

For any $P, Q \in \ker\varphi$ and $f \in \bar{k}(E_1)$ we have

$$\tau_{P\oplus Q}^*(f) = f \circ \tau_{P\oplus Q} = f \circ \tau_Q \circ \tau_P = \tau_P^*(f \circ \tau_Q) = \tau_P^*(\tau_Q^*(f)),$$

so $\tau_{P\oplus Q}^* = \tau_P^* \circ \tau_Q^*$, and the map $P \mapsto \tau_P^*$ is a group homomorphism. It is clearly injective, since if $P \neq Q$ then $P \ominus Q \neq O$ and $\tau_{P\ominus Q}^* = \tau_P^* \circ (\tau_Q^*)^{-1}$ is not the identity map (apply it to any nonconstant $f \in \bar{k}(E_1)$). □

Corollary 24.5. *For any isogeny $\varphi\colon E_1 \to E_2$ of elliptic curves, $\#\ker\varphi$ divides $\deg\varphi$. In particular, the kernel of an isogeny is finite.*

Proof. By definition, $\deg\varphi = [\bar{k}(E_1) : \varphi^*(\bar{k}(E_2))]$, and we know from Galois theory that the order of the automorphism group of a finite extension divides the degree of the extension. Since $\ker\varphi$ injects into $\operatorname{Aut}(\bar{k}(E_1)/\varphi^*(\bar{k}(E_2)))$, its order must divide $\deg\varphi$. □

Remark 24.6. In fact, the homomorphism in Lemma 24.4 is an isomorphism, and the corollary implies that when φ is separable we have $\#\ker\varphi = \deg\varphi$; see [1, III.4.10].

24.2 Torsion points on elliptic curves

Definition 24.7. Let E/k be an elliptic curve and let n be a positive integer. The *multiplication-by-n* map $[n]\colon E(\bar{k}) \to E(\bar{k})$ is the group homomorphism defined by

$$nP = P \oplus P \oplus \cdots \oplus P.$$

The points $P \in E(\bar{k})$ for which $nP = O$ are called *n-torsion points*. They form a subgroup of $E(\bar{k})$ denoted $E[n]$.

If $\varphi\colon E_1 \to E_2$ is an isogeny, then we know from Corollary 24.5 that $n = \deg\varphi$ is a multiple of the order of $\ker\varphi$. It follows that every point in $\ker\varphi$ is an n-torsion point. By definition, $[n]$ is a group homomorphism. We now show that $[n]$ is an isogeny.

Theorem 24.8. *The multiplication-by-n map on an elliptic curve E/k is an isogeny.*

Proof assuming $\operatorname{char}(k) \neq 2$*:* The case $n = 1$ is clear, and for $n = 2$ the map $P \mapsto P \oplus P$ is a rational map, hence a morphism (by Theorem 18.6, a rational map from a smooth projective curve is a morphism), since it can be defined in terms of rational functions of the coordinates of P via the algebraic formulas for the group operation on $E(\bar{k})$. More generally, given any morphism $\phi\colon E \to E$, plugging the coordinate functions of ϕ into the formulas for the group law yields a morphism that sends P to $\phi(P) \oplus P$. It follows by induction that $[n]$ is a morphism, and it clearly fixes the identity element O.

It remains to show that $[n]$ is surjective. For this it suffices to show that it does not map every point to O, since a morphism of smooth projective curves is either surjective or constant (by Corollary 18.7). We have already seen that there are exactly 4 points in $E(\bar{k})$ that are fixed by the negation map, three of which have order 2 (in short Weierstrass form, these are the point at infinity and the 3 points whose y-coordinate is zero). For n odd, $[n]$ cannot map a point of order 2 to O, so $[n]$ is surjective for n odd. For $n = 2^k m$ with m odd we may write $[n] = [2] \circ \cdots \circ [2] \circ [m]$. We already know that $[m]$ is surjective, so it suffices to show that $[2]$ is. But $[2]$ cannot map any of the infinitely many points in $E(\bar{k})$ that are not one of the 4 points fixed by the negation map to O, so $[2]$ must be surjective. □

Remark 24.9. Note that in characteristic 2 there are not four 2-torsion points, in fact there may be none. But one can modify the proof above to use 3-torsion points instead.

Corollary 24.10. *Let E/k be an elliptic curve. For any positive integer n, the number of n-torsion points in $E(\bar{k})$ is finite.*

Remark 24.11. In fact one can show that the number of n-torsion points divides n^2, and for n not divisible by $\mathrm{char}(k)$, is equal to n^2.

24.3 Torsion points on elliptic curves over $\mathbb{Q}$

Let E be an elliptic curve $\mathbb{Q}$, which we may assume is given by a short Weierstrass equation

$$E\colon y^2 = x^3 + a_4 x + a_6,$$

with $a_4, a_6 \in \mathbb{Q}$. Let d be the LCM of the denominators of a_4 and a_6. After multiplying both sides by d^6 and replacing y by $d^{3n}y$ and x by $d^{2n}x$, we may assume $a_4, a_6 \in \mathbb{Z}$. Since E is non-singular, we must have $\Delta = \Delta(E) := -16(4a_4^3 + 27a_6^2) \neq 0$.[1]

For each prime p the equation for E also defines an elliptic curve over $\mathbb{Q}_p$. For the sake of simplicity we will focus our attention on primes p that do not divide Δ, but everything we do below can be extended to arbitrary p (as we will indicate as we go along). Let E^0 denote the elliptic curve over $\mathbb{Q}_p$ obtained by base extension from $\mathbb{Q}$ to $\mathbb{Q}_p$. Let $\overline{E}/\mathbb{F}_p$ denote the curve over $\mathbb{F}_p$ obtained by reducing the equation for E modulo p. Here we are assuming $\Delta \not\equiv 0 \bmod p$ so that the reduced equation has no singular points, meaning that $\overline{E}$ is an elliptic curve. We say that E has *good reduction* at p when this holds.

The reduction map $E^0(\mathbb{Q}_p) \to \overline{E}(\mathbb{F}_p)$ is a group homomorphism, and we define $E^1(\mathbb{Q}_p)$ to be its kernel; these are the points that reduce to $(0:1:0)$ modulo p. In fact, $E^1(\mathbb{Q}_p)$ can be defined as the kernel of the reduction map regardless of whether E has good reduction at p or not and one can show that the points in $E^1(\mathbb{Q}_p)$ still form a group.

The points in $E^1(\mathbb{Q}_p)$ are precisely the points in $E^0(\mathbb{Q}_p)$ that can be represented as $(x:y:z)$, with $v_p(x), v_p(z) > 0$ and $v_p(y) = 0$; equivalently, the points with $v_p(x/y) > 0$ (note that $v_p(x/y)$ does not depend on how the coordinates are scaled). For all positive integers n we thus define

$$E^n(\mathbb{Q}_p) = \left\{(x:y:z) \in E^0(\mathbb{Q}_p) : v_p(x/y) \geq n\right\},$$

and note that this agrees with our previous definition of $E^1(\mathbb{Q}_p)$.

[1] The leading factor of -16 appears for technical reasons that we won't explain here, but it is useful to have a factor of 2 in Δ because a short Weierstrass equation always has singular points in characteristic 2.

Lemma 24.12. *For $n > 0$, each of the sets $E^{n+1}(\mathbb{Q}_p)$ is an index p subgroup of $E^n(\mathbb{Q}_p)$.*

Proof. Containment is clear from the definition, but we need to show that the sets $E^n(\mathbb{Q}_p)$ are actually groups. For $O = (0 : 1 : 0)$ we have $v_p(x/y) = \infty$, so $O \in E^n(\mathbb{Q}_p)$ for all n. Any affine point $P \in E^n(\mathbb{Q}_p) - E^{n+1}(\mathbb{Q}_p)$ has $v_p(x/y) = n$, and and after dividing through by z can be written as $(x : y : 1)$ with $v_p(y) < 0$. Since $a_4, a_6 \in \mathbb{Z}_p$, the equation $y^2 = x^3+a_4x+a_6$ implies $3v_p(x) = 2v_p(y)$, so

$$n = v_p(x/y) = v_p(x) - v_p(y) = -v_p(y)/3,$$

and therefore $v_p(y) = -3n$ and $v_p(x) = -2n$. After multiplying through by p^{3n} we can write $P = (p^n x_0 : y_0 : p^{3n})$ with $x_0, y_0 \in \mathbb{Z}_p^\times$. We then have

$$\begin{aligned} p^{3n}y_0^2 &= p^{3n}x_0^3 + a_4p^{7n}x_0 + a_6p^{9n} \\ y_0^2 &= x_0^3 + a_4p^{4n}x_0 + a_6p^{6n}. \end{aligned}$$

After reducing mod p we obtain an affine point $(\overline{x_0} : \overline{y_0} : 1)$ whose coordinates are all nonzero and which lies on the singular variety $C_0/\mathbb{F}_p$ defined by

$$y^2z = x^3,$$

which also contains the reduction of $O = (0 : 1 : 0)$. If we consider the image of the group law on $E^0(\mathbb{Q}_p)$ on $C_0(\mathbb{F}_p)$, we still have an operation defined by the rule that three colinear points sum to zero. We claim that this makes the set S of nonsingular points in $C_0(\mathbb{F}_p)$ into a group of order p. To show this, we first determine S. We have

$$\begin{aligned} (\partial/\partial x)(y^2z - x^3) &= -3x^2, \\ (\partial/\partial y)(y^2z - x^3) &= 2yz, \\ (\partial/\partial z)(y^2z - x^3) &= y^2. \end{aligned}$$

It follows that a point in $C_0(\mathbb{F}_p)$ is singular if and only if its y-coordinate is zero. In particular all of the reductions of points in $E^n(\mathbb{Q}_p)$ are non-singular, for any $n \geq 1$. Every non-singular point in $C_0(\mathbb{F}_p)$ can be written as $(x : 1 : x^3)$, and this gives a bijection from $\mathbb{F}_p$ to S defined by $x \mapsto (x : 1 : x^3)$. Thus the set S has order p and it contains the identity element. It is clearly closed under negation, and we now show it is closed under addition. If P and Q are two elements of S not both equal to $(0 : 1 : 0)$, then at least one of them has non-zero z-coordinate and the line L defined by P and Q can be written in the form $z = ax + by$. Plugging this into the curve equation gives

$$y^2(ax + by) = x^3,$$

and it is then clear that the third point R in $C_0 \cap L$ must have nonzero y-coordinate, since $y_0 = 0 \Rightarrow x_0 = 0 \Rightarrow z_0 = 0$ for any $(x_0 : y_0 : z_0) \in C_0 \cap L$. Since P and Q are both in $C_0(\mathbb{F}_p)$, so is R, thus R lies in S, as does its negation, which is $P \oplus Q$. Therefore the reduction map $E^n(\mathbb{Q}_p) \to C_0(\mathbb{F}_p)$ defines a group homomorphism from $E^n(\mathbb{Q}_p)$ to S, and its kernel is $E^{n+1}(\mathbb{Q}_p)$, an index p subgroup of $E^n(\mathbb{Q}_p)$. □

Definition 24.13. The infinite chain of groups

$$E^0(\mathbb{Q}_p) \supset E^1(\mathbb{Q}_p) \supset E^2(\mathbb{Q}_p) \supset \cdots$$

is called the *p-adic filtration* of $E/\mathbb{Q}$.

Theorem 24.14. *Let $E/\mathbb{Q}$ be an elliptic curve and let p be a prime not dividing $\Delta(E)$. The p-adic filtration of E satisfies*

(1) $E^0(\mathbb{Q}_p)/E^1(\mathbb{Q}_p) \simeq \overline{E}(\mathbb{F}_p)$;

(2) $E^n(\mathbb{Q}_p)/E^{n+1}(\mathbb{Q}_p) \simeq E(\mathbb{F}_p) \simeq \mathbb{Z}/p\mathbb{Z}$ *for all* $n > 0$;

(3) $\cap_n E^n = \{O\}$.

Proof. The group $E^1(\mathbb{Q}_p)$ is, by definition, the kernel of the reduction map from $E^0(\mathbb{Q}_p)$ to $\overline{E}(\mathbb{F}_p)$. To prove (1) we just need to show that the reduction map is surjective.

Let $P = (a_1 : a_2 : a_3)$ be a point in $\overline{E}(\mathbb{F}_p)$ with the $a_i \in \mathbb{Z}/p\mathbb{Z}$. The point P is non-singular, so at least one of the partial derivatives of the curve equation $f(x_1, x_2, x_3) = 0$ for E does not vanish. Without loss of generality, suppose $\partial f/\partial x_1$ does not vanish at P. If we pick an arbitrary point $\hat{P} = (\hat{a}_1 : \hat{a}_2 : \hat{a}_3)$ with coefficients $\hat{a}_i \in \mathbb{Z}_p$ such that $\hat{a}_i \equiv a_i \bmod p$, we can apply Hensel's to the polynomial $g(t) = f(t, \hat{a}_2, \hat{a}_3)$ using $a_1 \in \mathbb{Z}/p\mathbb{Z}$ as our initial solution, which satisfies $g'(a_i) \neq 0$. This yields a point in $E^0(\mathbb{Q}_p)$ that reduces to P, thus the reduction map is surjective, which proves (1).

Property (2) follows from the lemma above. For (3), note that $E^1(\mathbb{Q}_p)$ contains only points with nonzero y-coordinate, and the only such point with $v_p(x/y) = \infty$ is O; every other other point $(x : y : z) \in E^1(\mathbb{Q}_p)$ lies in $E^n(\mathbb{Q}_p) - E^{n+1}(\mathbb{Q}_p)$, where $n = v_p(x/y)$. □

Remark 24.15. Theorem 24.14 can be extended to all primes p by replacing $\overline{E}(\mathbb{F}_p)$ in (1) with the set S of non-singular points on the reduction of E modulo p. As in the proof of Lemma 24.12, one can show that S always contains O and is closed under the group operation, but there are now three different group structures that can arise:

1. A cyclic group of order p isomorphic to the additive group of $\mathbb{F}_p$; in this case E is said to as *additive reduction* at p.

2. A cyclic group of order $p-1$ isomorphic to the multiplicative group of $\mathbb{F}_p$; in this case E is said to have *split multiplicative reduction* at p.

