Sums of two squares revisited

I was at one time an amateur numerical analyst; then I found, almost by chance, that I could make a living out of it. However, I am still an amateur number theorist with no expectation of ever getting anything from it except sheer enjoyment. Hence, I am displaying my arrogance by again letting number theory find its way into these Miniatures. However, my lack of any deep knowledge will force me to stick to questions accessible to other amateurs. In Miniature number 7, I showed using a pigeon-hole proof, that primes of the form $4n+1$ can be written as sums of two squares. Today I will give a constructive argument drawn to my attention by Alf van der Poorten of Macquarie University. But is the 'Hermite-Serret' algorithm, as it is known, really constructive? It depends on whether you regard the discovery of an $x \in \{1, 2, \ldots, p-1\}$ such that $x^2 \equiv -1 \pmod{p}$ to be a trival problem or not, because this is the starting point for the construction. Start with two relatively prime positive integers, a_0 and a_1 with $a_0 > a_1$ and with the property that there exists b_1 such that $a_0 b_1 = a_1^2 + 1$. Of course a_0 has the role of p, but is not necessarily assumed to be prime, and a_1 has the role of x. In this slightly generalized version of the Hermite-Serret construction, use the Euclidean algorithm to form a sequence a_0, a_1, a_2, …, a_{m+1} and identify the first two members of the sequence, say a_k, a_{k+1} which are each less than $\sqrt{a_0}$. Then the theorem belonging to the algorithm, states that $a_k^2 + a_{k+1}^2 = a_0$. Let n_1, n_2, …, n_m denote the quotients arising in the Euclidean algorithm so that

$$a_{i+1} = a_{i-1} - n_i a_i, \qquad i = 1, 2, \ldots, m.$$

Along with the a sequence, introduce a b sequence that satisfies exactly the same relationship between successive members, starting with $b_0 = a_1$ and with the b_1 already introduced. Thus, we have

$$\begin{bmatrix} a_i & a_{i+1} \\ b_i & b_{i+1} \end{bmatrix} = \begin{bmatrix} a_{i-1} & a_i \\ b_{i-1} & b_i \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & -n_i \end{bmatrix}, \quad i = 1, 2, \ldots,$$

which we will write in the form $X_i = X_{i-1} N_i$. Because $\det(N_i) = -1$ and $\det(X_0) = a_0 b_1 - a_1 b_0 = a_0 b_1 - a_1^2 = 1$, it follows that $a_i b_{i+1} - a_{i+1} b_i = \det(X_i) = (-1)^i$. Coming back to the n sequence we see that

$$\frac{a_0}{a_1} = \frac{a_0}{b_0} = \begin{bmatrix} n_1, & n_2, & \ldots, & n_m \end{bmatrix} = n_1 + \frac{1}{n_2+} \frac{1}{n_3+} \cdots \frac{1}{n_m},$$

where we can force the length m of the continued fraction to be even because if $n_m > 1$ then

$$\begin{bmatrix} n_1, & n_2, & \ldots, & n_m \end{bmatrix} = \begin{bmatrix} n_1, & n_2, & \ldots, & n_m - 1, & 1 \end{bmatrix}.$$

Let $k = m/2$. We will show that $a_{k-1} > \sqrt{a_0} > a_k$ and that $a_k^2 + a_{k+1}^2 = a_0$, thus justifying the Hermite-Serret algorithm.

It is useful to look at the *reversed* continued fraction $\begin{bmatrix} n_m, & n_{m-1}, & \ldots, & n_1 \end{bmatrix}$. The numerator and denominator of convergent number m are the elements in the first row of the matrix product

$$\begin{bmatrix} n_m & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} n_{m-1} & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} n_{m-2} & 1 \\ 1 & 0 \end{bmatrix} \cdots \begin{bmatrix} n_1 & 1 \\ 1 & 0 \end{bmatrix}.$$

Thus, if convergent number m is N/D, then

$$\begin{bmatrix} N & D \end{bmatrix} = \begin{bmatrix} 1 & 0 \end{bmatrix} N_m^{-1} N_{m-1}^{-1} N_{m-2}^{-1} \cdots N_1^{-1},$$

or, what is equivalent,

$$\begin{bmatrix} 1 & 0 \end{bmatrix} = \begin{bmatrix} N & D \end{bmatrix} N_1 N_2 N_3 \cdots N_m,$$

implying that $N = a_0$ and $D = a_1$. Thus, the value of $\begin{bmatrix} n_m, & n_{m-1}, & \ldots, & n_1 \end{bmatrix}$ is a_0/a_1 and the reversed sequence is identical to the forward sequence. Now calculate $a_k^2 + a_{k+1}^2$. We find

$$a_k^2 + a_{k+1}^2 = \begin{bmatrix} a_k & a_{k+1} & v \end{bmatrix} \begin{bmatrix} a_k & a_{k+1} \end{bmatrix}^T = \begin{bmatrix} a_0 & a_1 \end{bmatrix} N_1 N_2 \cdots N_k N_k^T N_{k-1}^T \cdots N_1^T \begin{bmatrix} a_0 & a_1 \end{bmatrix}^T.$$

Because each of the N matrices is symmetric and because the sequence of such matrices is palandromic, this expression can be written as

$$\begin{bmatrix} a_0 & a_1 \end{bmatrix} N_1 N_2 \cdots N_k N_{k+1} N_{k+2} \cdots N_m \begin{bmatrix} a_0 & a_1 \end{bmatrix}^T = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} a_0 & a_1 \end{bmatrix}^T = a_0.$$

The inequality $\sqrt{a_0} > a_k$ is an immediate consequence and $a_{k-1} > \sqrt{a_0}$ follows from $a_{k-1}^2 = (n_k a_k + a_{k+1})^2 > a_k^2 + a_{k+1}^2 = a_0$. There is room for a single example: $a_0 = 29$, $a_1 = 17$. The continued fraction is $\begin{bmatrix} 1, & 1, & 2, & 2, & 2 \end{bmatrix}$, which is stretched to the palandromic sequence of even length $\begin{bmatrix} 1, & 1, & 2, & 2, & 1, & 1 \end{bmatrix}$. The a and b sequences are $\begin{bmatrix} 29, & 17, & 12, & 5, & 2, & 1, & 1, & 0 \end{bmatrix}$ and $\begin{bmatrix} 17, & 10, & 7, & 3, & 1, & 1, & 0, & 1 \end{bmatrix}$, giving a sequence of approximations to $29/17$ and the solution to the sum of squares problem: $29 = 5^2 + 2^2$.

John Butcher, butcher@math.auckland.ac.nz