## 1.3 26/06/17 - Introduction III: Fermat for Exponent 4 and Infinite Descent; Review of Projective Geometry I: Basics

Let's first finish up our discussion of the Fermat problem for the exponent $n = 4$.

### 1.3.A Fermat for Exponent 4, Proof 1

Let's try to solve the equation
$$X^4 + Y^4 = Z^4.$$

This is the only case for which Fermat is believed to have a proof, and Fermat's proof uses a very important strategy, known now as "Fermat's method of infinite descent". Let us now review (in modern terminology) Fermat's proof. In fact, we will show something stronger.

**Theorem 1.3.1** (Fermat)**.** Let $X, Y, W \in \mathbb{Z}_{\geq 0}$ be such that $X^4 + Y^4 = W^2$. Then $XYW = 0$.

*Proof.* Suppose to the contrary there is a solution $(X, Y, W)$ with $X, Y, W > 0$, and pick such a solution with the smallest possible $W$. This implies in particular that $X^2, Y^2, W$ are pairwise coprime (why?) and hence using 1.2.3 that (possibly up to permuting $X$ and $Y$, which we do), there are unique coprime $m, n \in \mathbb{Z}$ of different parity with $m > n > 0$ such that
$$(X^2, Y^2, W) = (m^2 - n^2, 2mn, m^2 + n^2);$$
in particular, $X$ is odd. Now $X^2 = m^2 - n^2$, so that $X^2 + n^2 = m^2$. Again $X, n, m$ are pairwise coprime positive integers with $X$ odd, so 1.2.3 applied once again gives us unique coprime $p, q \in \mathbb{Z}$ of different parity with $p > q > 0$ and such that
$$(X, n, m) = (p^2 - q^2, 2pq, p^2 + q^2).$$
In particular, $n$ is even and hence $m$ is odd. Finally, $m = p^2 + q^2$ implies that $m, p$, and $q$ are pairwise coprime. Therefore
$$Y^2 = 2mn = 4mpq$$
implies that $p, q, m$ are all squares. Write $(p, q, m) = (X_1^2, Y_1^2, W_1^2)$ for some integers $X_1, Y_1, W_1 > 0$. Then
$$X_1^4 + Y_1^4 = p^2 + q^2 = m = W_1^2,$$
so that $(X_1, Y_1, W_1)$ is also a solution. Finally,
$$W_1 \leq W_1^2 = m \leq m^2 < m^2 + n^2 = W$$
tells us that this is a "smaller" solution, giving us the required contradiction. ∎

Can you see why this is called "infinite descent"?

The machinery of elliptic curves (specifically descent on them) is a vast modern generalization of this technique due to Fermat; this is where the name "descent" in the theory of elliptic curves comes from. Now Fermat's equation for $n = 4$ does not lend itself directly to the technology of elliptic cuvres, because it not one: $X^4 + Y^4 = Z^4$ defines a *quartic* curve in the plane which is a *canonical curve of genus two* (whatever that means). However, there is a clever way to massage it to give points on an elliptic curve; it is this curve that we do "descent" on. Since it's fun to figure out how to do this massaging, I will leave this as an exercise (2.1.12). We will figure out how to do descent on that curve–and hence prove Fermat's result in a second way–in a more general context much later.

In closing, let me remark that the case $n \geq 5$ of Fermat's theorem is genuinely very difficult, and it is beyond my abilities to go into the details here (see the history on the Wikipedia page, for instance). However, Andrew Wiles's proof from 1994 of it still nontrivially uses the theory of elliptic curves, following an idea going back to Frey from 1984. If we have time and interest, I am happy to explain the outline of the proof towards the end of the course; let me know if this is something you would like to see.

To fruitfully study plane curves, we need some basic understanding of the projective geometry of the plane. Let's take a detour to rapidly review what we need; a much fuller account can be found at [4].