3. A cyclic group of order $p+1$ isomorphic to the subgroup of the multiplicative group of a quadratic extension of $\mathbb{F}_p$ consisting of elements of norm one; in this case E is said to have *non-split multiplicative reduction* at p.

Parts (2) and (3) of the theorem remain true for all primes p (as we will now assume).

Corollary 24.16. *Suppose $P = (x : y : 1)$ is an affine point in $E^0(\mathbb{Q}_p)$ with finite order prime to p. Then $x, y \in \mathbb{Z}_p$.*

Proof. Suppose not. Then both x and y must have negative p-adic valuations in order to satisfiy $y^2 = x^3 + a_4x + a_6$ with $a_4, a_6 \in \mathbb{Z}_p$, and we must have $2v_p(x) = 3v_p(y)$, so $v_p(x/y) > 0$. Let $n = v_p(x/y)$. Then $P \in E^n(\mathbb{Q}_p) - E^{n+1}(\mathbb{Q}_p)$, and the image of P in

$$E^n(\mathbb{Q}_p)/E^{n+1}(\mathbb{Q}_p) \simeq \mathbb{Z}/p\mathbb{Z}$$

is not zero, hence it has order p. The order m of P is prime to p, so the image of mP in $E^n(\mathbb{Q}_p)/E^{n+1}(\mathbb{Q}_p)$ is also nonzero. Thus $mP \notin E^{n+1}(\mathbb{Q}_p)$, but this is a contradiction, because $mP = O \in E^{n+1}(\mathbb{Q}_p)$. □

Lemma 24.17. *Suppose P_1, P_2, P_3 are colinear points in $E^n(\mathbb{Q}_p)$, for some $n > 0$, with $P_i = (x_i : 1 : z_i)$. Then $v_p(x_1 + x_2 + x_3) \geq 5n$.*

Proof. We have already seen that for $P_i \in E^n(\mathbb{Q}_p)$ we have $x_i \in p^n\mathbb{Z}_p$ and $z_i \in p^{3n}\mathbb{Z}_p$. Fixing $y = 1$, if $P_1 \neq P_2$ then the equation of the line through P_1 and P_2 in the x-z plane has the form $z = \alpha x + \beta$ with $\alpha = (z_2 - z_1)/(x_2 - x_1)$. Using the curve equation $z = x^3 + a_3xz^2 + z^3$ (with $y = 1$), we can rewrite α as

$$\begin{aligned}
\alpha &= \frac{z_2 - z_1}{x_2 - x_1} \\
&= \frac{z_2 - z_1}{x_2 - x_1} \cdot \frac{1 - a_4z_2^2 - a_6(z_2^2 + z_1z_2 + z_1^2)}{1 - a_4z_2^2 - a_6(z_2^2 + z_1z_2 + z_1^2)} \\
&= \frac{(z_2 - a_6z_2^3) - (z_1 - a_4x_1z_1^2 - a_6z_1^3) - a_4x_1z_2^2}{(x_1 - x_2)(1 - a_4z_2^2 - a_6(z_2^2 + z_1z_2 + z_1^2))} \\
&= \frac{(x_2^3 + a_4x_2z_2^2) - x_1^3 - a_4x_1z_2^2}{(x_1 - x_2)(1 - a_4z_2^2 - a_6(z_2^2 + z_1z_2 + z_1^2))} \\
&= \frac{(x_2 - x_1)(x_2^2 + x_1x_2 + x_1^2 + a_4z_2^2)}{(x_1 - x_2)(1 - a_4z_2^2 - a_6(z_2^2 + z_1z_2 + z_1^2))} \\
&= \frac{x_2^2 + x_1x_2 + x_1^2 + a_4z_2^2}{1 - a_4z_2^2 - a_6(z_2^2 + z_1z_2 + z_1^2)}.
\end{aligned}$$

The key point is that the denominator of α is then a p-adic unit. It follows that $\alpha \in p^{2n}\mathbb{Z}_p$, and then $\beta = z_1 - \alpha x_1 \in p^{3n}\mathbb{Z}_p$. Substituting $z = \alpha x + \beta$ into the curve equation gives

$$\alpha x + \beta = x^3 + a_4x(\alpha x + \beta)^2 + b(\alpha x + \beta)^3.$$

We know that x_1, x_2, x_3 are the three roots of the cubic defined by the equation above, thus $x_1 + x_2 + x_3$ is equal to the coefficient of the quadratic term divided by the coefficient of the cubic term. Therefore

$$x_1 + x_2 + x_3 = \frac{2a_4\alpha\beta + 3a_6\alpha^2\beta}{1 + a_4\alpha^2 + a_6\beta^3} \in p^{5n}\mathbb{Z}_p.$$

The case $P_1 = P_2$ is similar. □

Corollary 24.18. *The group $E^1(\mathbb{Q}_p)$ is torsion-free.*

Proof. By the previous corollary we only need to consider the case of a point of order p. So suppose $pP = 0$ for some $P \in E^n(\mathbb{Q}_p) - E^{n+1}(\mathbb{Q}_p)$. Consider the map

$$E^n(\mathbb{Q}_p) \to p^n\mathbb{Z}_p/p^{5n}\mathbb{Z}_p$$

that sends $P := (x : 1 : z)$ to the reduction of x in $p^n\mathbb{Z}_p/p^{5n}\mathbb{Z}_p$. By the lemma above, this is a homomorphism of abelian groups, so it sends pP to the reduction of $px \in p^{n+1}\mathbb{Z}_p - p^{n+2}\mathbb{Z}_p$ modulo $p^{5n}\mathbb{Z}_p$, which is not zero, a contradiction. □

Corollary 24.19. *Let $E/\mathbb{Q}$ be an elliptic curve and let p be a prime of good reduction. The torsion subgroup of $E(\mathbb{Q})$ injects into $\overline{E}(\mathbb{F}_p)$. in particular, the torsion subgroup is finite.*

Proof. The group $E(\mathbb{Q})$ is isomorphic to a subgroup of $E^0(\mathbb{Q}_p)$ and $\overline{E}(\mathbb{F}_p) = E^0(\mathbb{Q}_p)/E^1(\mathbb{Q}_p)$. But $E^1(\mathbb{Q}_p)$ is torison free, so the torsion subgroup of $E(\mathbb{Q})$ injects into $\overline{E}(\mathbb{F}_p)$. □

Now that we know that each elliptic curve over $\mathbb{Q}$ has a finite number of rational torsion points, one might ask whether there is a *uniform* upper bound that applies to every elliptic curve over $\mathbb{Q}$. It's not *a priori* clear that this should be the case; one might suppose that by varying the elliptic curve we could get an arbitrarily large number of rational torsion points. But this is not the case; an elliptic curve over $\mathbb{Q}$ can have at most 16 rational points of finite order. This follows from a celebrated theorem of Mazur that tells us exactly which rational torsion subgroups can (and do) arise for elliptic curves defined over $\mathbb{Q}$.

Theorem 24.20 (Mazur)**.** *Let $E/\mathbb{Q}$ be an elliptic curve. The torsion subgroup of $E(\mathbb{Q})$ is isomorphic to one of the fifteen subgroups listed below:*

$$\mathbb{Z}/n\mathbb{Z} \quad (n = 1, 2, 3, \ldots, 9, 10, 12), \qquad\qquad \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2n\mathbb{Z} \quad (n = 1, 2, 3, 4).$$

The proof of this theorem is well beyond the scope of this course.[2] However, as a further refinement of our results above, we can prove the Nagell-Lutz Theorem.

Theorem 24.21 (Nagell-Lutz)**.** *Let $P = (x_0 : y_0 : 1)$ be an affine point of finite order on the elliptic curve $y^2 = x^3 + a_4x + a_6$ over $\mathbb{Q}$, with $a_4, a_6 \in Z$. Then $x_0, y_0 \in \mathbb{Z}$, and if $y_0 \neq 0$ then y_0^2 divides $4a_4^3 + 27a_6^2$.*

Proof. For any prime p, if $v_p(x_0) < 0$ then $2v_p(y_0) = 3v_p(x_0)$ and $v_p(x_0/y_0) > 0$. It follows that $P \in E_1(\mathbb{Q})$, but then P cannot be a torsion point. So $v_p(x) \geq 0$ for all primes p. Thus x_0 is an integer, and so is $y_0^2 = x_0^3 + a_4x_0 + a_6$, and therefore y.

If P has order 2 then $y_0 = 0$; otherwise, the x-coordinate of $2P$ is an integer equal to $\lambda^2 - 2x_0$, where $\lambda = (3x_0^2 + a_4)/(2y_0)$ is the slope of the tangent at P. Thus $4y_0^2$ and therefore y_0^2 divides $\lambda^2 = (3x_0^2 + a_4)^2$, as well as $x_0^3 + a_4x_0 + a_6$. We now note that

$$\begin{aligned}(3x_0^2 + 4a_4)(3x_0^2 + a_4)^2 &= 27x_0^6 + 54a_4x_0^4 + 27a_4^2x_0^2 + 4a_4^3\\(3x_0^2 + 4a_4)(3x_0^2 + a_4)^2 &= 27(x_0^3 + a_4x_0)^2 + 4a_4^3\\0 &\equiv 27a_6^2 + 4a_4^3 \bmod y_0^2,\end{aligned}$$

since $(x_0^3 + a_4x_0) \equiv -a_6 \bmod y_0^2$, thus y_0 divides $4a_4^3 + 27a_6^2$. □

The Nagell-Lutz theorem gives an effective method for enumerating all of the torsion points in $E(\mathbb{Q})$ that is quite practical when the coefficients a_4 and a_6 are small. By factoring $D = 4a_4^3 + 27a_6^2$, one can determine all the squares y_0^2 that divide D. By considering each of these, along with $y_0 = 0$, one then checks whether there exists an integral solution x_0 to $y_0^2 = x_0^3 + ax_0 + a_6$ (note that such an x_0 must be a divisor of $a_6 - y_0^2$).

This yields a list of candidate torsion points $P = (x_0 : y_0 : 1)$ that are all points in $E(\mathbb{Q})$, but do not necessarily all have finite order. To determine which do, one computes multiples nP for increasing values of n (by adding the point P at each step, using the group law on E), checking at each step whether $nP = O$. If at any stage it is found that the affine coordinates of nP are not integers then nP, and therefore P, cannot be a torsion point, and in any case we know from Mazur's theorem that if $nP \neq O$ for any $n \leq 12$ then P is not a torsion point; alternatively, we also know that n must divide $\#\overline{E}(\mathbb{F}_p)$, where p is the least prime that does not divide $\Delta(E)$.

However, this method is not practical in general, both because it requires us to factor D, and because D might have a very large number of square divisors (if D is, say, the product

[2] There was recently a graduate seminar at Harvard devoted entirely to the proof of Mazur's theorem; see http://www.math.harvard.edu/~chaoli/MazurTorsionSeminar.html for notes and references.

of the squares of the first 100 primes, then we have 2^{100} values of y_0 to consider). But Cororllary 24.19 gives us a much more efficient alternative that can be implemented to run in quasi-linear time (roughly proportional to the number of bits it takes to represent a_4 and a_6 on a computer).

We first determine the least odd prime p that does not divide D; we don't need to factor D to do this and we will always have p bounded by $O(\log D) = O(\log \max(|a_4|, |a_6|)$. We then exhaustively compute the set $\overline{E}(\mathbb{F}_p)$, which clearly has cardinality at most $2p$ (in fact, at most $p+1+2\sqrt{p}$). For each integer $m > 1$ there is an *m-division polynomial* $f_m \in \mathbb{Z}[x]$ with the property that $P = (x_0, y_0) \in E(\overline{\mathbb{Q}})$ satisfies $mP = 0$ if and only if $f_m(x_0) = 0$. The polynomials f_m can be explicitly computed using formulas for the group law on E and have integer coefficients that depend on the integer coefficients of E and degree bounded by m^2. If $\overline{P} = (\overline{x}_0, \overline{y}_0)$ is a point of order m in $\overline{E}(\mathbb{F}_p)$ then $f(\overline{x}_0) \equiv 0 \bmod p$, and we can use Hensel's lemma to efficiently "lift" the root $\overline{x}_0$ of f_m modulo p to a root x_0 of f_m modulo p^n, where n is chosen so that p^n is more than twice as large as the absolute value of the x-coordinate of any torsion point in $E(\mathbb{Q})$; the fact that y_0^2 must divide D gives us an upper bound on both y_0 and x_0. We choose a representative $x_0 \in \mathbb{Z}$ with $|x_0| < p^n/2$ and check whether $f_m(x_0) = 0$; if so then $x_0^3 + a_4x + a_6$ must be the square of an integer $y_0 \equiv \overline{y}_0 \bmod p$ (which we can also compute using Hensel lifting) and $(x_0, y_0) \in E(\mathbb{Q})$ is a torsion point. Repeating this process for each $P \in \overline{E}(\mathbb{F}_p)$ yields the torsion subgroup of $E(\mathbb{Q})$. But we know from Mazur's theorem that we only need to consider the points $\overline{P} \in \overline{E}(\mathbb{F}_p)$ of order $m \leq 12$, which means there at only $O(1)$ points to consider; here we are using the fact that $\overline{E}(\mathbb{F}_p)$ is generated by at most two elements, which we will not prove here. Provided we use fast algorithms for integer multiplication in our implementation of Hensel lifting, this yields a quasi-linear running time.

References

[1] J. H. Silverman, *The arithmetic of elliptic curves*, Springer, 2009.

18.782 Introduction to Arithmetic Geometry **Fall 2013**
Lecture #25 **12/05/2013**

25.1 Overview of Mordell's theorem

In the last lecture we proved that the torsion subgroup of the rational points on an elliptic curve $E/\mathbb{Q}$ is finite. In this lecture we will prove a special case of Mordell's theorem, which states that $E(\mathbb{Q})$ is finitely generated. By the structure theorem for finitely generated abelian groups, this implies

$$E(\mathbb{Q}) \simeq \mathbb{Z}^r \oplus T,$$

where $\mathbb{Z}^r$ is a free abelian group of rank r, and T is the (necessarily finite) torsion subgroup.[1] Thus Mordell's theorem provides an alternative proof that T is finite, but unlike our earlier proof, it does not provide an explicit method for computing T. Indeed, Mordell's theorem is notably *ineffective*; it does not give us a way to compute a set of generators for $E(\mathbb{Q})$, or even to determine the rank r. It is a major open question as to whether there exists an algorithm to compute r; it is also not known whether r can be uniformly bounded.[2]

Mordell's theorem was generalized to number fields (finite extensions of $\mathbb{Q}$) and to abelian varieties (recall that elliptic curves are abelian varieties of dimension one) by André Weil and is often called the Mordell-Weil theorem. All known proofs of Mordell's theorem (and its generalizations) essentially amount to two proving two things:

(a) $E(\mathbb{Q})/2E(\mathbb{Q})$ is a finite group.

