Math 126 Number Theory

Prof. D. Joyce, Clark University

24 Apr 2006

Last time. Introduction to Pell equations. We saw some examples of how to find solutions to Pell equations using continued fractions.

Today. Theory of Pell equations. We'll see some theorems that show the methods work that we looked at last time.

We'll start with a theorem that shows how to find infinitely many solutions if you just have one to start with. Here's the statement of the theorem, followed by a lemma that we'll use to prove it, then we'll have the proof of the lemma and the theorem.

Theorem. If the equation $x^2 - dy^2 = 1$ has one solution, then it has infinitely many solutions, while if the equation $x^2 - dy^2 = -1$ has one solution, then not only does it have infinitely many solutions but also $x^2 - dy^2 = 1$ has infinitely many solutions.

Lemma. If (a, b) is a solution to the Diophantine equation

$$x^2 - dy^2 = c,$$

then solutions to the equations

$$x^2 - dy^2 = c^n$$

are recursively defined by

$$\begin{aligned} x_1 &= a \\ y_1 &= b \\ x_{n+1} &= ax_n + dby_n \\ y_{n+1} &= bx_n + ay_n \end{aligned}$$

Proof: We'll prove this inductively. The base case, $n = 1$ is given.

For the inductive step, we assume that (x_n, y_n) is a solution of $x^2 - dy^2 = c^n$, and we'll show that (x_{n+1}, y_{n+1}) is a solution of $x^2 - dy^2 = c^{n+1}$. One continued equation will do it.

$$\begin{aligned} & x_{n+1}^2 - dy_{n+1}^2 \\ = \ & (ax_n + dby_n)^2 - d(bx_n + ay_n)^2 \\ = \ & (a^2x_n^2 + 2dabx_ny_n + d^2b^2y_n^2) \\ & - d(b^2x_n^2 + 2abx_ny_n + a^2y_n^2) \\ = \ & a^2x_n^2 + 2dabx_ny_n + d^2b^2y_n^2 \\ & - db^2x_n^2 - 2abx_ny_n - a^2y_n^2 \\ = \ & a^2x_n^2 + d^2b^2y_n^2 - da^2y_n^2 - db^2x_n^2 \\ = \ & (a^2 - db^2)(x_n^2 - dy_n^2) \\ = \ & c \cdot c^n = c^{n+1} \end{aligned}$$

Q.E.D.

Proof of the theorem: Let (a, b) be a solution of the first equation $x^2 - dy^2 = 1$. Then the all the solutions (x_n, y_n) provided by the lemma are solutions of the same equation.

Now let(a, b) be a solution of the second equation $x^2 - dy^2 = -1$. Then (x_n, y_n) will be a solution of the first equation when n is even, but of the second equation when n is odd, since $(-1)^n$ is either 1 or -1 depending on the parity of n. Q.E.D.

Now we know how to get more solutions if we have one, but we still have to find one. That's where the continued fractions come in. We'll talk about that next time. But there's more we can get out of this theorem.

An ancient method for finding square roots, and Newton's method. Let's look at just the first step, going from $(a, b) = (x_1, y_1)$ to (x_2, y_2). Note that the formulas in the lemma tell us that

$$(x_2, y_2) = (a^2 + db^2, 2ab)$$

We can think of (a,b) as giving one approximation of $\sqrt{d}$ since the equation $a^2 - db^2 = c$ is equivalent to

$$\left(\frac{a}{b}\right)^2 = d + \frac{c}{b^2}.$$

That says $(a/b)^2$ is near d, so a/b is near $\sqrt{d}$. The second point (x_2, y_2) satisfies the same equation $x^2 - dy^2 = c$, so it's also an approximation of $\sqrt{d}$, but a much better one since the error is c/y_2^2 instead of c/b^2 and y_2 is larger than $y_1 = b$.

If we want to quickly find better approximations to $\sqrt{d}$, then rather than next computing (x_3, y_3), we can treat (x_2, y_2) as the starting point, that is, take it to be (a,b), and apply the process. When you do that, you'll get the (x_4, y_4) in one step. Then if you take (x_4, y_4) as your starting point, you'll next get (x_8, y_8). Thus, you'll skip over lots of intermediate approximations of $\sqrt{d}$ and quickly get to very good approximations.

So, the one step we're looking at is replacing $\frac{a}{b}$ by $\dfrac{a^2 + db^2}{2ab}$. We can rewrite this as $\frac{1}{2}(\frac{a}{b} + d\,\frac{b}{a})$. Let's use the single variable s for the number $\frac{a}{b}$. Then, our first approximation s of $\sqrt{d}$ gives us a second approximation $\frac{1}{2}(s + d/s)$. That's a reasonable improvement on the approximation and one that has been used since ancient times, perhaps even by the Old Babylonians 4000 years ago.

Here's a different way of deriving this improvement on the approximation. Suppose we have an approximation s for $\sqrt{d}$. If s is actually less than $\sqrt{d}$, that is, if s^2 is less than d, then d/s will be greater than $\sqrt{d}$; therefore their average $\frac{1}{2}(s + d/s)$ should be closer to $\sqrt{d}$. Likewise if s is actually greater than $\sqrt{d}$, then d/s will be less than $\sqrt{d}$, and, in that case too, their average should be closer to $\sqrt{d}$.

We can see what's going on graphically if we move on to the 1600s and use a little analytic geometry and calculus. Finding the square root of d is the same as solving the equation $x^2 = d$, and that, in turn, is the same as finding a root of the polynomial $f(x) = x^2 - d$. We'll graph this polynomial in class and see how Newton's method is used to solve it. I've got a brief introduction to Newton's method at `http://aleph0.clarku.edu/~djoyce/newton/` that we'll look at.

Editorial. Mathematics is really very well-connected. A topic like this, Pell's equation, which seems to belong to number theory, has strong connections to geometry, computation, linear algebra, and calculus. That's not unusual. The many connections among the subjects in mathematics are always there, but we tend to study mathematics one subject at a time. There are many good reasons for doing that, but when we do that, we hide the connections.