### 1.3.B Basics

**Example 1.3.2.** Fix a field $K$, and for $s \in K$ consider the intersection of the two lines

$$sy = x \text{ and } x = 1.$$

When $s \neq 0$, there is a unique intersection point, namely $(1, 1/s)$, but when $s = 0$, there is no intersection at all, and the lines are parallel. What has happened is that as $s \to 0$, the intersection point has "escaped to infinity".

Such behavior can be extremely troublesome while doing geometry, and it is helpful to develop basic vocabulary to deal with "points at infinity". The modern approach to doing this involves working with *homogeneous coordinates on projective space*.

**Definition 1.3.3.** Let $K$ be a field and $n \in \mathbb{Z}_{\geq 0}$. We define the *set of $K$-points of projective $n$-space over $K$* to be

$$\mathbb{P}^n(K) := \{[X_0 : X_1 : \cdots : X_n] : X_0, \ldots, X_n \in K \text{ not all zero}\} / \sim$$

where $\sim$ means that we identify points with the same coordinates up to scaling by a nonzero number, i.e., for all $X_0, \ldots, X_n \in K$, not all zero, and $\lambda \in K^\times$ we have

$$[X_0 : \cdots : X_n] \sim [\lambda X_0 : \cdots : \lambda X_n].$$

The coordinates $X_0, \ldots, X_n$ are called *homogeneous coordinates*, and are well-defined only up to simultaneous scaling.

**Unimportant Remark 1.3.4.** For those who have seen linear algebra, the set $\mathbb{P}^n(K)$ can be identified with the set of one-dimensional $K$-linear subspaces of $K^{n+1}$ (how?). For those who have seen a little topology, when $K = \mathbb{R}$ or $K = \mathbb{C}$ (or more generally any topological field such as $K = \mathbb{Q}_p$), then $\mathbb{P}^n(K)$ can be given a natural topology, which for $K = \mathbb{R}, \mathbb{C}, \mathbb{Q}_p$ even makes it a(n) (analytic) manifold.

**Example 1.3.5.**

(a) When $n = 0$, the space $\mathbb{P}^0(K)$ is just one point.

(b) When $n = 1$, the points of $\mathbb{P}^1(K)$ look like $[S : T]$, where $S, T \in K$ are not both zero, and we can simultaneously scale $S$ and $T$ by the same nonzero element $\lambda$ of $K$. In particular, when $S \neq 0$, then scaling by $\lambda = S^{-1}$ gives the unique representative $[1 : t]$ where $t := T/S \in K$. In this way, $\mathbb{A}^1(K)$ sits naturally in $\mathbb{P}^1(K)$. The only point not contained in it is the point $[0 : 1]$ corresponding to $S = 0$; this can be thought of as corresponding to "$t = 1/0 = \infty$" and will often be called the *point at infinity* and denoted $\infty$. Therefore, $\mathbb{P}^1(K) = \mathbb{A}^1(K) \amalg \{\infty\}$ is the "one-point compactification" of $\mathbb{A}^1(K)$. (Warning: This is not true for higher $n$.) This allows us to handle infinity easily–no actual division by zero needed.

(c) When $n = 2$, we are looking at triples $[X : Y : Z]$ of $X, Y, Z \in K$, not all zero, up to simultaneous scaling. Again, if $Z \neq 0$, then scaling by $\lambda = Z^{-1}$ gives us the unique representative $[x : y : 1]$ where $(x, y) = (X/Z, Y/Z)$, and again we get an $\mathbb{A}^2(K) \subset \mathbb{P}^2(K)$. The complement corresponds to $Z = 0$, which is the set of points $[X : Y : 0]$, in natural bijection with $\mathbb{P}^1(K)$ itself. This locus is called the *line at infinity*; the point $[X : Y : 0]$ should be thought of as the point infinitely far away in the direction of the vector $[X : Y]$. In this way, $\mathbb{P}^2(K) = \mathbb{A}^2(K) \amalg \mathbb{P}^1(K)$. Said yet another way, we have

$$\big\{[X : Y : Z] \in \mathbb{P}^2(K) : Z \neq 0\big\} = \big\{[X/Z : Y/Z : 1] \in \mathbb{P}^2(K)\big\} = \{(x, y) \in \mathbb{A}^2(K)\},$$

where again $(x, y) = (X/Z, Y/Z)$.

