Mathematical Apology 5
$\sqrt{-1}$ in finite fields and sums of two squares

Professor John Butcher, The University of Auckland

Some positive integers cannot be written as the sum of the squares of two integers, for example 3, 6, 7, 11, 14, 15, 19 cannot; some integers can be written as the some of two squares in exactly one way, for example $2 = 1^2 + 1^2$, $5 = 2^2 + 1^2$, $8 = 2^2 + 2^2$, $10 = 3^2 + 1^2$; and some integers can be written as the sum of two squares in two or more different ways, for example $65 = 8^2 + 1^2 = 7^2 + 4^2$. In this apology we will try to find reasons behind these observations. But it pays us first to look a little away from the main question.

A field is a set of objects (known as numbers) on which the operations of $+$, $\times$, $-$ and $\div$ can be performed under certain rules. Most clubs have ethical standards required of its members and the Field Club is no exception. Both addition and multiplication must be commutative and associative, the distributive rule must hold, there must be a zero number which cannot be divided by and zero multiplied by anything must be zero and added to anything must leave it unchanged. There is some further fine print in the rules but I will mention only one extra requirement, because some applications for membership (for example, the application by the set $\{0, 1, 2, 3, 4, 5\}$, where addition and multiplication are defined modulo 6), have had to be declined because of the rule "there cannot be divisors of zero". In turning down the particular application from the 6 member set, it was pointed out that $2 \times 3 = 6 \equiv 0 \pmod 6$ and no exceptions can be made even for close relatives of existing club members.

The well-known fields of rational, real and complex numbers are the mainstays of this club and are always present to make up a quorum at meetings. Within the last couple of hundred years they were joined by some upstart "finite fields". These newcomers were discovered by Evariste Galois before he threw his life away in a duel in 1832 at the age of 20. There is a Galois Field with p^n members for any prime number p and any positive integer n. We will concentrate, however, just on the fields with a prime number of members.

The field with p members, denoted by $GF(p)$, is made up from the numbers $\{0, 1, 2, \ldots, p-1\}$ with addition and multiplication defined modulo p. There is obviously a negative of each number a, equal to $p - a$ if $a \neq 0$, and this makes subtraction possible. To find the reciprocal of a number, so as to perform division, solve the Diophantine equation $ax - yp = 1$, using the Euclidean algorithm. But even if we don't explicitly find a^{-1} for $a \neq 0$, we know it exists, because if we multiply a by each member of $\{1, 2, \ldots, p-1\}$ we can easily see that we get this same set back again (with the order changed) and one of the products must be 1. In passing, this enables us to prove one of the great theorems of mathematics, with the understated name of "Fermat's Little Theorem". We have seen that the sets $\{1, 2, \ldots, p-1\}$ and $\{1a, 2a, \ldots, (p-1)a\}$ are the same and hence the products of their members are the same. This means that $1 \cdot 2 \cdot \cdots \cdot (p-1) = 1 \cdot 2 \cdot \cdots \cdot (p-1)a^{p-1}$, and cancelling the factors 1, 2, …, $p - 1$ from both sides assures us that

$$a^{p-1} \equiv 1 \pmod p.$$

One way of interpreting the Fermat theorem is to consider the polynomial equation

$$x^{p-1} - 1 = 0, \qquad (1)$$

in the field $GF(p)$ as having each of the $p-1$ numbers $\{1, 2, \ldots, p-1\}$ as solutions. If p is an *odd* prime, then we can factorize (1) as

$$x^{p-1} - 1 = (x^{(p-1)/2} - 1)(x^{(p-1)/2} + 1),$$

so that exactly half of the members of $\{1, 2, \ldots, p-1\}$ satisfy

$$x^{(p-1)/2} = 1 \tag{2}$$

and the remainder satisfy

$$x^{(p-1)/2} = -1. \tag{3}$$

While we have the set $\{1, 2, \ldots, p-1\}$ in our hands, with $p \neq 2$, let us square every member. We get only half as many distinct squares because a and $p-a$ have the same square. The non-zero members of $GF(p)$ which are in this set of squares, are known as "quadratic residues", and the remainder are known as "non-residues". If $x = t^2$, then

$$x^{(p-1)/2} = t^{p-1} = 1,$$

so that for the quadratic residues it is (2) that is satisfied, rather than (3). Because there are exactly the same numbers of solutions to these two equations, and exactly the same number of quadratic residues and non-residues, it follows that if x a non-residue it satisfies (3).