(b) For any fixed $Q \in E(\mathbb{Q})$, the *height* of $2P + Q$ is greater than the height of P for all but finitely many P.

We note that there is nothing special about 2 here, any integer $n > 1$ works.

We will explain what (b) means in a moment, but let us first note that we really do need some sort of (b); it is not enough to just prove (a). To see why, consider the additive abelian group $\mathbb{Q}$. the quotient $\mathbb{Q}/2\mathbb{Q}$ is certainly finite (it is the trivial group), but $\mathbb{Q}$ is not finitely generated. To see this, note that for any finite $S \subseteq \mathbb{Q}$, we can pick a prime p such that under the canonical embedding $\mathbb{Q} \subseteq \mathbb{Q}_p$ we have $S \subseteq \mathbb{Z}_p$, and therefore $\langle S \rangle \subseteq \mathbb{Z}_p$, but we never have $\mathbb{Q} \not\subseteq \mathbb{Z}_p$.

The *height* of a projective point $P = (x : y : z)$ with $x, y, z \in \mathbb{Z}$ sharing no common factor is defined as

$$H(P) := \max(|x|, |y|, |z|),$$

where $|\ |$ is the usual archimedean absolute value on $\mathbb{Q}$. The height $H(P)$ is a positive integer that is independent of the representation of the representation of P, and for any bound B, the set

$$\{P \in E(\mathbb{Q}) : H(P) \le B\}$$

is finite, since it cannot possibly have more than $(2B + 1)^3$ elements. We will actually use a slightly more precise notion of height, the *canonical height*, which we will define later.

Now let us suppose that we have proved (a) and (b), and see why this implies that $E(\mathbb{Q})$ is finitely generated. Since $E(\mathbb{Q})/2E(\mathbb{Q})$ is finite, for any sufficiently large B the finite set $S = \{P \in E(\mathbb{Q}) : H(P) \le B\}$ must contain a set of representatives for $E(\mathbb{Q})/2E(\mathbb{Q})$, and

[1] Any finitely generated *abelian* torsion group must be finite; this does not hold for nonabelian groups.

[2] Most number theorists think not, but there are some notable dissenters.

Andrew V. Sutherland

we can pick B so that (b) holds for all $Q \in S$ and $P \notin S$. If S does not generate $E(\mathbb{Q})$, then there is a point $P_0 \in E(\mathbb{Q}) - \langle S\rangle$ of minimal height $H(P_0)$. Since S contains a set of representatives for $E(\mathbb{Q})/2E(\mathbb{Q})$, we can write P_0 in the form

$$P_0 = 2P + Q,$$

for some $Q \in S$ and $P \in E(\mathbb{Q})$. Since $P_0 \notin \langle S\rangle$, we must have $P \notin \langle S\rangle$, but (b) implies $H(P) < H(P_0)$, contradicting the minimality of $H(P_0)$. So the set $E(\mathbb{Q}) - \langle S\rangle$ must be empty and S is a finite set of generators for $E(\mathbb{Q})$.

We should note that this argument does not yield an algorithm to compute S because we do not have an effective bound on B (we know B exists, but not how big it is).

25.2 Elliptic curves with a rational point of order 2

In order to simplify the presentation, we will restrict our attention to elliptic curves $E/\mathbb{Q}$ that have a rational point of order 2 (to prove the general case one can work over a cubic extension of $\mathbb{Q}$ for which this is true). In short Weierstrass form any point of order 2 is an affine point of the form $(x_0, 0)$. After replacing x with $x + x_0$ we obtain an equation for E of the form

$$E\colon y^2 = x(x^2 + ax + b),$$

on which $P = (0, 0)$ is a point of order two. Since E is not singular, the cubic on the RHS has no repeated roots, which implies

$$b \neq 0, \qquad a^2 - 4b \neq 0.$$

The algebraic equations for the group law on curves of this form are slightly different than for curves in short Weierstrass form; the formula for the inverse of a point is the same, we simply negate the y-coordinate, but the formulas for addition and doubling are slightly different. To add two affine points $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$ with $x_1 \neq x_2$, as in Lecture 23 we consider the line L through P_1 and P_2 with equation

$$L\colon (y - y_1) = \lambda(x - x_1),$$

where $\lambda = (y_2 - y_1)/(x_2 - x_1)$. Solving for y and plugging into equation for E, we have

$$\begin{aligned} \lambda^2 x^2 &= x(x^2 + ax + b) \\ 0 &= x^3 + (a - \lambda^2)x^2 + \cdots \end{aligned}$$

The x-coordinate x_3 of the third point in the intersection $L \cap E$ is a root of the cubic on the RHS, as are x_1 and x_2, and the sum $x_1 + x_2 + x_3$ must be equal to the negation of the quadratic coefficient. Thus

$$\begin{aligned} x_3 &= \lambda^2 - a - x_1 - x_2, \\ y_3 &= \lambda(x_1 - x_3) - y_1, \end{aligned}$$

where we computed y_3 by plugging x_3 into the equation for L and negating the result. The doubling formula for $P_1 = P_2$ is the same, except now $\lambda = (3x^2 + 2ax + b)/(2y)$.

25.3 2-isogenies

In order to prove that $E(\mathbb{Q})/2E(\mathbb{Q})$ is finite, we need to understand the image of the multiplication-by-2 map $[2]$. We could use the doubling formula derived above to do this, but it turns out to be simpler to decompose $[2]$ as a composition of two isogenies

$$[2] = \hat{\varphi} \circ \varphi,$$

where $\varphi\colon E \to E'$ and $\hat{\varphi}\colon E' \to E$ for some elliptic curve E' that we will determine. The kernel of φ will be $\{O, P\}$, where $P = (0,0)$ is our rational point of order 2. Similarly, the kernel of $\hat{\varphi}$ will be $\{O', P'\}$, where O' is the distinguished point on E' and P' is a rational point of order 2 on E'.

Recall from Lecture 24 that for any isogeny $\varphi\colon E \to E'$ we have an injective map

$$\ker\varphi \to \operatorname{Aut}(\overline{\mathbb{Q}}(E)/\varphi^*(\overline{\mathbb{Q}}(E')))$$

defined by $P \mapsto \tau_P^*$, where τ_P is the translation-by-P morphism. In our present situation there is only one non-trivial point in the kernel of φ, the point $P = (0,0)$, and it is rational, so we can work over $\mathbb{Q}$. We can determine both E' and the morphism φ by computing $\varphi^*(\mathbb{Q}(E'))$ as the fixed field of the automorphism $\tau_P^*\colon \mathbb{Q}(E) \to \mathbb{Q}(E)$.

Remark 25.1. This strategy applies in general to any separable isogeny with a cyclic kernel (a *cyclic isogeny*), all we need is a point P that generates the kernel.

For an affine point $Q = (x,y)$ not equal to $P = (0,0)$ the x-coordinate of $\tau_P(Q) = P+Q$ is $\lambda^2 - a - x$, where $\lambda = y/x$ is the slope of line throught P and Q. Using the curve equation for E, we can simplify this to

$$\lambda^2 - a - x = \frac{y^2 - ax^2 - x^3}{x^2} = \frac{bx}{x^2} = \frac{b}{x}.$$

The y-coordinate of $\tau_P(Q)$ is then $\lambda(0 - b/x) - 0 = -by/x^2$. Thus for $Q \not\in \{O, P\}$ the map τ_P is given by

$$(x,y) \mapsto (b/x, -by/x^2).$$

To compute the fixed field of τ_P^*, note that if we regard the slope $\lambda = y/x$ as a function in $\mathbb{Q}(E)$, then composition with τ_P merely changes its sign. Thus

$$\tau_P^*(\lambda^2) = \left(\frac{-by/x^2}{b/x}\right)^2 = \left(\frac{-y}{x}\right)^2 = \lambda^2.$$

We also note that the point $Q + \tau_P(Q)$ is fixed by τ_P, hence the sum of the y-coordinates of Q and $\tau_P(Q)$ is fixed by τ_P (when represented as affine points $(x:y:1)$). Thus

$$\tau_P^*(y - by/x^2) = \tau_P^*\left(\frac{x^2y - by}{x^2}\right) = \frac{(b/x)^2(-by/x^2) - b(-by/x^2)}{(b/x)^2} = y - by/x^2.$$

Note that $\lambda^2 = y^2/x^2 = x(x^2 + ax + b)/x^2 = x + a + b/x$, so let us define

$$X = x + a + b/x \qquad \text{and} \qquad Y = y(1 - b/x^2)$$

Then $\mathbb{Q}(X,Y)$ is a subfield of $E(\mathbb{Q}) = \mathbb{Q}(x,y)$ fixed by τ_P^*, hence a subfield of $\varphi^*(\mathbb{Q}(E'))$, and we claim that it is a subfield of index 2. To see this, note that

$$x = (X + Y\sqrt{X} - a)/2 \qquad \text{and} \qquad y = x\sqrt{X},$$

thus $[\mathbb{Q}(E):\mathbb{Q}(X,Y)]\le 2$ and $[\mathbb{Q}(E):\mathbb{Q}(X,Y)]\ne 1$ because $\mathbb{Q}(E)$ contains $x/y=\sqrt{X}$ and $\mathbb{Q}(X,Y)$ does not. We also know that $[\mathbb{Q}(E):\varphi^*(\mathbb{Q}(E'))]\ge 2$, since $\ker\varphi\subseteq\mathbb{Q}(E)$ has order 2 and injects into $\operatorname{Aut}(\mathbb{Q}(E)/\varphi^*(\mathbb{Q}(E)))$, therefore $\varphi^*(\mathbb{Q}(E'))=\mathbb{Q}(X,Y)$.

Since φ^* is a field embedding, we have $\mathbb{Q}(E')\simeq\mathbb{Q}(X,Y)$. We now know the function field of E'; to compute an equation for E' we just need a relation between X and Y.

$$\begin{aligned}Y^2&=y^2(1-b/x^2)^2\\&=x(x^2+ax+b)(1-2b/x^2+b^2/x^4)\\&=X(x^2-2b+b^2/x^2)\\&=X\big((x+b/x)^2-4b\big)\\&=X\big((X-a)^2-4b\big)\\&=X(X^2-2aX+a^2-4b).\end{aligned}$$

Let us now define $A=-2a$ and $B=a^2-4b$. Then the equation

$$Y^2=X(X^2+AX+B)$$

has the same form as that of E, and since $B=a^2-4b\ne 0$ and $A^2-4B=16b\ne 0$, it defines an elliptic curve E' with distinguished point $O'=(0:1:0)$, and the affine point $P'=(0,0)$ has order 2. The 2-isogeny $\varphi\colon E\to E'$ sends O to O' and each affine point (x,y) on E to $(X,Y)=(x+a+b/x,\ y(1-b/x^2))$ on E'.

Since E' has the same form has E, we can repeat the process above to compute the 2-isogeny $\hat\varphi\colon E'\to E$ that sends O' to O and (X,Y) to $(X+A+B/X, Y(1-B/X^2))$. One can then verify that

$$[2]=\hat\varphi\circ\varphi,$$

by composing $\hat\varphi$ and φ and comparing the result to the doubling formula on E.

But we can see this more directly by noting that $\ker(\hat\varphi\circ\varphi)=E[2]$ and

$$\deg(\hat\varphi\circ\varphi)=\deg\hat\varphi\deg\varphi=2\cdot 2=4=\#E[2]=\#\ker(\hat\varphi\circ\varphi).$$

Thus the injective homomorphism $E[2]\to\operatorname{Aut}(\overline{\mathbb{Q}}(E)/(\hat\varphi\circ\varphi)^*(\overline{\mathbb{Q}}(E)))$ is an isomorphism, and the same holds for $\operatorname{Aut}(\overline{\mathbb{Q}}(E)/[2]^*\overline{\mathbb{Q}}(E))$. Since we are in characteristic zero, both extensions are separable, and it follows from Galois theory that there is a unique subfield of $\overline{\mathbb{Q}}(E)$ fixed by the automorphism group $\{\tau_P^*:P\in E[2]\}$. Thus the function field embeddings $(\hat\varphi\circ\varphi)^*$ and $[2]^*$ are equal, and the corresponding morphisms must be equal (by the functorial equivalence of smooth projective curves and their function fields).

Remark 25.2. The construction and argument above applies quite generally. Given any finite subgroup H of $E(\bar k)$ there is a unique elliptic curve E' and separable isogeny $E\to E'$ with H as its kernel; see [2, Prop. III.4.12].

25.4 The weak Mordell-Weil theorem

We are now ready to prove that $E(\mathbb{Q})/2E(\mathbb{Q})$ is finite (in the case that $E(\mathbb{Q})$ has a rational point of order 2). This is a special case of what is known as the weak Mordell-Weil theorem, which says that $E(k)/nE(k)$ is finite, for any positive integer n and any number field k. Our strategy is to prove that $E(\mathbb{Q})/\varphi(E(\mathbb{Q}))$ is finite, where $\varphi\colon E\to E'$ is the 2-isogeny from the previous section. This will also show that $E'(\mathbb{Q})/\hat\varphi(E(\mathbb{Q}))$ is finite, and it will follow that $E/2E(\mathbb{Q})$ is finite.

We begin by characterizing the image of φ in $E'(\mathbb{Q})$.

Lemma 25.3. *An affine point $(X, Y) \in E'(\mathbb{Q})$ lies in the image of φ if and only if either $X \in \mathbb{Q}^{\times 2}$, or $X = 0$ and $a^2 - 4b \in \mathbb{Q}^{\times 2}$.*

Proof. Suppose $(X, Y) = \varphi(x, y)$. T If $X \neq 0$ then $X = (y/x)^2 \in \mathbb{Q}^{\times 2}$. If $X = 0$ then $x(x^2 + ax + b) = 0$, and $x \neq 0$ (since $\varphi(0, 0) = O'$), so $x^2 + ax + b = 0$ has a rational solution, which implies $a^2 - 4b \in \mathbb{Q}^{\times 2}$.