**Unimportant Remark 1.3.6.**

(a) Continuing in the above fashion, we can describe a decomposition of $\mathbb{P}^n(K)$ of the form

$$\mathbb{P}^n(K) = \mathbb{A}^n(K) \amalg \mathbb{P}^{n-1}(K) = \mathbb{A}^n(K) \amalg \mathbb{A}^{n-1}(K) \amalg \cdots \amalg \mathbb{A}^1(K) \amalg \mathbb{A}^0(K).$$

(b) We will probably not seriously need $\mathbb{P}^n(K)$ for $n \geq 3$ in this course, but it may be helpful to note that much of what follows will apply with minor modifications to the general case.

(c) If you have seen barycentric coordinates before, then homogeneous coordinates on projective space $\mathbb{P}^2$ can be thought of as barycentric coordinates with the coordinate (or pure weight) points at $[1:0:0], [0:1:0]$, and $[0:0:1]$.

All faults of affine geometry are rectified in projective geometry, and this boils down to basic linear algebra/matrix theory; as an example, let's see how to fix 1.3.2.

**Example 1.3.7.** The "projectivization" of the lines of 1.3.2 are the lines

$$sY = X \text{ and } X = Z$$

in $\mathbb{P}^2(K)$. These two *always* have a unique intersection point, namely $[s:1:s]$, in $\mathbb{P}^2(K)$. Note that for $s = 0$, this is the point $[0:1:0]$ at infinity along the vertical direction, as expected.

**Example 1.3.8.** The "projectivization" of the parametrization in §1.2.A is the map $\mathbb{P}^1 \to \mathbb{P}^2$ given by

$$[S:T] \mapsto [S^2 - T^2 : 2ST : S^2 + T^2],$$

with image $X^2 + Y^2 = Z^2$. (Why is this map well-defined? For this to be true, we have to check that $S^2 - T^2, 2ST, S^2 + T^2$ are never simultaneously zero. This requires our base field to have characteristic other than 2.) When $\operatorname{ch} K \neq 2$, this gives us a *genuine* bijection (even isomorphism) between $\mathbb{P}^1(K)$ and $C(K)$, where $C = \mathbb{V}(X^2 + Y^2 - Z^2) \subset \mathbb{P}^2_K$. This way of looking at it solves various mysteries: for instance, the unique point mapping to $P = (-1, 0) = [-1:0:1]$ is precisely $[S:T] = [0:1]$, also known as "$t = \infty$". Similarly, over $K = \mathbb{C}$ the points $[1:\pm i]$ map to the points $[1:\pm i:0]$ in $\mathbb{P}^2(\mathbb{C})$, known as the "circular points at infinity"; these are not visible over $\mathbb{Q}$ or $\mathbb{R}$.

Let us study solutions of polynomial equations on projective spaces; for simplicity, we will stick to $n = 3$ and use homogeneous coordinates $X, Y, Z$. On projective space, the vanishing locus of an arbitrary $F \in k[X, Y, Z]$ does not make sense in general: if $F = X^2 - Y$, for instance, then $(1, 1, 0)$ lies on $F$ but $(2, 2, 0)$ does not. However, the one case in which this problem does *not* arise is when $F$ is *homogeneous*, i.e., when all monomials appearing in $F$ have the same degree, say $d \in \mathbb{Z}_{\geq 0}$ (c.f. 2.1.11). In this case we have

$$F(\lambda X, \lambda Y, \lambda Z) = \lambda^d F(X, Y, Z)$$

for $\lambda \in K^\times$, so it makes sense to speak of the *curve* $C = \mathbb{V}(F)$ cut out by $F$ in projective space. In other words, $C$ consists of $K$-points

$$C(K) := \{[X:Y:Z] \in \mathbb{P}^2(K) : F(X, Y, Z) = 0\}.$$

**Example 1.3.9.** If $F = Y^2 Z - X^3 - Z^3$ and $C = \mathbb{V}(F) \subset \mathbb{P}^2_K$, then $\mathscr{O} := [0:1:0] \in C(K)$ but $[1:0:0] \notin C(K)$.