This enables us to answer a question suggested by the title of this apology. When does $\sqrt{-1}$ exist in $GF(p)$? In other words, for which odd primes p is -1 (or, equivalently, $p-1$) a perfect square modulo p?

From our brief introduction to quadratic residues, we see that everything hinges on the value of $(-1)^{(p-1)/2}$. But this is easy to calculate. If the prime p has the form $p = 4n+1$, then -1 is a quadratic residue, because $(-1)^{(p-1)/2} = (-1)^{2n} = +1$. If p has the form $4n+3$, then it is a non-residue, because $(-1)^{(p-1)/2} = (-1)^{2n+1} = -1$.

Finally, we return to the question of the possible solutions of the Diophantine equation

$$x^2 + y^2 = N.$$

Suppose that N is equal to a prime p. We already know the answer if $p = 2$, because $2 = 1^2 + 1^2$. If $p = 4n+3$, there is no hope because the squares of x and y are each congruent to either 0 or 1 (mod 4) and therefore the total of x^2 and y^2 cannot be 3 (mod 4).

If $p = 4n+1$ we can find a number t in $GF(p)$ such that $t^2 = -1$. This means if we choose any non-zero x in $GF(p)$ and define $y = tx$, then $x^2 + y^2 = 0$. If we now interpret x as an integer, rather than as a member of $GF(p)$, and restrict its value so that $0 < x^2 < p$, then y can be interpreted as the remainder when tx is divided by p. With this interpretation, we know that $x^2 + y^2 = np$, for some positive integer n, and we want to prove that $n = 1$ arises at least once. Write x_1, x_2, …, x_k for the values of x for which we do this calculation, where k is the integer part of $\sqrt{p}$, and let y_1, y_2, …, y_k denote the corresponding values of y. If the values of y_i, $i = 1, 2, \ldots, k$ are sorted into increasing order, and the numbering adjusted to reflect this ordering, then we have

$$0 < y_1 < y_2 < \cdots < y_k < p.$$

Because

$$(y_1 - 0) + (y_2 - y_1) + (y_3 - y_2) + \cdots + (p - y_k) = p,$$

it follows that at least one of the terms on the left-hand side must be less than $\sqrt{p}$ because the arithmetic mean of these terms is

$$\frac{p}{k+1} < \frac{p}{\sqrt{p}} = \sqrt{p}.$$

If $y_1 < \sqrt{p}$, then $x_1^2 + y_1^2 = p$ because this sum of squares is a multiple of p and is less than $2p$. If, for some $i = 2, 3, \dots, k-1$, $y_i - y_{i-1} < \sqrt{p}$, it similarly follows that $(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2 = p$ and if $p - y_k < \sqrt{p}$, then $x_k^2 + (p - y_k)^2 = p$.

Having concluded that $x^2+y^2 = p$ is possible only if $p = 2$ or $p \equiv 1 \pmod 4$, it seems reasonable to leave the question of asking when $x^2 + y^2 = N$ has solutions for N a *composite* positive integer, as an exercise. The following identities should be a help

$$\begin{aligned} 2(x^2 + y^2) &= (x+y)^2 + (x-y)^2, \\ (x_1^2 + y_1^2)(x_2^2 + y_2^2) &= (x_1x_2 \pm y_1y_2)^2 + (x_1y_2 \mp y_1x_2)^2. \end{aligned}$$

Author

Emeritus Professor John Butcher, Department of Mathematics, The University of Auckland, Private Bag 92019, Auckland. Phone (09) 3737599, extn 8747, Fax (09) 3737457,
e-mail butcher@math.auckland.ac.nz