Conversely, if $X \in \mathbb{Q}^{\times 2}$ then $x = (X + Y\sqrt{X} - a)/2$ and $y = x\sqrt{X}$ gives a point $(x, y) \in E(\mathbb{Q})$ for which $\varphi(x, y) = (X, Y)$, and if $X = 0$ and $a^2 - 4b \in \mathbb{Q}^{\times 2}$, then $x^2 + ax + b$ has a nonzero rational root x for which $\varphi(x, 0) = (0, 0) = (X, Y)$. □

Now let us define the map $\pi\colon E'(\mathbb{Q}) \to \mathbb{Q}^\times/\mathbb{Q}^{\times 2}$ by

$$(X, Y) \mapsto \begin{cases} X & \text{if } X \neq 0, \\ a^2 - 4b & \text{if } X = 0, \end{cases}$$

and let $\pi(O') = 1$.

Lemma 25.4. *The map $\pi\colon E'(\mathbb{Q}) \to \mathbb{Q}^\times/\mathbb{Q}^{\times 2}$ is a group homomorphism.*

Proof. By definition, $\pi(O') = 1$, so π preserves the identity element and behaves correctly on sums involving O'. and since $\pi(P) = \pi(-P)$ and the square classes of X and $1/X$ are the same, π preserves inverses. We just need to verify $\pi(P + Q) = \pi(P)\pi(Q)$ for affine points P, Q that are not inverses.

So let P and Q be affine points whose sum is an affine point R, let $Y = \ell X + m$ be the line L containing P and Q (the line L is not vertical because $P + Q = R \neq O'$). Plugging the equation for Y given by L into the equation for E' gives

$$\begin{aligned} (\ell X + m)^2 &= X(X^2 + AX + B) \\ 0 &= X^3 + (A - \ell^2)x^2 + (B - \ell m)x - m^2. \end{aligned}$$

The X-coordinates X_1, X_2, X_3 of P, Q, R are all roots of the cubic on the RHS, hence their product is equal to m^2, the negation of the constant term. Thus $X_1X_2X_3$ is a square, which means that $\pi(P)\pi(Q)\pi(P + Q) = 1$, and therefore $\pi(P)\pi(Q) = 1/\pi(P + Q) = \pi(P + Q)$, since $\pi(P + Q)$ and $1/\pi(P + Q)$ are in the same square-class of $\mathbb{Q}^\times$. □

Lemma 25.5. *The image of $\pi\colon E'(\mathbb{Q}) \to \mathbb{Q}^\times/\mathbb{Q}^{\times 2}$ is finite.*

Proof. Let (X, Y) be an affine point in $E'(\mathbb{Q})$ with $X \neq 0$, and let $r \in \mathbb{Z}$ be a square-free integer representative of the square-class $\pi(X, Y)$. We will show that r must divide B, which clearly implies that $\operatorname{im} \pi$ is finite. The equation $Y^2 = X(X + aX + B)$ for E' implies that X and $X + aX + B$ lie in the same square-class, thus

$$\begin{aligned} X^2 + AX + B &= rs^2 \\ X &= rt^2, \end{aligned}$$

for some $s, t \in \mathbb{Q}^\times$. Let us write $t = \ell/m$ with $\ell, m \in \mathbb{Z}$ relatively prime. Plugging $X = rt^2$ into the first equation gives

$$\begin{aligned} r^2t^4 + Art^2 + B &= rs^2 \\ r^2\ell^4/m^4 + Ar\ell^2/m^2 + B &= rs^2 \\ r^2\ell^4 + Ar\ell^2m^2 + Bm^4 &= rm^4s^2, \end{aligned}$$

and since the LHS is an integer, so is the RHS. Let p be any prime dividing r. Then p must divide Bm^4, since it divides every other term. If p divides m then p^3 must divide $r^2\ell^4$, since it divides every other term, but then p divides ℓ, since r is squarefree, which is impossible because ℓ and m are relatively prime. So p does not divide m and therefore must divide B. This holds for every prime divisor of the squarefree integer r, so r divides B as claimed. □

Corollary 25.6. *$E'(\mathbb{Q})/\varphi(E(\mathbb{Q}))$ and $E(\mathbb{Q})/\hat{\varphi}(E(\mathbb{Q}))$ are finite.*

Proof. Lemma 25.3 implies that $\ker\pi = \varphi(E(\mathbb{Q}))$, thus $E'(\mathbb{Q})/\varphi(E(\mathbb{Q})) \simeq \operatorname{im}\pi$ is finite, and this remains true if we replace E with E' and φ with $\hat{\varphi}$. □

Corollary 25.7. *$E(\mathbb{Q})/2E(\mathbb{Q})$ is finite.*

Proof. The fact that $[2] = \hat{\varphi}\circ\varphi$ implies that each $\hat{\varphi}(E'(\mathbb{Q}))$-coset in $E(\mathbb{Q})$ can be partitioned into $2E(\mathbb{Q})$-cosets. Two points P and Q in the same $\hat{\varphi}(E'(\mathbb{Q}))$-coset lie in the same $2E(\mathbb{Q})$-coset if and only if $(P-Q) \in 2E(\mathbb{Q}) = (\hat{\varphi}\circ\varphi)(E(\mathbb{Q}))$, equivalently, $\hat{\varphi}^{-1}(P-Q) \in \varphi(E(\mathbb{Q})$. Thus the number of $2E(\mathbb{Q})$-cosets in each $\hat{\varphi}(E'(\mathbb{Q}))$-coset is precisely $E'(\mathbb{Q})/\varphi(E(\mathbb{Q}))$, thus

$$\#E(\mathbb{Q})/2E(\mathbb{Q}) = \#E(\mathbb{Q})/\hat{\varphi}(E'(\mathbb{Q}))\ \#E'(\mathbb{Q})/\varphi(E(\mathbb{Q}))$$

is finite. □

Remark 25.8. The only place in our work above where we really used the fact that we are working over $\mathbb{Q}$, as opposed to a general number field, is in the proof of Lemma 25.5. Specifically, we used the fact that the ring of integers $\mathbb{Z}$ of $\mathbb{Q}$ is a UFD, and that its unit group $\mathbb{Z}^\times$ is finite. Neither is true of the ring of integers $\mathcal{O}_k$ of a number field k, in general, but there are analogous facts that one can use; specifically, $\mathcal{O}_k$ is a Dedekind domain, hence ideals can be unique factored into prime ideals, the class number of $\mathcal{O}_k$ is finite, and $\mathcal{O}_k^\times$ is finitely generated. We also assumed that E has a rational point of order 2, but after a base extension to a number field we can assume this without loss of generality.

25.5 Height functions

Let k be any number field. Recall from Lecture 6 that (up to equivalence) the absolute values of k consist of non-archimedean absolute values, one for each prime ideal $\mathfrak{p}$ of the ring of integers $\mathcal{O}_k$ (these are the *finite places* of k), and archimedean absolute values, one for each embedding of k into $\mathbb{R}$ and one for each conjugate pair of embeddings of k into $\mathbb{C}$ (these are the *infinite places* of k). Let $\mathcal{P}_k$ denote the set of (finite and infinite) places of k.

For each place $p \in \mathcal{P}_k$ we want to normalize the associated absolute value $|\ |_p$ so that

(a) The product formula $\prod_{p\in\mathcal{P}_k}|x|_p = 1$ holds for all $x \in k^\times$.

(b) For any number field $k' \subseteq k$ and any place p of k' we have $\prod_{q|p}|x|_q = |x|_p$, where $q|p$ means that the restriction of $|\ |_q$ to k' is equivalent to $|\ |_p$.

Both requirements are satisfied by using the standard normalization for $\mathbb{Q}$, with

$$|x|_p = p^{-v_p(x)}$$

for $p < \infty$ and $|x|_\infty = |x|$, and then for each $q \in \mathcal{P}_k$ with $q|p$ defining

$$|x|_q = |N_{k_q/\mathbb{Q}_p}(x)|_p^{1/[k:\mathbb{Q}]},$$

where k_q and $\mathbb{Q}_p$ denote the completions of k at q and $\mathbb{Q}$ at p, respectively.[3]

Definition 25.9. The (absolute) *height* of a projective point $P = (x_0 : \cdots : x_n) \in \mathbb{P}^n(\overline{\mathbb{Q}})$ is

$$H(P) := \prod_{p \in \mathcal{P}_k} \max_i |x_i|_p,$$

where $k = \mathbb{Q}(x_0, \ldots, x_n)$. For any $\lambda \in \overline{\mathbb{Q}}^\times$, if we let $k = \mathbb{Q}(x_0, \ldots, x_n\lambda)$, then

$$\prod_{p \in \mathcal{P}_k} \max_i |\lambda x_i|_p = \prod_{p \in \mathcal{P}_k} \max_i(|\lambda|_p |x_i|_p) = \prod_{p \in \mathcal{P}_k} |\lambda|_p \prod_p \max_i |x_i| = \prod_{p \in \mathcal{P}_k} \max_i |x_i|,$$

thus $H(P)$ is well defined (it does not depend on a particular choice of $x_0, \ldots, x_n$).

For $k = \mathbb{Q}$ we can write $P = (x_0 : \cdots : x_n)$ with the $x_i \in \mathbb{Z}$ having no common factor. Then $\max |x_i|_p = 1$ for $p < \infty$ and $H(P) = \max_i |x_i|_\infty$; this agrees with the definition we gave earlier.

Lemma 25.10. *For all $P = (x_0 : \cdots : x_n) \in \mathbb{P}^n(\overline{\mathbb{Q}})$ we have $H(P) \geq 1$.*

Proof. Pick a nonzero x_j and let $k = \mathbb{Q}(x_0, \ldots, x_n)$. Then

$$H(P) = \prod_{p \in \mathcal{P}_k} \max_i |x_i|_p \;\geq\; \prod_{p \in \mathcal{P}_k} |x_j|_p = 1. \qquad \square$$

Definition 25.11. The *logarithmic height* of $P \in \mathbb{P}^n(\overline{\mathbb{Q}})$ is the nonnegative real number

$$h(P) := \log H(P).$$

We now consider how the height of a point changes when we apply a morphism to it. We will show that there for any fixed morphism $\phi \colon \mathbb{P}^m \to \mathbb{P}^n$ there are constants c and d (depending on ϕ) such that for any point $P \in \mathbb{P}^m(\overline{\mathbb{Q}})$ we have

$$dh(P) - c \;\leq\; h(\phi(P)) \;\leq\; dh(P) + c.$$

This can be written more succinctly write as

$$h(\phi(P)) = dh(P) + O(1),$$

where the $O(1)$ term indicates a bounded real function of P (the function $h(\phi(P)) - dh(P)$).

We first prove the upper bound; this is easy.

Lemma 25.12. *Let k be a number field and let $\phi \colon \mathbb{P}^n \to \mathbb{P}^m$ be a morphism $(\phi_0 \colon \cdots : \phi_n)$ defined by homogeneous polynomials $\phi_i \in k[x_0, \ldots, x_n]$ of degree d. There is a constant c such that*

$$h(\phi(P)) \leq dh(P) + c$$

for all $P \in \mathbb{P}^n(\bar{k})$.

[3] The correctness of this definition relies on some standard results from algebraic number theory that we will not prove here; the details are not important, all we need to know is that a normalization satisfying both (a) and (b) exists, see [1, p. 9] or [2, pp. 225-227] for a more detailed exposition.

Proof. Let $c = N\prod_p \max_j |c_j|_p$, where c_j ranges over coefficients that appear in any ϕ_i, and N bounds the number of monomials appearing in any ϕ_i. If $P = (a_0 : \ldots : a_n)$ and $k = \mathbb{Q}(a_0, \ldots, a_n)$, then

$$H(\phi(P)) = \prod_{p\in P_k} \max_i |\phi_i(P)|_p \ \le\ \prod_{p\in P_k} \max_{i,j} |c_j a_i^d|_p = cH(P)^d,$$

by the multiplicativity of $|\ |_p$ and the triangle inequality. The lemma follows. □

We now make a few remarks about the morphism $\phi\colon \mathbb{P}^n \to \mathbb{P}^m$ appearing in the lemma. Morphisms with domain $\mathbb{P}^n$ are tightly constrained, more so than projective morphisms in general, because the ideal of $\mathbb{P}^n$ (as a variety), is trivial; this means that the polynomials defining ϕ are essentially unique up to scaling. This has several consequences.

- The polynomials ϕ_i defining ϕ cannot have a common zero in $\mathbb{P}^n(\bar{k})$; otherwise there would be a point at which ϕ is not defined. This requirement is not explicitly stated because it is implied by the definition of a morphism as a regular map.
- The image of ϕ in $\mathbb{P}^m$ is either a point (in which case $d = 0$), or a subvariety of dimension n; if this were not the case then the polynomials defining ϕ would have a common zero in $\mathbb{P}^n(\bar{k})$. The fact that $\operatorname{im}\phi$ is a variety follows from the fact that projective varieties are complete (so every morphism is a closed map). In particular, if ϕ is non-constant then we must have $m \ge n$.
- If ϕ is non-constant, then $d = [k(\mathbb{P}^n) : \phi^*(k(\operatorname{im}\phi))]$ is equal to the degree of the ϕ_i. In particular, if $d = 1$ then ϕ is a bijection from $\mathbb{P}^n$ to its image. Note that this agrees with out definition of the degree of a morphism of curves.

Corollary 25.13. *It ϕ is any automorphism of $\mathbb{P}^n$, then*

$$h(\phi(P)) = h(P) + O(1). \tag{1}$$

Proof. We must have $d = 1$, and we can apply Lemma 25.12 to ϕ^{-1} as well. □

The corollary achieves our goal in the case $d = 1$ and $m = n$. If $d = 1$ and $m > n$, after applying a suitable automorphism to $\mathbb{P}^m$ we can assume that $\operatorname{im}\phi$ is the linear subvariety of $\mathbb{P}^m$ defined by $x_{n+1} = x_{n+2} = \cdots = x_{m+1} = 0$, and it is clear that the orthogonal projection $(x_0 : \cdots : x_m) \mapsto (x_0 : \cdots : x_n)$ does not change the height of any point in this subvariety. It follows that (2) holds whenever $d = 1$, whether $m = n$ or not.