Clearly, replacing $F$ by a nonzero scalar multiple $\mu F$ for $\mu \in K^\times$ is not going to change the vanishing locus, and in this way we get an identification between curves and nonzero homogeneous polynomials in $K[X, Y, Z]$ up to scaling. We take this to be our definition.

> **Definition 1.3.10.** A *projective plane curve* $C$ over a field $K$ is an equivalence class of nonzero homogeneous polynomials $F \in K[X, Y, Z]$ under the equivalence relation of scaling: $F$ and $\mu F$ for $\mu \in K^\times$ define the same curve $C = \mathbb{V}(F) = \mathbb{V}(\mu F)$. A representative $F$ of this equivalence class is said to be a *defining equation* of the curve. The degree of any defining equation is then called the *degree* of the curve.

We emphasize that this is an *algebraic* definition; the curve is determined by its equation and not its set of $K$-points, of which there may not be any (e.g., $C = \mathbb{V}(X^2 + Y^2 + Z^2)$ over $K = \mathbb{R}$). A curve of degree one (resp. two, three, four, etc.) is called a *line* (resp. *conic*, *cubic*, *quartic*, etc.). However, it is still natural to think of the curve $C$–to a first approximation–in terms of its $K$-points $C(K)$.

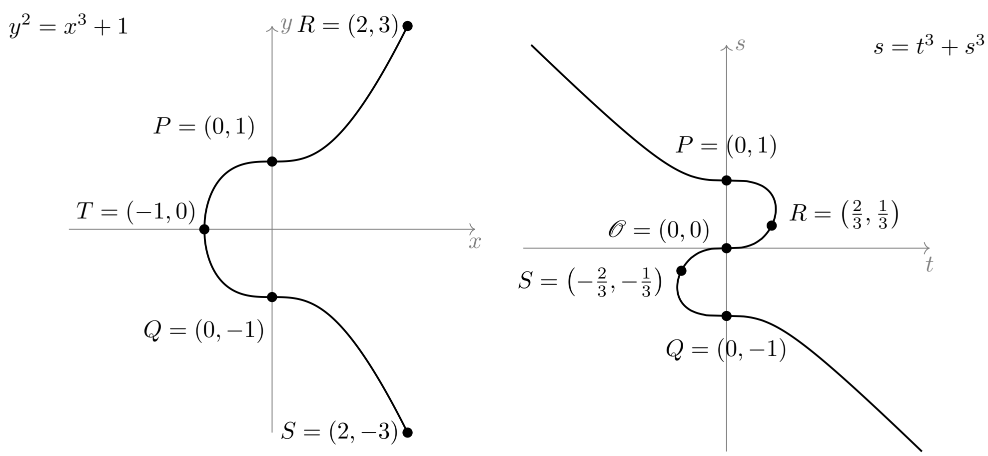
How do we visualize a projective curve $C$? Let us now briefly discuss something we have seen in practice already. Given a projective curve $C = \mathbb{V}(F) \subset \mathbb{P}^2_K$ over some field $K$ and of degree $d \in \mathbb{Z}_{\geq 1}$, we can dehomogenize it with respect to any of the variables $X, Y$, or $Z$ to get an affine curve, where dehomogenization, say, with respect to the variable $Z$ amounts to either dividing throughout by $Z^d$ and using coordinates $x := X/Z$ and $y := Y/Z$, or equivalently looking at points in the affine patch $Z \neq 0$, where we can normalize to $Z = 1$. Algebraically, this amounts to replacing $F \in K[X, Y, Z]$ by $f(x, y) := F(x, y, 1) \in K[x, y]$. In other words,

$$
\begin{aligned}
C(K) \cap \{Z \neq 0\} &= \{[X : Y : Z] \in \mathbb{P}^2(K) : F(X, Y, Z) = 0 \text{ and } Z \neq 0\} \\
&= \{(x, y) \in \mathbb{A}^2(K) : F(x, y, 1) = 0\} \\
&= \{(x, y) \in \mathbb{A}^2(K) : f(x, y) = 0\}.
\end{aligned}
$$

There is nothing special about $Z$ here, and we can equally well dehomogenize with respect to say $Y$, by taking coordinates $t := X/Y$ and $s := Z/Y$ instead. These different "local patches" of the curve can then be glued together to get a fuller picture of the projective curve.

