Math 445
Handy facts since the second exam

Don't forget the handy facts from the first two exams!

Rational points on curves

For more general curves, defined by polynomials $f(x, y) = 0$ of higher degree, looking at how lines meet the curve can provide a wealth of information.

Notation: $\mathcal{C}_f(\mathbb{R}) = \{(x, y) \in \mathbb{R}^2 : f(x, y) = 0\}$ For the most part, we will focus on *cubic* polynomials f, whose highest degree monomial is x^3, x^2y, xy^2, and/or y^3.

If L is a line (generically, defined by an equation $ax + by + c = 0$, a or $b \neq 0$), which we usually write $y = mx + b = L(x)$ (if not vertical: $x = c$), any point on both L and $\mathcal{C}_f(\mathbb{R})$ satisfies $p(x) = f(x, mx + b) = 0$. This is a polynomial of degree $d =$ the (total) degree of f. It therefore has exactly d (complex) roots (counting multiplicity), *unless* it is identically 0. So if L meets $\mathcal{C}_f(\mathbb{R})$ in more than d points (counting multiplicity; we will consider this shortly), then $p \equiv 0$, i.e., $P \in L$ implies $P \in \mathcal{C}_f(\mathbb{R})$, i.e., $L \subseteq \mathcal{C}_f(\mathbb{R})$. Even more, if L is defined by $L(x, y) = ax + by + c = 0$ and L meets $\mathcal{C}_f(\mathbb{R})$ in more than d points, then $f(x, y) = L(x, y)K(x, y)$ for some polynomial K (of degree $d - 1$).

More generally, we can refine this by considering points in $\mathcal{C}_f(\mathbb{R})$ with *multiplicity*. In the one-variable case, a point a is a solution to $p(x) = 0$ with multiplicity m if a is a root of both p <u>and</u> the first $m-1$ derivatives of p (this is equivalent to $p(x)$ having factor $(x-a)^m$). In the multivariable case, $P = (a, b)$ is a root of $f(x, y)$ with multiplicity m if P is a root of both f and every partial derivative $(\partial/\partial x)^i(\partial/\partial y)^j(f)$ for $i+j < m$. For the most part we will worry about multiplicity 2, i.e., P is a root of $f, \partial f/\partial x$, and $\partial f/\partial y$. Such a point is called a *double point* of $\mathcal{C}_f(\mathbb{R})$. More generally, a point of a curve $\mathcal{C}_f(\mathbb{R})$ of multiplicity greater than 1 is called a *singular point*. A curve with no singular points is called *smooth*.

As with quadratic curves, we can use knowledge of some rational points in $\mathcal{C}_f(\mathbb{R})$, for f a cubic polynomial, to find more, using lines. The idea now is to use the line through <u>two</u> solutions to find a third. Such a line will have rational slope, given by an equation $y = L(x) = mx + b$. If we look to solve the polynomial $p(x) = f(x, L(x)) = 0$, our two points provide two solutions; the third root (which we can find by factoring) will give us the third solution (plugging into $y = L(x)$ to recover its y-value). This is known as the *chord method*.

If the cubic curve $\mathcal{C}_f(\mathbb{R})$ has a double point P, we don't even need a second point; P will serve for both (so long as it has rational coordinates!). Even line with rational slope through P will give a third, rational, point; in fact, every rational solution can be found this way (remembering the line with infinite slope...).

If we only know one rational point $P = (a, b)$ in $\mathcal{C}_f(\mathbb{R})$, we can still find a line for which $p(x) = f(x, L(x)) = 0$ has a double root; the tangent line to $\mathcal{C}_f(\mathbb{R})$, is defined by the equation $f_x(P)(x-a)+f_y(P)(y-b) = 0$, which implies (via the chain rule) $p(a) = p'(a) = 0$, i.e., a is a double root. [If f has rational coefficients, this line has rational slope.] The third root then gives us a new rational point in $\mathcal{C}_f(\mathbb{R})$. This method is known as the *tangent method*.

Projective space.