We now prove the general case

Theorem 25.14. *Let k be a number field and let $\phi\colon \mathbb{P}^n \to \mathbb{P}^m$ be a morphism $(\phi_0 \colon \cdots : \phi_n)$ defined by homogeneous polynomials $\phi_i \in k[x_0, \ldots, x_n]$ of degree d. Then*

$$h(\phi(P)) = dh(P) + O(1). \tag{2}$$

Proof. If $d = 0$ then ϕ is constant and the theorem holds trivially, so we assume $d > 0$. We will decompose ϕ as the composition of four morphisms: a morphism $\psi\colon \mathbb{P}^n \to \mathbb{P}^N$, an automorphism of $\mathbb{P}^N$, an orthogonal projection $\mathbb{P}^N \to \mathbb{P}^n \subseteq \mathbb{P}^m$, and an automorphism of $\mathbb{P}^m$. All but the morphism ψ change the logarithmic height of a point P by at most an additive constant that does not depend on P, and we will show that $h(\psi(P)) = dh(P)$.

The map $\psi = (\psi_0\colon \cdots : \psi_N)$ is defined as follows. We let $N = \binom{n+d}{d} - 1$, and take $\psi_0, \ldots, \psi_N$ to be the distinct monomials of degree d in the variables $x_0, \ldots, x_n$, in some order. Clearly the ψ_N have no common zero in $\mathbb{P}^n(\overline{\mathbb{Q}})$, so ψ defines a morphism $\mathbb{P}^n \to \mathbb{P}^N$. Let $P = (a_0\colon \cdots : a_n)$ be any point in $\mathbb{P}^n$, and let $k = \mathbb{Q}(a_0, \ldots, a_n)$. For each $p \in \mathcal{P}_k$,

$$\max_i |\psi_i(P)|_p = \max_j |a_j^d|_p = \max_j |a_j|_p^d = (\max_j |a_j|_p)^d,$$

and it follows that

$$H(\psi(P)) = \prod_{p \in \mathcal{P}_k} \max_i |\psi_i(P)|_p = \prod_{p \in \mathcal{P}_k} (\max_j |a_j|_p)^d = H(P)^d.$$

Thus $h(\psi(P)) = dh(P)$ as claimed. We now note that each ϕ_i is a linear combination of the ψ_j, thus ϕ induces an automorphism $\hat{\phi}\colon \mathbb{P}^N \to \mathbb{P}^N$, and after applying a second automorphism of $\mathbb{P}^N$ we may assume that the image of $\hat{\phi} \circ \psi$ in $\mathbb{P}^N$ is the variety defined by $x_{n+1} = \cdots = x_N = 0$. Taking the orthogonal projection from $\mathbb{P}^N$ to $\mathbb{P}^n$ embedded in $\mathbb{P}^m$ as the locus of $x_{n+1} = \cdots = x_m = 0$ does not change the height of any point, and we may then apply an automorphism of $\mathbb{P}^m$ to map this embedded copy of $\mathbb{P}^n$ to $\operatorname{im} \phi$. □

Remark 25.15. For an alternative proof of Theorem 25.14 using the Nullstellensatz, see [2, VIII.5.6].

Lemma 25.16. *Let $k/\mathbb{Q}$ be a finite Galois extension. Then $h(P^\sigma) = h(P)$ for all $P \in \mathbb{P}^n(k)$ and $\sigma \in \operatorname{Gal}(k/\mathbb{Q})$.*

Proof. The action of σ permutes $\mathcal{P}_k$, so if $P = (x_0 : \cdots : x_n)$ with $x_i \in k$, then

$$H(P^\sigma) = \prod_{p \in \mathcal{P}_k} \max_i |x_i^\sigma|_p = \prod_{p^\sigma \in \mathcal{P}_k} \max_i |x_i^\sigma|_{p^\sigma} = \prod_{p^\sigma \in \mathcal{P}_k} \max_i |x_i|_p = \prod_{p \in \mathcal{P}_k} \max_i |x_i|_p = H(P).$$

□

Remark 25.17. Lemma 25.16 also holds for $k = \overline{\mathbb{Q}}$.

Theorem 25.18 (Northcott)**.** *For any positive integers B, d, and n, the set*

$$\{P \in \mathbb{P}^n(k) : h(P) \le B \text{ and } [k : \mathbb{Q}] \le d\}$$

is finite.

Proof. Let $P = (x_0 : \cdots : x_n) \in \mathbb{P}^n(k)$ with $[k : \mathbb{Q}] \le d$. We can view each x_i as a point $P_i = (x_i : 1)$ in $\mathbb{P}^1(k)$, and we have

$$H(P) = \prod_{p \in \mathcal{P}_k} \max |x_i|_p \ge \max_i \prod_{p \in \mathcal{P}} \max_i(|x_i|_p, 1) = \max_i H(P_i).$$

Thus it suffices to consider the case $n = 1$, and we may assume $P = (x : 1)$ and $k = \mathbb{Q}(x)$.

Without loss of generality we may replace k by its Galois closure, so let $k/\mathbb{Q}$ be Galois with $\operatorname{Gal}(k/\mathbb{Q}) = \{\sigma_1, \ldots \sigma_d\}$. The point $Q = (x^{\sigma_1} : \cdots : x^{\sigma_d}) \in \mathbb{P}^{d-1}(k)$ is fixed by $\operatorname{Gal}(k/\mathbb{Q})$, hence by $\operatorname{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$, so $Q \in \mathbb{P}^{d-1}(\mathbb{Q})$. By Lemma 25.16, $h(Q) = h(P)$, so we have reduced to the case $k = \mathbb{Q}$, and by the argument above we can also assume $n = 1$.

The set $\{P \in \mathbb{P}^1(\mathbb{Q}) : h(P) \le B\}$ is clearly finite; each P can be represented as a pair of relatively prime integers of which only finitely many have absolute value at most e^B. □

25.6 Canonical height functions on elliptic curves

Theorem 25.19 (Tate)**.** *Let S be a set and let $r > 1$ a real number. Let $\phi\colon X \to X$ and $h\colon X \to \mathbb{R}$ be functions such that $h \circ \phi = rh + O(1)$, and let*

$$\hat{h}_\phi(x) := \lim_{n\to\infty} \frac{1}{r^n} h(\phi^n(x)).$$

Then $\hat{h}_\phi$ is the unique function $S \to \mathbb{R}$ for which

- *(i)* $\hat{h}_\phi = h + O(1)$*;*
- *(ii)* $\hat{h}_\phi \circ \phi = r\hat{h}_\phi$*.*

Proof. Choose c so that $|\frac{1}{r}h(\phi(x)) - h(x)| \le \frac{c}{r}$ for all $x \in S$. For all $n > 1$ we have

$$\left|\frac{1}{r^n}h(\phi^n(x)) - \frac{1}{r^{n-1}}h(\phi^{n-1}(x))\right| = \frac{1}{r^{n-1}}\left|\frac{1}{r}h(\phi(\phi^{n-1}(x)) - h(\phi^{n-1}(x))\right| \le \frac{c}{r^{n-1}},$$

thus for all $x \in S$ the sequence $\frac{1}{r^n}h(\phi^n(x))$ converges, so $\hat{h}_\phi$ is well defined.

For all $x \in S$ we have

$$|\hat{h}_\phi(x) - h(x)| \le \sum_{n=1}^{\infty}\left|\frac{1}{r^n}h(\phi^n(x)) - \frac{1}{r^{n-1}}h(\phi^{n-1}(x))\right| \le \sum_{n=1}^{\infty}\frac{c}{r^n} = \frac{c}{r-1},$$

so (i) holds. Property (ii) is clear, and for uniqueness we note that if $f = h + O(1)$ and $f \circ \phi = rf$ then applying the construction above with h replaced by f yields $\hat{f}_\phi = \hat{h}_\phi$, but it is also clear that $\hat{f}_\phi = f$, so $f = \hat{h}_\phi$. □

We now want to apply Theorem 25.19 to the set $S = E(\overline{\mathbb{Q}})$ with $\phi = [2]$ the multiplication-by-2 map and $r = 4$, It might seem natural to let h be the height function on the projective plane $\mathbb{P}^2$ containing our elliptic curve E, but as E is a one-dimensional variety, it is better to work with $\mathbb{P}^1$, so we will use the image of E under the projection $\mathbb{P}^2 \to \mathbb{P}^1$ defined by $(x : y : z) \mapsto (x : z)$.

To understand how $[2]$ operates on $\pi(E)$, we recall the formula to double an affine point $P = (x_1 : y_1 : 1)$ with $y_1 \ne 0$ computes the x-coordinate of $2P = (x_3 : y_3 : 1)$ via $x_3 = \lambda^2 - 2x_1$, with

$$\lambda^2 = \left(\frac{3x_1^2 + a_4}{2y_1}\right)^2 = \frac{9x_1^4 + 6a_4x_1^2 + a_4^2}{4y^2} = \frac{9x_1^4 + 6a_4x_1^2 + a_4^2}{4x_1^3 + 4a_4x_1 + 4a_6},$$

where we have used the curve equation $y^2 = x^3 + a_4x + a_6$ to get a formula that only depends on x_1. We then have

$$x_3 = \frac{9x_1^4 + 6a_4x_1^2 + a_4^2}{4x_1^3 + 4a_4x_1 + a_6} - 2x_1 = \frac{x_1^4 + 2a_4x_1^2 - 8a_6x_1 + a_4^2}{4x_1^3 + 4a_4x_1 + a_6}.$$

Putting this in projective form, we now define the map $\phi\colon \mathbb{P}^1 \to \mathbb{P}^1$ by

$$\phi(x : z) = (x^4 + 2a_4x^2z^2 - 8a_6xz^3 + a_4^2z^4 : 4x^3z + 4a_4xz^3 + a_6z^4).$$

The fact that $4a_4^3 + 27a_6^2 \ne 0$ ensures that the polynomials defining ϕ have no common zero in $\mathbb{P}^1(\overline{\mathbb{Q}})$, thus $\phi\colon \mathbb{P}^1 \to \mathbb{P}^1$ is a morphism of degree 4, and Theorem 25.14 implies that

$$h(\phi(P)) = 4h(P) + O(1).$$

Definition 25.20. Let E be an elliptic curve over a number field k. The *canonical height*

$$\hat{h}\colon E(\bar{k}) \to \mathbb{R}$$

is the function $\hat{h} = \hat{h}_\phi \circ \pi$, where $\hat{h}_\phi$ is the function given by Theorem 25.19, with $\phi\colon \mathbb{P}^1 \to \mathbb{P}^1$ as above and h the absolute height on $\mathbb{P}^1$. It satisfies $\hat{h}(2P) = 4\hat{h}(P)$ for all $P \in E(\mathbb{Q})$.

Theorem 25.21. *Let E be an elliptic curve over a number field k. For any bound B the set $\{P \in E(k) : \hat{h}(P) \leq B\}$ is finite.*

Proof. This follows immediately from Northcott's theorem and Theorem 25.19 part (i). □

Theorem 25.22 (Parallelogram Law)**.** *Let $\hat{h}$ be the canonical height function of an elliptic curve E over a number field k. Then for all $P, Q \in E(\bar{k})$ we have*

$$\hat{h}(P+Q) + \hat{h}(P-Q) = 2\hat{h}(P) + 2\hat{h}(Q)$$

Proof. This is a straight-forward but tedious calculation that we omit; see [2, VIII.6.2]. □

25.7 Proof of the Mordell's Theorem

With all the pieces in place we now complete the proof of Mordell's theorem for an elliptic curve $E/\mathbb{Q}$ with a rational point of order 2.

Theorem 25.23. *Let $E/\mathbb{Q}$ be an elliptic curve with a rational point of order 2. Then $E(\mathbb{Q})$ is finitely generated.*

Proof. By the weak Mordell-Weil theorem that we proved in §25.4 for this case we know that $E(\mathbb{Q})/2E(\mathbb{Q})$ is finite. So let us choose a bound B such that the set

$$S\colon = \{P \in E(\mathbb{Q}) : \hat{h}(P) \leq B\}$$

contains a set S_0 of representatives for $E(\mathbb{Q})/2E(\mathbb{Q})$. We claim that S generates $E(\mathbb{Q})$.

Suppose for the sake of obtaining a contradiction that this is not the case. Then there is a point $Q \in E(\mathbb{Q}) - \langle S \rangle$ of minimal height $\hat{h}(Q)$; the fact that every set of bounded height is finite implies that $\hat{h}$ takes on discrete values, so such a Q exists. There is then a point $P \in S_0 \subset S$ such that $Q = P + 2R$ for some $R \in E(\mathbb{Q})$. Since $Q \not\in \langle S \rangle$, we must have $R \not\in \langle S \rangle$, so $\hat{h}(R) \geq \hat{h}(Q)$, by the minimality of $\hat{h}(Q)$. By the parallelogram law,

$$\begin{aligned}
2\hat{h}(P) &= \hat{h}(Q+P) + \hat{h}(Q-P) - 2\hat{h}(Q) \\
&\geq 0 + \hat{h}(2R) - 2\hat{h}(Q) \\
&= 4\hat{h}(R) - 2\hat{h}(Q) \\
&\geq 2\hat{h}(Q)
\end{aligned}$$

So $\hat{h}(Q) \leq \hat{h}(P) \leq B$ and therefore $Q \in S$, a contradiction. □

References

[1] J-P. Serre, *Lectures on the Mordell-Weil theorem*, 3rd edition, Springer Fachmedien Wiesbaden, 1997.

[2] J. H. Silverman, *The arithmetic of elliptic curves*, Springer, 2009.

18.782 Introduction to Arithmetic Geometry **Fall 2013**
Lecture #26 **12/10/2013**

26.1 Genus 1 curves with no rational points

Let C/k be a (smooth, projective, geometrically irreducible) curve of genus 1 over a perfect field k. Let n be the least positive integer for which $\operatorname{Div}_k C$ contains an effective divisor D of degree n (such divisors exist; take the pole divisor of any non-constant function in $k(C)$, for example). If C has a k-rational point, then $n = 1$ and C is an elliptic curve. We now consider the case where C does *not* have a rational point, so $n > 1$. We have $\deg(D) = n > 2g-2 = 0$, so the Riemann-Roch theorem implies

$$\ell(D) = \deg(D) + 1 - g = n,$$

and for any positive integer m we have

$$\ell(mD) = \deg(mD) + 1 - g = mn.$$

We now analyze the situation for some specific values of n.