**Example 1.3.11.** Consider the projective curve $C := \mathbb{V}(Y^2Z - X^3 - Z^3) \subset \mathbb{P}^2_K$ (for any field $K$). The affine patch $Z \neq 0$ looks like the familiar curve $y^2 = x^3 + 1$, and the affine patch $Y \neq 0$ looks like $s = t^3 + s^3$. The first patch has all points on $C$ except for $\mathscr{O} = [0 : 1 : 0]$; the second patch has all points except for $[-\omega^j : 0 : 1]$ for $j = 0, 1, 2$, where $\omega = \zeta_3$ is a primitive cube root of unity. (Some of these point may not exist in $C(K)$. If $K$ has characteristic three, there is only one point $[-1 : 0 : 1]$.) These two patches are related by the coordinate change

$$
(t, s) = \left(\frac{x}{y}, \frac{1}{y}\right) \text{ and } (x, y) = \left(\frac{t}{s}, \frac{1}{s}\right).
$$




(a) The affine chart $Z \neq 0$. (b) The affine chart $Y \neq 0$.

Figure 1.3.12: Two affine patches of the curve $Y^2Z = X^3 + Z^3$ with a few marked points. Note that the point $\mathscr{O}$ is not visible in the first patch, and the point $T$ is not visible in the second patch, but together they cover everything. Note also that the line at infinity in the first patch becomes the $t$-axis $s = 0$ in the second patch, and that this line has a "triple" intersection with the curve at $\mathscr{O}$, i.e., is an inflectional tangent to $C$ at $\mathscr{O}$.

I would recommend playing with Desmos3D and plotting $Y^2Z = X^3 + Z^3$ and seeing how it intersects the loci $Z = 1$, $Y = 1$, and $X^2 + Y^2 + Z^2 = 1$ to see how these pictures are related to each other.

Going in the other direction, given an affine curve $C \subset \mathbb{A}^2_K$ given by the vanishing of say $f(x, y) \in K[x, y]$, we can look at its *projective closure*. This corresponds to replacing $f$ by its *homogenization* $F \in K[X, Y, Z]$ obtained by replacing $x$ by $X/Z$ and $y$ by $Y/Z$ and clearing denominators, or (equivalently) by replacing $x$ by $X$ and $y$ by $Y$, and inserting suitable powers of $Z$ in the monomials to get a homogeneous polynomial of lowest possible degree. This projective closure is usually denoted by $\overline{C} \subset \mathbb{P}^2_K$. We can recover $C$ from $\overline{C}$ by dehomogenizing with respect to $Z$ once again.

**Example 1.3.13.** The line $2x - 5y + 3 = 0$ in $\mathbb{A}^2_K$ has projective closure $2X - 5Y + 3Z = 0$ in $\mathbb{P}^2_K$. To obtain the points at infinity on this projective closure, we set $Z = 0$ to get the unique point $[5 : 2 : 0]$ in the direction of this line. Make sure this makes sense to you!

**Example 1.3.14.** The rectangular hyperbola $x^2 - y^2 = 1$ in $\mathbb{A}^2_K$ has projective closure $X^2 - Y^2 = Z^2$ in $\mathbb{P}^2_K$. To obtain the points at infinity of this projective closure, we set $Z = 0$ to get the two points $[1 : \pm 1 : 0]$. Of course, this is as expected: these correspond to the asymptotes of the hyperbola.

**Exercise 1.3.15.** Are the operations of homogenization and dehomogenization inverses?