There are some situations when this approach seems to break down; for example with an equation like

$$f(x,y) = y^2 - (x^3 - 5x + 3)$$

the line through the solutions $(1,1)$ and $(1,-1)$ (i.e., the vertical line $x = 1$), meets $\mathcal{C}_f(\mathbb{R})$ in only two points. [Plug in $x = 1$ to verify this.] It is going to be very important to us, however, that this approach not break down, and so we will take the (at the moment somewhat irrational) step of "inventing" new solutions, to cover these cases. The idea is to think of our solutions as living in a larger space; real projective space $\mathbb{P}_2(\mathbb{R})$. The idea is to first *projectivize* our equation, replacing $f(x,y) = 0$ with the *homogeneous* equation

$$F(X,Y,Z) = Z^3 f(X/Z, Y/Z) = 0$$

For example, our equation becomes $Y^2Z - (X^3 - 5XZ^2 + 3Z^3) = 0$. Such an equation has the property that (X,Y,Z) is a solution implies (aX, aY, aZ) for any a, i.e., solutions are "really" lines of solutions through the origin. $\mathbb{P}_2(\mathbb{R})$ is nothing more than this; it is the set of all lines through the origin in $\mathbb{R}^3$; since exact values of the coordinates are unimportant, we write points in $\mathbb{P}_2(\mathbb{R})$ as $X : Y : Z$ rather than (X,Y,Z). Since any solution to $F(X,Y,Z) = 0$ can be replaced with a constant multiple, any solution with $Z \neq 0$ has a corresponding solution with $Z = 1$. But if $X : Y : 1$ is a solution, then $f(X,Y) = 0$, i.e., it gives an ordinary solution in $\mathcal{C}_f(\mathbb{R})$. The solutions with $Z = 0$ do not have any corresponding solutions in $\mathcal{C}_f(\mathbb{R})$, as we have originally interpreted it; they are our extra solutions "at infinity" in $\mathbb{P}_2(\mathbb{R})$. They are found by projectivizing f, and setting $Z = 0$. In our example above, the point $0 : 1 : 0$ is a solution. It is the third point on our vertical line. It can in fact be interpreted as the vertical line; points in $\mathbb{R}^2$ correspond to lines $X : Y : Z$ with $Z \neq 0$, by looking at where the line meets the plane $Z = 1$ in $\mathbb{R}^3$. The points $X : Y : 0$, on the other hand, are lines in the XY-plane in $\mathbb{R}^3$, with slope Y/X. So $0 : 1 : 0$ corresponds to the vertical line in the XY-plane. In general, $a : b : 0$ is the point in $\mathbb{P}_2(\mathbb{R})$ where all lines of slope b/a meet!

Elliptic curves.

The type of curve where these tools prove the most useful are the *elliptic curves*. A cubic curve $\mathcal{C}_f(\mathbb{R})$ is called elliptic if it has no singular point (in $\mathbb{P}_2(\mathbb{R})$), and f has no linear factor (i.e., $\mathcal{C}_f(\mathbb{R})$ contains no line). The quickest test for this is to verify both of these properties over the complex numbers $\mathbb{C}$; and for this, we have the useful fact that

The polynomial $f(x,y) = y^2 - q(x)$ (q cubic) defines an elliptic curve over $\mathbb{C}$ if and only if q has no repeated root (over $\mathbb{C}$). For such a curve, we have that any line through two points A, B of $\mathcal{C}_f(\mathbb{R})$ intersects $\mathcal{C}_f(\mathbb{R})$ in a unique third point (in $\mathbb{P}_2(\mathbb{R})$), which find as above. We denote this third point AB. [When $A = B$, it is understood that the line meant is the tangent line to $\mathcal{C}_f(\mathbb{R})$ at A.] But this turns out to be, by itself, not terribly useful as a binary operation; it is, for example, not associative. [It has the useful properties, however, that AB=BA, and AB=C implies AC=B and BC=A.] To make a useful operation, we proceed as follows.

Pick any point in $\mathcal{C}_f(\mathbb{R})$; call it $\underline{0}$. Then given $A, B \in \mathcal{C}_f(\mathbb{R})$, we first find AB as above, and then find $\underline{0}(AB)$, and call it $\underline{0}(AB) = A + B$. This product, it turns out, is much more well behaved:

(1) $A+B=B+A$ for all $A,B\in\mathcal{C}_f(\mathbb{R})$
(2) $A+\underline{0}=A$ for all $A\in\mathcal{C}_f(\mathbb{R})$
(3) For every $A\in\mathcal{C}_f(\mathbb{R})$, there is a unique $B\in\mathcal{C}_f(\mathbb{R})$ with $A+B=\underline{0}$
[In fact, $B=A(\underline{00})$.]
(4) For all $A,B,C\in\mathcal{C}_f(\mathbb{R})$, $(A+B)+C=A+(B+C)$

The last fact is the most involved to verify; it use the fact:

If f and g are cubic polynomials, f has no linear factor, $P_1,\dots,P_9$ are distinct points in $\mathcal{C}_f(\mathbb{R})\cap\mathcal{C}_g(\mathbb{R})$ and P_1,P_2,P_3 lie in a line L, then there is a quadratic polynomial $q(x,y)$ so that $P_4,\dots,P_9\in\mathcal{C}_q(\mathbb{R})$.