26.1.1 The case $n = 2$

We have $\ell(D) = 2$, so let $\{1, x\}$ be a basis for $\mathcal{L}(D)$. Then $\ell(2D) = 4$, so in addition to $\{1, x, x^2\}$, the Riemann-Roch space $\mathcal{L}(2D)$ contains a fourth linearly independent function y. We then have $\{1, x, x^2, y, xy, x^3\}$ as a basis for $\mathcal{L}(3D)$, but $\mathcal{L}(4D)$ is an 8-dimensional vector space containing the 9 functions $\{1, x, x^2, y, xy, x^3, x^2y, x^4, y^2\}$, so there is a linear relation among them, and this linear relation must have nonzero coefficient on both y^2 and x^4. Assuming we are not in characteristic 2, we can complete the square in y to obtain an equation of the form

$$y^2 = f(x)$$

where f is a quartic polynomial over k. The polynomial f must be squarefree, and it cannot have any k-rational roots (otherwise we would have a rational point). Note that the homegenization of this equation is singular at $(0 : 1 : 0)$, but its desingularization is a curve in $\mathbb{P}^3$. Using the same argument as used on the problem set for hyperelliptic curves, one can show that every curve defined by an equation of this form has genus 1.

26.1.2 The case $n = 3$

We have $\ell(D) = 3$, so let $\{1, x, y\}$ be a basis for $\mathcal{L}(D)$. The 10 functions

$$\{1, x, y, x^2, xy, y^2, x^3, x^2y, xy^2, y^3\}$$

all lie in the 9-dimensional Riemann-Roch space $\mathcal{L}(3D)$, hence there is a linear relation among them that defines a plane cubic curve without any rational points. Conversely, every plane cubic curve has genus 1, since over a finite extension of k we can put the curve in Weierstrass form, which we have already proved has genus 1 (recall that genus is preserved under base extension of a perfect field). An example of a plane cubic curve with no rational points was given on the problem set, and here is another one:

$$3x^3 + 4y^3 + 5z^3 = 0.$$

Andrew V. Sutherland

Unlike the example on the problem set, this curve has a rational point locally everywhere, that is, over every completion of $\mathbb{Q}$. As noted back in Lecture 3, every geometrically irreducible plane curve has rational points modulo p for all sufficiently large primes p, and in this example the only primes that we need to check are 2, 3, and 5; it is easy to check that there are rational solutions modulo each of these primes, and modulo 3^3. Using Hensel's lemma, solutions modulo p (or p^3, for $p = 3$) can be lifted to $\mathbb{Q}_p$, and there are clearly solutions over $\mathbb{R} = \mathbb{Q}_\infty$

26.1.3 The case $n = 4$

We have $\ell(D) = 4$, so let $\{1, x, y, z\}$ be a basis for $\mathcal{L}(D)$. The 10 functions

$$\{1, x, y, z, x^2, y^2, z^2, xy, xz, yz\}$$

all lie in the 8-dimensional Riemann-Roch space $\mathcal{L}(2D)$, hence there are *two* independent linear relations among them, each corresponding to a quadratic form in $\mathbb{P}^3$, and C is the intersection of two quadric hypersurfaces (its clear that C is contained in the intersection, and one can show that it is equal to the intersection by comparing degrees).

26.2 The case $n > 4$

One can continue in a similar fashion for $n > 4$; indeed, by a theorem of Lang and Tate, over $\mathbb{Q}$ there are genus 1 curves that exhibit every possible value of n. But the situation becomes quite complicated already for $n = 5$: we have $\{1, w, x, y, z\}$ as a basis for $\mathcal{L}(D)$ and in $\mathcal{L}(2D)$ we get 15 functions in a Riemann-Roch space of dimension 10.[1]

26.3 Twists of elliptic curves

A genus one curve C/k with no k-rational points is not an elliptic curve, but for some finite extension L/k the set $C(L)$ will be nonempty; thus if base-extend C to L, we obtain an elliptic curve over L. We will show, this elliptic curve can be defined by a Weierstrass equation whose coefficients actually lie in k, so it is also the base-extension of an elliptic curve E/k. The curves E and C are clearly not isomorphic over k, since E has a k-rational point and C does not, but they become isomorphic when we base-extend to L. In other words, the isomorphism $\varphi\colon C \to E$ is defined over L, but not over k, so the distinguished k-rational point O on E is the image of an L-rational point on C that is not defined over k.

Definition 26.1. Two varieties defined over a field k that are related by an isomorphism defined over $\bar{k}$ are said to be *twists* of each other.

In order to characterize the curves that are twists of a given elliptic curve E/k, we introduce the *j-invariant*. For simplicity, we will assume henceforth that $\operatorname{char}(k) \neq 2, 3$, so that we can put our elliptic curves in short Weierstrass form. But the j-invariant can also be defined in terms of a general Weierstrass equation and except where we explicitly note otherwise, all the theorems we will prove are true in any characteristic.

[1] Note that while every curve can be smoothly embedded in $\mathbb{P}^3$, this embedding will not necessarily be defined over k. Over k, $\mathbb{P}^{n-1}$ is the best we can do.

Definition 26.2. Let E/k be an elliptic curve with Weierstrass equation $y^2 = x^3+a_4x+a_6$. The *j-invariant* of E is

$$j(E) := 1728\frac{4a_4^3}{4a_4^3 + 27a_6^2}.$$

Note that the denominator is always nonzero, since $\Delta(E) = -16(4a^3 + 27a_6^2) \neq 0$.

Theorem 26.3. *For every $j \in k$ there exists an elliptic curve E/k with $j(E) = j$.*

Proof. We define such an E/k via an equation $y^2 = x^3 + a_4x + a_6$ as follows. If $j = 0$, let $a_4 = 0$ and $a_6 = 1$, and if $j = 1728$, let $a_4 = 1$ and $a_6 = 0$. Otherwise, let $a_4 = 3j(1728 - j)$ and $a_6 = 2j(1728 - j)^2$. One can check that $\Delta(E) \neq 0$ and $j(E) = j$ in each case. □

Theorem 26.4. *Two elliptic curves defined over k have the same j-invariant if and only if they are isomorphic over $\bar{k}$.*

Proof. For the forward implication, let $y^2 = x^3 + a_4x + a_6$ and $y^2 = x^3 + a_4'x + a_6'$ be Weierstrass equations for elliptic curve E/k and E'/k, respectively, with $j(E) = j(E') = j$. If $j = 0$ then $a_4 = a_4' = 0$, and we can make $a_6 = a_6'$ by a linear change of variables defined over a suitable extension of k, hence $E \simeq_{\bar{k}} E'$. If $j = 1728$ then $a_6 = a_6' = 0$, and we can similarly make $a_4 = a_4'$ via a change of variables over a suitable extension of k. Otherwise, over a suitable extension of k we can make a_4 and a_4' both equal to 1, and then $j(E) = j(E') \Rightarrow a_6 = a_6'$. Thus in every case, $j(E) = j(E') \Rightarrow E \simeq_{\bar{k}} E'$.

For the reverse implication, we note that the cubic $x^3 + a_4x + a_6$ is uniquely determined by its roots, which are precisely the x-coordinates $\{x_1, x_2, x_3\}$ of the three points of order 2 in $E(\bar{k})$. If $E \simeq_{\bar{k}} E'$, then both curves can be embedded in $\mathbb{P}^2$ so that $E[2] = E'[2]$, and they will then have the same Weierstrass equation, hence the same j-invariant. □

Corollary 26.5. *Let C/k be a genus one curve and let O and O' be any two points in $C(\bar{k})$. Then the elliptic curves (C, O) and (C, O') over $\bar{k}$ have the same j-invariant.*

Proof. The translation-by-O' map on (C, O) is an isomorphism from (C, O) to (C, O'). □

It follows from the corollary that the j-invariant of an elliptic curve (E, O) is independent of the choice of O, it depends only on the curve E.

Definition 26.6. Let C/k be a curve of genus one. The j-invariant $j(C)$ of C is the j-invariant of the elliptic curve (C, O) over $\bar{k}$, for any $O \in C(\bar{k})$.

Theorem 26.7. *Let C/k be a curve of genus one. Then $j(C) \in k$.*

Proof. Let us pick $O \in C(L)$, where L is some finite Galois extension L/k, and let E/L be the elliptic curve (C, O). Then E is isomorphic to the base extension of C to L, so let $\varphi\colon C \to E$ be the isomorphism (which is defined over L). For any $\sigma \in \operatorname{Gal}(L/k)$ there is an isomorphism $\varphi^\sigma\colon C^\sigma \to E^\sigma$. But C is defined over k, so $C^\sigma = C$, and therefore $E^\sigma \simeq_L E$, so $j(E^\sigma) = j(E)$. But then $j(E)^\sigma = j(E^\sigma) = j(E)$ for all $\sigma \in \operatorname{Gal}(L/k)$, so $j(E) \in k$. □

Corollary 26.8. *Every genus one curve C/k is a twist of an elliptic curve E/k.*

The corollary does not uniquely determine E, not even up to k-isomorphism; it is possible for two elliptic curves defined over k to be twists without being isomorphic over k. For example, for any $d \in k^\times$ the elliptic curves defined by the Weierstrass equations

$$E\colon y^2 = x^3 + a_4x + a_6$$

and

$$E_d\colon y^2 = x^3 + d^2a_4x + d^3a_6$$

have the same j-invariant and are related by the isomorphism $(x, y) \mapsto (x/d, y/d^{3/2})$, which is defined over $k(\sqrt{d})$. But unless $d \in k^{\times 2}$, they are not isomorphic over k; the curves E and E_d are said to be *quadratic twists* of each other. More generally, we have the following.

Lemma 26.9. *Let $E\colon y^2 = x^3 + a_4x + a_6$ and $E'\colon y^2 = x^3 + a_4'x + a_6'$ be elliptic curves defined over k, with $j(E) = j(E')$. Then for some $\lambda \in \bar{k}^\times$ we have $a_4' = \lambda^4 a_4$ and $a_6' = \lambda^6 a_6$. Moreover, the degree of $k(\lambda)/k$ divides* 2,4,6 *when $a_4a_6 \neq 0$, $a_6 = 0$, $a_4 = 0$, respectively.*

Proof. We first assume $a_4a_6 \neq 0$. From the definition of the j-invariant, we have

$$\begin{aligned}(4a_4'^3 + 27a_6'^2)a_4^3 &= (4a_4^3 + 27a_6^2)a_4'^3\\ 4 + 27(a_6'^2/a_4'^3) &= 4 + 27(a_6^2/a_4^3)\\ a_6'^2a_4^3 &= a_6^2a_4'^3.\end{aligned}$$

If we let $\lambda = \sqrt{(a_6'a_4)/(a_6a_4')}$ then we have $a_4' = \lambda^4a_4$ and $a_6' = \lambda^6a_6$ as desired. When $a_6 = 0$ we may simply take $\lambda = \sqrt[4]{a_4'/a_4}$, and when $a_4 = 0$ we may take $\lambda = \sqrt[6]{a_6'/a_6}$. □

We now want to distinguish (up to k-isomorphism) a particular elliptic curve E/k that is a twist of a given genus one curve C/k. For any twist E/k of C/k we have an isomorphism $\phi\colon C \to E$ that is defined over some extension L/k of k that lies in $\bar{k}$. Every $\sigma \in \operatorname{Gal}(\bar{k}/k)$ defines an isomorphism $\phi^\sigma\colon C^\sigma \to E^\sigma$, and since C and E are both defined over k, we have $C^\sigma = C$ and $E^\sigma = E$, so in fact ϕ^σ is an isomorphism from C to E. The map

$$\varphi_\sigma := \phi^\sigma \circ \phi^{-1}$$

is then an isomorphism from E to itself. Every such isomorphism can be written as

$$\varphi_\sigma = \tau_{P_\sigma} \circ \varepsilon_\sigma,$$

where $P_\sigma = \varphi_\sigma(O)$ and ε_σ is an isomorphism that fixes the distinguished point $O \in E(k)$. Both τ_P and ε_σ are isomorphisms from E to itself, but ε_σ is also an isogeny, which is not true of τ_{P_σ} unless it is the identity map.

Definition 26.10. An *automorphism* of an elliptic curve E is an isomorphism $E \to E$ that is also an isogeny. The set of automorphisms of E form a group $\operatorname{Aut}(E)$ under composition.

Theorem 26.11. *Let k be a field of characteristic not equal to* 2 *or* 3.[2] *The automorphism group of an elliptic curve E/k is a cyclic group of order* 6, 4, *or* 2, *depending on whether $j(E)$ is equal to* 0, 1728, *or neither, respectively.*

Proof. We may assume E/k is in short Weierstrass form. Any automorphism ε^* of the function field $k(E)$ must preserve the Riemann-Roch space $\mathcal{L}(O)$, which has $\{1, x\}$, as a basis, and also the Weierstrass coefficients a_4 and a_6. It follows from Lemma 26.9 that $\varepsilon^*(x) = \lambda^{-2}x$, where λ is a 6th, 4th, or 2nd root of unity, as $j(E) = 0, 1728$, or neither, and we must then have $\varepsilon^*(y) = \lambda^{-3}y$. This uniquely determines ε^* and therefore ε. □

[2] Over a field of characteristic 2 or 3 one can have automorphism groups of order 24 or 12, respectively; this occurs precisely when $j(E) = 0 = 1728$.

Theorem 26.12. *Let C/k be a genus one curve. There is an elliptic curve E/k related to C/k by an isomorphism $\phi\colon C \to E$ such that for every automorphism $\sigma \in \operatorname{Gal}(\bar{k}/k)$ the isomorphism $\varphi_\sigma\colon E \to E$ defined by $\varphi_\sigma := \phi^\sigma \circ \phi^{-1}$ is a translation-by-P_σ map for some $P_\sigma \in E(\bar{k})$. The curve E is unique up to k-isomorphism.*

Proof. To simplify matters we assume $j(C) \neq 0, 1728$ and $\operatorname{char}(k) \neq 2, 3$. We first pick a point $Q_0 \in C(\bar{k})$ and let E be the elliptic curve (C, Q_0). We have $j(E) = j(C) \in k$, so we can put E in short Weirestrass form with coefficients $a_4, a_6 \in k$, and we have an isomorphism $\phi\colon C \to E$ that sends Q_0 to $O := (0:1:0)$, but it need not be the case that φ_σ is a translation-by-P_σ map for every $\sigma \in \operatorname{Gal}(\bar{k}/k)$.

We can write each of the isomorphisms $\varphi_\sigma = \phi^\sigma \circ \phi^{-1}$ as

$$\varphi_\sigma = \tau_{P_\sigma} \circ \varepsilon_\sigma,$$

where τ_{P_σ} is translation by $P_\sigma = Q_0^\sigma - Q_0$, and $\varepsilon_\sigma \in \operatorname{Aut}(E)$.

Since $j(E) \neq 0, 1728$, we have $\#\operatorname{Aut}(E) = 2$. The group $\operatorname{Aut}(E)$ clearly contains the identity map $[1]$ and the negation map $[-1]$, so $\operatorname{Aut}(E) = \{[\pm 1]\}$. The Galois group $\operatorname{Gal}(\bar{k}/k)$ acts on $\operatorname{Aut}(E)$ trivially, since both $[1]$ and $[-1]$ are defined over k.

If we apply an automorphism $\rho \in \operatorname{Gal}(\bar{k}/k)$ to φ_σ we obtain

$$\varphi_\sigma^\rho = (\phi^\sigma)^\rho \circ (\phi^{-1})^\rho = (\phi^{\rho\sigma}) \circ \phi^{-1} \circ \phi \circ (\phi^\rho)^{-1} = \varphi_{\rho\sigma} \circ \varphi_\rho^{-1}.$$

Thus

$$\varphi_{\rho\sigma} = \varphi_\sigma^\rho \circ \varphi_\rho = (\tau_{P_\sigma} \circ \varepsilon_\sigma)^\rho \circ (\tau_{P_\rho} \circ \varepsilon_\rho) = \tau_{P_\sigma^\rho + P_\rho} \circ (\varepsilon_\sigma^\rho \circ \varepsilon_\rho) = \tau_{P_{\sigma\rho}} \circ \varepsilon_\sigma \circ \varepsilon_\rho,$$

since ρ fixes ε_σ. But we also have $\varphi_{\rho\sigma} = \tau_{P_{\rho\sigma}} \circ \varepsilon_{\rho\sigma}$, thus $\varepsilon_{\rho\sigma} = \varepsilon_\sigma \circ \varepsilon_\rho = \varepsilon_\rho \circ \varepsilon_\sigma$, since $\operatorname{Aut}(E)$ is commutative. The map $\sigma \to \varepsilon_\sigma$ is thus a group homomorphism $\pi\colon \operatorname{Gal}(\bar{k}/k) \to \operatorname{Aut}(E)$. If the kernel of π is all of $\operatorname{Gal}(\bar{k}/k)$, then every ε_σ is trivial and φ_σ is translation-by-P_σ for all $\sigma \in \operatorname{Gal}(\bar{k}/k)$, as desired.

Otherwise the kernel if π is an index-2 subgroup of $\operatorname{Gal}(\bar{k}/k)$ whose fixed field is a quadratic extension $k(\sqrt{d})/k$ for some $d \in k^\times$. In this case let us consider the quadratic twist E_d of E by d, as defined above, and let $\chi_d\colon E \to E_d$ be the isomorphism $(x, y) \mapsto (x/d, y/d^{3/2})$. We then have an isomorphism $\phi_d = \chi_d \circ \phi$ from C to E_d, and for each $\sigma \in \operatorname{Gal}(\bar{k}/k)$ an isomorphism

$$\tilde{\varphi}_\sigma = \phi_d^\sigma \circ \phi_d^{-1} = (\chi_d \circ \phi)^\sigma \circ (\chi_d \circ \phi)^{-1} = \chi_d^\sigma \circ \phi^\sigma \circ \phi^{-1} \circ \chi_d^{-1} = \chi_d^\sigma \circ \varphi_\sigma \circ \chi_d^{-1}.$$

If $\varepsilon_\sigma = [1]$ then σ fixes $k(\sqrt{d})$ and therefore $\chi_d^\sigma = \chi_d$ and $\tilde{\varphi}_\sigma$ is just translation by $\chi_d(P_\sigma)$, since in this case $\varphi_\sigma = \tau_{P_\sigma}$ and χ_d commutes group operations on E and E_d (since it is an isogeny). If $\varepsilon_\sigma = [-1]$ then $\sigma(\sqrt{d}) = -\sqrt{d}$ and $\chi_d^\sigma = \chi_d \circ [-1]$, and now $\varphi_\sigma = \tau_{P_\sigma} \circ [-1]$. We then have

$$\tilde{\varphi}_\sigma = (\chi_d \circ [-1]) \circ (\tau_{P_\sigma} \circ [-1]) \circ \chi_d^{-1},$$

and now $\tilde{\varphi}_\sigma$ is translation by $\chi_d(-P_\sigma)$. Thus in every case $\tilde{\varphi}_\sigma$ is a translation map, so replacing E by E_d and ϕ by ϕ_d yields the desired result.

If $\phi'\colon C \to E'$ is another isomorphism with the same property then after composing with a suitable translation if necessary we can assume$\phi'(Q_0)$ is the point $O = (0:1:0)$ on E'. The map $\phi' \circ \phi^{-1}$ is then an isomorphism from E to E' that is fixed by every $\sigma \in \operatorname{Gal}(\bar{k}/k)$, hence defined over k, so E is unique up to k-isomorphism. □

Definition 26.13. The elliptic curve E/k given by Theorem 26.12 is the *Jacobian* of the genus one curve C/k; it is determined only up to k-isomorphism, so we call any elliptic curve that is k-isomorphic to E "the" Jacobian of C.

Note that if C is in fact an elliptic curve, then it is its own Jacobian.

We now want to give an alternative characterization of the Jacobian in terms of the Picard group. We will show that the Jacobian of a genus one curve C/k is isomorphic to $\operatorname{Pic}^0 C$; more precisely, for every algebraic extension L/k we have $E(L) \simeq \operatorname{Pic}^0_L C$ (as abelian groups). This characterization of the Jacobian has the virtue that it applies to curves of any genus; although we will not prove this, for each curve C/k of genus g there is an abelian variety A/k of dimension g such that $A(L) \simeq \operatorname{Pic}^0_L C$ for all algebraic extensions L/k.

In order to to prove this for curves of genus one, we first introduce the notion of a principal homogeneous space.

26.4 Principal homogeneous spaces (torsors)

Recall that an *action* of a group G on a set S is a map $G \times S \to S$ such that the identity acts trivially and the action of gh is the same as the action of h followed by the action of g. With the action written on the left, this means $(gh)s = g(hs)$, or on the right, $s^{(gh)} = (s^h)^g$, where $g, h \in G$ and $s \in S$. Below are various properties that group actions may have:

- *faithful*: no two elements of G act the same way on *every* $s \in S$ ($\forall s(gs = hs) \Rightarrow g = h$).
- *free*: no two elements of G act in the same way on *any* $s \in S$ ($\exists s(gs = hs) \Rightarrow g = h$).
- *transitive*: for every $s, t \in S$ there is a $g \in G$ such that $gs = t$.
- *regular*: free and transitive; for all $s, t \in S$ there is a *unique* $g \in G$ with $gs = t$.

Note that free implies faithful, so long as $S \neq \emptyset$.

Definition 26.14. A nonempty set S equipped with a regular group action by an abelian group G is a *principal homogeneous space for* G, also known as a G*-torsor*.

Since a G-torsor S is being acted upon by an abelian group, it is customary to write the action additively on the right. So for any $s \in S$ and $g \in G$ we write $s + g$ to denote the action of g on S (which is another element t of S). Conversely, for any $s, t \in S$ we write $t - s$ to denote the unique $g \in G$ for which $t = s + g$.

As a trivial example of a G-torsor, we can take G acting on itself. More generally, any G-torsor S is necessarily in bijection with G. In fact, we can make S into a group isomorphic to G as follows: pick any element $s_0 \in S$, and define the bijection $\phi\colon G \to S$ by $\phi(g) = s_0 + g$. Declaring ϕ to be a group homomorphism makes S into a group; the group operation is given by $\phi(g) + \phi(h) = \phi(g + h)$, and ϕ is an isomorphism with the map $s \mapsto s - s_0$ as its inverse.

A good analogy for the relationship between G and S is the relationship between a vector space and affine space. A G-torsor is effectively a group with no distinguished identity element, just as affine space is effectively a vector space with no distinguished origin.

26.5 Principal homogeneous spaces of elliptic curves

The notion of a G-torsor S defined above is entirely generic; we now specialize to the case where $G = E(\bar{k})$ is the group of points on an elliptic curve E/k and $S = C(\bar{k})$ is the set of points on a curve C/k. In this setting we add the additional requirement that the action is given by a morphism of varieties. More formally, we make the following definition.

Definition 26.15. Let E/k be an elliptic curve. A *principal homogeneous space for* E (or E*-torsor*), is a genus one curve C/k such that the set $C(\bar{k})$ is an $E(\bar{k})$-torsor and the map $C \times E \to C$ defined by $(Q, P) \mapsto Q + P$ is a morphism of varieties that is defined over k.

Note that if C/k is an E-torsor and L/k is any algebraic extension over which C has an L-rational point P, then the set $C(L)$ is an $E(L)$-torsor and the elliptic curves (E, O) and (C, P) are isomorphic over L via the translation-by-P map. In particular, we always have $j(C) = j(E)$. If C has a k-rational point then C and E are isomorphic over k, and in general E is the Jacobian of C, as we now prove.

Theorem 26.16. *Let C/k be a curve of genus one and let E/k be an elliptic curve. Then C is an E-torsor if and only if E is the Jacobian of C.*

Proof. Suppose C is an E-torsor, let O be the distinguished point of E and pick any $Q_0 \in C(\bar{k})$. Then we have an isomorphism $\phi\colon C \to E$ that sends to Q_0 to O defined by $Q \mapsto Q - Q_0$, where $Q - Q_0$ denotes the unique element of $E(\bar{k})$ that sends Q to Q_0. For any $\sigma \in \operatorname{Gal}(\bar{k}/k)$, the map $\varphi_\sigma = \phi^\sigma \circ \phi^{-1}$ is given by $P \mapsto (Q_0 + P) - Q_0^\sigma$, and is thus translation by $P_\sigma = Q_0 - Q_0^\sigma$. So E is the Jacobian of C (up to k-isomorphism).

Now suppose E is the Jacobian of C and let $\phi\colon C \to E$ be the isomorphism from C to E given by Theorem 26.12. Then $P \in E(\bar{k})$ acts on $Q \in C(\bar{k})$ via $Q \mapsto \phi^{-1}(\phi(Q) + P)$, and this action is regular, since ϕ and translation-by-P are both isomorphisms. Thus $C(\bar{k})$ is an $E(\bar{k})$-torsor, and the map $\mu\colon C \times E \to C$ given by the action of E is clearly a morphism of varieties, since both ϕ and the group operation $E \times E \to E$ are.

To show that μ is defined over k, we check that $\mu^\sigma = \mu$ for all $\sigma \in \operatorname{Gal}(\bar{k}/k)$. The group operation $E \times E \to E$ is defined over k, hence invariant under the action of σ, and for any $Q \in C$ and $P \in E$ we have

$$\begin{aligned}
\mu^\sigma(Q, P) &= (\phi^{-1})^\sigma(\phi^\sigma(Q) + P)\\
&= (\phi^{-1})^\sigma((\varphi_\sigma \circ \phi)(Q) + P)\\
&= (\phi^{-1})^\sigma(\phi(Q) + P_\sigma + P)\\
&= \phi^{-1}(\phi(Q) + P_\sigma + P - P_\sigma)\\
&= \phi^{-1}(\phi(Q) + P)\\
&= \mu(Q, P),
\end{aligned}$$

where we have used $\varphi_\sigma = \phi^\sigma \circ \phi^{-1}$ to derive $\phi^\sigma = \varphi_\sigma \circ \phi$ and $(\phi^{-1})^\sigma = (\phi^\sigma)^{-1} = \phi^{-1} \circ \varphi_\sigma^{-1}$, and applied $\varphi_\sigma(P) = P + P_\sigma$ and $\varphi_\sigma^{-1}(P) = P - P_\sigma$. □

Theorem 26.17. *Let C/k be an E-torsor and let $Q_0 \in C(\bar{k})$. The map $\pi\colon \operatorname{Div}^0_{\bar{k}} C \to E(\bar{k})$ defined by*

$$\sum_i n_i P_i \mapsto \sum_i n(P_i - Q_0)$$

is a surjective homorphism whose kernel consists of the principal divisors, and it is independent of the choice of Q_0. Moreover, for any extension L/k in $\bar{k}$ the map π commutes with every element of $\operatorname{Gal}(\bar{k}/L)$ and therefore induces a canonical isomorphism $\operatorname{Pic}^0_L C \simeq E(L)$.

Note that in the definition of π, the sum on the LHS is a formal sum denoting a divisor, while the sum on the RHS is addition in the abelian group $E(\bar{k})$, where each term $P_i - Q_0$ denotes the unique element of $E(\bar{k})$ whose action sends Q_0 to P_i.

Proof. The map π is clearly a group homomorphism. To see that it is surjective, for any point $P \in E(\bar{k})$, if we let $D = (Q_0 + P) - Q_0 \in \operatorname{Div}^0 C$ then

$$\pi(D) = ((Q_0 + P) - Q_0) - (Q_0 - Q_0) = P.$$

If $\pi(D) = \pi(\sum n_i P_i) = O$ for some $D \in \operatorname{Div}^0_{\bar{k}} C$, then the divisor $\sum_i n_i(P_i - Q_0)$ in $\operatorname{div}^0_{\bar{k}}(E)$ sums to O, hence is linearly equivalent to 0 and therefore a principal divisor. Since $\bar{k}(C) = \bar{k}(E)$, the same is true of D. Conversely, if $D \in \operatorname{div}^0_{\bar{k}} C$ is principal, so is the corresponding divisor in $\operatorname{Div}^0_{\bar{k}} E$, and therefore $\pi(D) = O$. Thus the kernel of π is precisely the group of principal divisors, hence π induces an isomorphism $\operatorname{Pic}^0_{\bar{k}} \to E(\bar{k})$.