[Typically, six points in the plane do not lie on a quadratic (other than the zero polynomial).]

When we apply this to the points $B, BC, C, AB, \underline{0}, \underline{0}(AB), A, \underline{0}(BC), (\underline{0}(AB))C$, and the polynomial $g(x,y)=L_1(x,y)L_2(x,y)L_3(x,y)$, where the line L_1 contains B, AB, A, L_2 contains $BC, \underline{0}, \underline{0}(BC)$, and L_3 contains $C, \underline{0}(AB), (\underline{0}(AB))C$, we find that the last six points lie on a quadratic. But the first three of these lie on a line, and so the last three do, as well. This implies that

$$(\underline{0}(AB))C=A(\underline{0}(BC))$$

and so $(A+B)+C=\underline{0}((\underline{0}(AB))C)=\underline{0}(A(\underline{0}(BC)))=A+(B+C)$. This argument really only applies if the nine points above are actually distinct. When they are not, we perturb the points $\underline{0},A,B,C$ slightly to make the nine points distinct, and apply "continuity".

Taken together, the four properties (1) through (4) tell us that $\mathcal{C}_f(\mathbb{R})$ is an *abelian group* under $+$. If we choose $\underline{0}$ to be a rational point (i.e., $\underline{0}\in\mathcal{C}_f(\mathbb{Q})$), then $\mathcal{C}_f(\mathbb{Q})$ forms a *subgroup* of $\mathcal{C}_f(\mathbb{R})$.

We can see that the actual choice of $\underline{0}$ certainly effects the definition of the addition, but it does not have a big effect on the structure of the resulting group; if we choose a different point $\underline{0}'$ and define $A\oplus B=\underline{0}'(AB)$, then $A\oplus B=A+B-\underline{0}'$, and therefore the function $\Phi:(\mathcal{C}_f(\mathbb{R}),\oplus)\to(\mathcal{C}_f(\mathbb{R}),+)$ defined by $\Phi(A)=A-\underline{0}'$, is an isomorphism of groups.

Focusing on elliptic curves of the form $f(x,y)=y^2-(x^3-ax-b)$ (which is all we will need for our applications), and using the point $0:1:0$ at infinity as $\underline{0}$, we can find explicit formulas for the addition of points. If $A=(x_1,y_1)$ and $B=(x_2,y_2)$, then (noting that $\underline{0}(x,y)=(x,-y)$) $A+B=\underline{0}(AB)$ will equal

$(m^2-x_1-x_2,-(y_1+m(m^2-2x_1-x_2)))$

where $m=\dfrac{y_2-y_1}{x_2-x_1}$, if $x_1\neq x_2$

$\underline{0}$

if $x_1=x_2$ and $y_1\neq y_2$

$(M^2-2x_1,-(y_1+M(M^2-3x_1)))$

where $M=\dfrac{3x_1^2-a}{2y_1}$, if $x_1=x_2$ and $y_1=y_2$

Factoring integers using elliptic curves

Elliptic curves have turned out to have many uses in the "real" world. We will look at one of them: providing the (to date) fastest known method to factor large integers. It uses the group operation on $\mathcal{C}_f(\mathbb{Q})$, and is based on the fact that for a finite group G, with order

n, every element $g \in G$ satisfies $n \cdot g = 0$.

Our approach, the Elliptic Curve Method, is modelled on another factoring algorithm due to Pollard, called the Pollard $(p-1)$-test. The idea is that if N is a (large) integer, with prime factor p, then by Fermat, for any a relatively prime to p, $p | a^{p-1}-1$, and so the g.c.d. $(a^{p-1}-1, N) > 1$. As usual, the problem is that we don't know p, but for this test we guess that $p-1$ consists of a product of fairly small primes, and test (a^n-1, N) for n a (large) product of fairly small numbers, in an effort to find a g.c.d. that is both greater than 1 and less than N, giving us a proper factor of N. In practice, we start with a randomly chosen a, and a sequence of fairly small numbers r_n, like $r_n = n$. We then form the sequence $a_1 = a$, $a_2 = a_1^{r_1} = a^{r_1}$, $a_3 = a_2^{r_2} = a^{r_1 r_2}$, and inductively, $a_{i+1} = a_i^{r_i} = a^{r_1 \cdots r_i}$. We then compute $g_i = (a_i - 1, N)$. Noting that $a_i - 1 | a_{i+1} - 1$ for every i, and so $g_i | g_{i+1}$ for every i, we typically, compute the g.c.d.'s only occasionally (since we expect to get $g_i = 1$ for awhile). This process will always eventually stop, since for any prime divisor p of N, $p-1$ will divide $r_1 \cdots r_n = 1 \cdot 2 \cdots n$ for some n, so $g_n > 1$. The first time this happens, however, it might be that $g_n = N$, and so the test fails; we then restart with a different a. The typical amount of time it take for this method to find a factor is on the order of the size of the smallest among the set of largest prime factors of $p-1$, where p ranges among all of the prime factors of N. The problem: this could be fairly large!

The elliptic curve method attempts to get around this problem. The basic idea behind the method above is that we are attempting to express the identity element in $\mathbb{Z}_p^*$ (the group of units of $\mathbb{Z}_p$), as a power of some number a, where the power is a product of fairly small numbers. [The fun part is that we are doing this without actually choosing p first!] The problem is that we are not guaranteed a p where products of small numbers will work. The ECM takes this problem and translates it into a framework where it is much more likely to work, using elliptic curves mod p.

The basic idea is to take the machinery we have developed for computing on elliptic curves, and do all of the calculations mod p, for some (unknown!) prime dividing N. In practice, this really means we do the calculations mod N. The basic fact is that, using the formulas for addition we have above (and really, it works in general), we can work out an addition formula for points in what we choose to call $\mathcal{C}_f(\mathbb{Z}_p)$. The formulas involve division, but mod p, we simply carry these out by instead multiplying by the invers (which we find by the Euclidean algorithm). We still need to know that this form of addition on $\mathcal{C}_f(\mathbb{Z}_p)$ gives us a group; but from the formulas, the needed properties can be verified directly (including associativity!).

To implement the ECM to find a factor of an integer N, we pick an elliptic curve $\mathcal{C}_f(\mathbb{Z}_p)$ by choosing values for a and b, and a point A on the curve. [Usually this is done the other way around; pick a point you want on the curve, such as $A = (1, 1)$, and choose the values of a and b accordingly.] $\mathcal{C}_f(\mathbb{Z}_p)$ is a group of some finite (but unknown) order; the idea is that we expect that for some choices of a and b, it has order a product of small primes, and so a calculation like the one in the Pollard $(p-1)$-test will quickly succeed. But this is where the fun starts!

The idea is to compute high multiples $r_1 \cdots r_n A$ of a point; we do this as we dealt with high powers long ago, by repeated doubling, and then adding together the necessary powers of 2 to get $r_1 \cdots r_n$. Our calculations are supposed to be carried out mod p, but they can't be; we don't know p. So instead we carry them out mod N (while pretending we are computing in $\mathcal{C}_f(\mathbb{Z}_p)$). But this will not always work; not every integer has an inverse mod N. So in our calculations we might occasionally fail to be able to compute a step. But this is a good thing! We will fail, because the quantity we need to invert, $x_2 - x_1$, is not relatively prime to N, i.e., $(x_2 - x_1, N) > 1$ (or, when doubling, $((2y_1), N) > 1$). Unless this is a multiple of N (i.e., since we are computing mod N, $x_2 = x_1$ or $N|y_1$), we have found what we sought; a proper factor of N! In point of fact, this is what the method is designed to do; we don't even want to find the order of A in $\mathcal{C}_f(\mathbb{Z}_p)$, since the order of this group really has no relation to N, it can, in fact, be any number between $p+1-2\sqrt{p}$ and $p+1+2\sqrt{p}$. What we really want to do is to discover that we can't compute the order, because the formulas break down and finds a factor of N, before the computation finishes. The point is that by varying the curve, we should relatively quickly stumble across one for which $\mathcal{C}_f(\mathbb{Z}_p)$ would have the kind of order that would allow us to compute it, if the computation were not going to break down. The basic idea is to find a curve where the calculation breaks down fairly quickly, and so we typically limit the size of $r_1 \cdots r_n$ (to around $\sqrt{N}$, so it is at least the expected size of $\mathcal{C}_f(\mathbb{Z}_p)$ for p the smallest prime dividing N), and vary the curve.