Now let $Q_1 \in C(\bar{k})$ and define $\pi'(\sum n_i P_i) = \sum n_i (P_i - Q_1)$. Then

$$\pi(D) - \pi'(D) = \sum_i n_i((P_i - Q_0) - (P_i - Q_1)) = \sum n_i(Q_1 - Q_0) = O,$$

since $\sum n_i = \deg(D) = 0$, thus $\pi' = \pi$ and π is independent of the choice of Q_0.

For any $\sigma \in \operatorname{Gal}(\bar{k}/k)$ and $D = \sum n_i P_i \in \operatorname{Div}^0_{\bar{k}} C$ we have

$$\pi(D)^\sigma = \sum_i n_i(P_i^\sigma - Q_0^\sigma) = \pi(D^\sigma).$$

It follows that $D \in \operatorname{Div}^0_L C$ if and only if $\pi(D) \in E(L)$, for any extension L/k in $\bar{k}$, thus π induces an isomorphism $\operatorname{Pic}^0_L C \to E(L)$ for every L/k in $\bar{k}$. □

26.6 The Weil-Châtelet group

Definition 26.18. Let E/k be an elliptic curve. Two E-torsors C/k and C'/k are *equivalent* if there is an isomorphism $\theta\colon C \to C'$ defined over k that is compatible with the action of E. This means that

$$\theta(Q + P) = \theta(Q) + P$$

holds for all $Q \in C(\bar{k})$ and $P \in E(\bar{k})$. The *Weil-Châtelet group* $\mathrm{WC}(E/k)$ is the set of equivalence classes of E-torsors under this equivalence relation.

The equivalence class of E is simply the set of elliptic curves that are k-isomorphic to E; this is the *trivial class* of $\mathrm{WC}(E/k)$, and it acts as the identity element under the group operation that we will define shortly.

Lemma 26.19. *If $\theta\colon C \to C'$ is an equivalence of E-torsors then*

$$\theta(P) - \theta(Q) = P - Q$$

for all $P, Q \in C$. Conversely, if $\theta\colon C \to C'$ is a k-isomorphism for which the above holds, then θ is an equivalence of E-torsors.

Proof. If θ is an equivalence of E-torsors, then

$$\begin{aligned}\theta(P)-\theta(Q) &= \theta(P)+(Q-P)-\theta(Q)+P-Q\\ &= \theta(P+(Q-P))-\theta(Q)+P-Q\\ &= P-Q.\end{aligned}$$

Conversely, if $\theta(P)-\theta(Q)=P-Q$ for all $P,Q\in C$, then for any $R\in E(\bar{k})$ we have $\theta(Q+R)-\theta(Q)=(Q+R)-Q=R$, and therefore $\theta(Q+R)=\theta(Q)+R$ for all $Q\in C$ and $R\in E(\bar{k})$, so θ is an equivalence of E-torsors. □

Recall from the proof of Theorems 26.12 and 26.16 that if C/k is an E-torsor (and therefore E is the Jacobian of C) then each $\sigma\in\operatorname{Gal}(\bar{k}/k)$ determines an isomorphism $\varphi_\sigma\colon E\to E$ that is a translation-by-P_σ map, where $P_\sigma=Q_0^\sigma-Q_0$ for some fixed $Q_0\in C(\bar{k})$. So we have a map $\alpha\colon\operatorname{Gal}(\bar{k}/k)\to E(\bar{k})$ defined by $\alpha(\sigma)=Q_0^\sigma-Q_0$. For any $\sigma,\tau\in\operatorname{Gal}(\bar{k}/k)$ we have

$$\alpha(\sigma)^\tau=(Q_0^\sigma-Q_0)^\tau=Q_0^{(\tau\sigma)}-Q_0^\tau=(Q_0^{\tau\sigma}-Q_0)-(Q_0^\tau-Q_0)=\alpha(\tau\sigma)-\alpha(\tau),$$

thus

$$\alpha(\tau\sigma)=\alpha(\tau)+\alpha(\sigma)^\tau,$$

and this holds for any choice of Q_0 used to define α. If $\alpha(\sigma)^\tau=\alpha(\sigma)$ then α is a group homomorphism, but in general this is not the case; the map α is known as a *crossed homomorphism*.

Definition 26.20. A map $\alpha\colon\operatorname{Gal}(\bar{k}/k)\to E(\bar{k})$ that satisfies

$$\alpha(\tau\sigma)=\alpha(\tau)+\alpha(\sigma)^\tau$$

for all $\sigma,\tau\in\operatorname{Gal}(\bar{k}/k)$ is called a *crossed homomorphism*.

If α and β are two crossed homomorphism then the map $(\alpha+\beta)(\sigma)=\alpha(\sigma)+\beta(\sigma)$ is also, since

$$(\alpha+\beta)(\tau\sigma)=\alpha(\tau\sigma)+\beta(\tau\sigma)=\alpha(\tau)+\alpha(\sigma)^\tau+\beta(\tau)+\beta(\sigma)^\tau=(\alpha+\beta)(\tau)+(\alpha+\beta)(\sigma)^\tau,$$

and addition of crossed homomorphism is clearly associative. The difference of two crossed homomorphisms is similarly a crossed homomorphism, and the map that sends every element of $\operatorname{Gal}(\bar{k}/k)$ to the distinguished point O acts as an additive identity. Thus the set of all crossed homomorphisms from $\operatorname{Gal}(\bar{k}/k)$ to $E(\bar{k})$ form an abelian group.

The crossed homomorphisms of the form $\sigma\mapsto Q_0^\sigma-Q_0$ that arise from an E-torsor C/k with $Q_0\in C(\bar{k})$ have the property that there is a finite normal extension L/k such that $\operatorname{Gal}(\bar{k}/L)=\alpha^{-1}(O)$; take L to be the normal closure of $k(Q_0)$.[3] Crossed homomorphisms with this property are said to be *continuous*.[4] Sums and negations of continuous crossed homomorphisms are clearly continuous, so they form a subgroup.

Now let us consider what happens when we pick a point $Q_1\in C(\bar{k})$ different from Q_0. Let α_0 be the crossed homomorphism $\sigma\mapsto Q_0^\sigma-Q_0$ and let α_1 be the crossed homomorphism $\sigma\mapsto Q_1^\sigma-Q_1$. Then their difference is defined by

$$\alpha_1(\sigma)-\alpha_0(\sigma)=(Q_1^\sigma-Q_1)-(Q_0^\sigma-Q_0)=(Q_1-Q_0)^\sigma-(Q_1-Q_0).$$

[3] Recall that we assume k to be perfect.

[4] If we give $\operatorname{Gal}(\bar{k}/k)$ the Krull topology and $E(\bar{k})$ the discrete topology this corresponds to the usual notion of continuity.

The crossed homomorphism $\alpha_1 - \alpha_0$ is defined in terms of $Q_1 - Q_0$ which is actually a point on $E(\bar{k})$, rather than $C(\bar{k})$. This is also true if we choose $Q_0 \in C_0(\bar{k})$ and $Q_1 \in C_1(\bar{k})$ where C_0 and C_1 are two equivalent E-torsors.

Definition 26.21. Crossed homomorphisms of the form $\sigma \mapsto P^\sigma - P$ with $P \in E(\bar{k})$ are *principal.* The principal crossed homomorphism form a subgroup, as do the continuous principal crossed homomorphisms.

Given our notion of equivalence for E-torsors, we do not wish to distinguish between principal crossed homomorphisms. This leads to the following definition.

Definition 26.22. Let E/k be an elliptic curve. The group of continuous crossed homomorphisms of E/k modulo its subgroup of principal crossed homomorphisms is the *first Galois-cohomology group* of $E(\bar{k})$. It is denoted by

$$H^1(\mathrm{Gal}(\bar{k}/k), E(\bar{k})).$$

For the sake of brevity we may also write $H^1(k, E)$.

Remark 26.23. More generally, if M is any abelian group on which $\mathrm{Gal}(\bar{k}/k)$ acts, one can define Galois cohomology groups $H^n(k, M)$ for each non-negative integer n. The group $H^0(k, M)$ is simply the subgroup of M fixed by $\mathrm{Gal}(\bar{k}/k)$; in our setting $H^0(k, E) = E(k)$.

We now use the group $H^1(k, E)$ to define a group operation on the $\mathrm{WC}(E/k)$.

Theorem 26.24. *Let E/k be an elliptic curve. There is a bijection between the Weil-Châtelet group $\mathrm{WC}(E/k)$ of E and its first cohomology group $H^1(k, E)$.*

Proof. We have already defined a map from $\mathrm{WC}(E/k)$ to $H^1(k, E)$; given an E-torsor C/k that represents an equivalence class in $\mathrm{WC}(E/k)$, we may pick any point $Q_0 \in C(\bar{k})$ to get a continuous crossed homomorphism $\sigma \mapsto Q_0^\sigma - Q_0$ that is uniquely determined modulo prinicipal crossed homomorphisms, hence it represents an element of $H^1(k, E)$. We just need to show that this map is injective and surjective.

We first prove that it is injective. Let C_1/k and C_2/k be E-torsors, pick $Q_1 \in C_1(\bar{k})$ and $Q_2 \in C_2(\bar{k})$, and suppose that the crossed homomorphism $\sigma \mapsto Q_1^\sigma - Q_1$ and $\sigma \mapsto Q_2^\sigma - Q_2$ are equivalent in $H^1(k, E)$. Then their difference is a principal crossed homomorphism $\sigma \mapsto P^\sigma - P$, for some $P \in E(\bar{k})$. Thus we have

$$(Q_1^\sigma - Q_1) - (Q_2^\sigma - Q_2) = P^\sigma - P$$

for all $\sigma \in \mathrm{Gal}(\bar{k}/k)$. Now define the map $\theta\colon C_1 \to C_2$ by

$$\theta(Q) = Q_1 + (Q - Q_2) - P.$$

It is clear that θ is an isomorphism, since C_1 and C_2 are both E-torsors, and it is defined over k, since for any $\sigma \in \mathrm{Gal}(\bar{k}/k)$ we have

$$\begin{aligned}
\theta(Q)^\sigma &= Q_1^\sigma + (Q^\sigma - Q_2^\sigma) - P^\sigma \\
&= Q_1 - (Q^\sigma - Q_2) - P + (Q_1^\sigma - Q_1) - (Q_2^\sigma - Q_2) - (P^\sigma - P) \\
&= Q_1 - (Q^\sigma - Q_2) - P \\
&= \theta(Q^\sigma)
\end{aligned}$$

Thus C_1 and C_2 lie in the same equivalence class in $\mathrm{WC}(E/k)$; this prove injectivity.

For surjectivity, let α be a continuous crossed homomorphism that represents an element of $H^1(k,E)$. We now define an action of $\mathrm{Gal}(\bar{k}/k)$ on the function field $\bar{k}(E)=\bar{k}(x,y)$ as follows: for any $\sigma\in\mathrm{Gal}(\bar{k}/k)$, the elements x^σ and y^σ are given by

$$(x,y)^\sigma=(x^\sigma,y^\sigma):=(x,y)+\alpha(\sigma),$$

where the $+$ indicates that we apply the algebraic formulas defining the group operation on $E(\bar{k})$ working with points in $\mathbb{P}^2(\bar{k}(E))$. To check that this defines a group action, we note that the identity clearly acts trivially, and for any $\sigma,\tau\in\mathrm{Gal}(\bar{k}/k)$ we have

$$(x,y)^{\tau\sigma}=(x,y)+\alpha(\tau\sigma)=(x,y)+\alpha(\tau)+\alpha(\sigma)^\tau=((x,y)+\alpha(\sigma))^\tau+\alpha(\tau)=((x,y)^\sigma)^\tau.$$

The fixed field of this action is is the function field of a curve C that is defined over k and isomorphic to E over $\bar{k}$. By construction, there is an isomorphism $\phi\colon C\to E$ such that for any $\sigma\in\mathrm{Gal}(\bar{k}/k)$ the automorphism $\varphi_\sigma=\phi^\sigma\circ\phi^{-1}$ is a translation by $P_\sigma=-\alpha(\sigma)$, thus E is the Jacobian of C, by Theorem 26.12, and therefore C is an E-torsor, by Theorem 26.16. Thus C represents an equivalence class of $\mathrm{WC}(E/k)$, and if we pick $Q_0=\phi^{-1}(O)$ then

$$\begin{aligned}Q_0^\sigma-Q_0&=(\phi^\sigma)^{-1}(O)-\phi^{-1}(O)\\&=\phi^{-1}(O+\alpha(\sigma))-\phi^{-1}(O)\\&=\alpha(\sigma),\end{aligned}$$

So the class of α in $H^1(k,E)$ is the image of the class of C in $\mathrm{WC}(E/k)$. □

The bijection given by the theorem maps the trivial class of $\mathrm{WC}(E/k)$ to the identity element of $H^1(k,E)$, thus we can define a group operation on $\mathrm{WC}(E/k)$ via this bijection.

Corollary 26.25. *The Weil-Châtelet group* $\mathrm{WC}(E/k)$ *is isomorphic to the group* $H^1(k,E)$.

Definition 26.26. Let E/k be an elliptic curve. The *Tate-Shafarevich* group $Ш(E)$ is the kernel of the map

$$\mathrm{WC}(E/k)\to\prod_p\mathrm{WC}(E_p/k_p),$$

where k_p ranges over the completions of k and E_p denotes the base extension of E to k_p.

The Tate-Shafarevich group contains precisely the equivalence classes in $\mathrm{WC}(E/k)$ that are locally trivial everywhere. These are the classes of curves C/k with Jacobian E/k that have a k_p-rational point at every completion k_p.

Definition 26.27. A curve C/k satisfies the *local-global principle* (or *Hasse principle*) if either $C(k)\neq\emptyset$ or $C(k_p)=\emptyset$ for some completion k_p.

Theorem 26.28. *Let C/k be a genus one curve with Jacobian E/k. A genus one curve C/k fails the local-global principle if and only if it represents a non-trivial element of* $Ш(E)$.

References

[1] J. H. Silverman, *The arithmetic of elliptic curves*, Springer, 2009.