MIDPOINT CRITERIA FOR SOLVING PELL'S EQUATION USING THE NEAREST SQUARE CONTINUED FRACTION

KEITH MATTHEWS, JOHN ROBERTSON, AND JIM WHITE

Abstract. We derive midpoint criteria for solving Pell’s equation \( x^2 - Dy^2 = \pm 1 \), using the nearest square continued fraction expansion of \( \sqrt{D} \). The period of the expansion is on average 70% that of the regular continued fraction. We derive similar criteria for the diophantine equation \( x^2 - xy - (D-1)y^2 = \pm 1 \), where \( D \equiv 1 \pmod{4} \). We also present some numerical results and conclude with a comparison of the computational performance of the regular, nearest square and nearest integer continued fraction algorithms.

1. Introduction

Euler gave two midpoint criteria for solving Pell’s equation \( x^2 - Dy^2 = \pm 1 \) using the regular continued fraction (RCF) expansion of \( \sqrt{D} \) (see [4, p. 358]). Suppose the simple continued fraction expansion for \( \sqrt{D} \) is periodic with period \( k \):

\[
\sqrt{D} = \begin{cases} \left[ a_0, a_1, \ldots, a_{h-1}, a_h, a_{h-1}, \ldots, a_1, 2a_0 \right] & \text{if } k = 2h - 1, \\
\left[ a_0, a_1, \ldots, a_{h-1}, a_h, a_{h-1}, \ldots, a_1, 2a_0 \right] & \text{if } k = 2h. 
\end{cases}
\]

Also if \( (P_n + \sqrt{D})/Q_n \) denotes the \( n \)-th complete quotient of the RCF expansion of \( \sqrt{D} \) and if \( Q_h = Q_{h-1} \), then \( k = 2h - 1 \), while if \( P_h = P_{h+1} \), then \( k = 2h \). Consequently we can detect the end of the half period. The smallest solution of \( x^2 - Dy^2 = \pm 1 \) is given by

\[
\eta = A_{k-1} + B_{k-1} \sqrt{D},
\]

where \( A_n/B_n \) is the \( n \)-th convergent to \( \sqrt{D} \). Euler observed that if \( k = 2h - 1 \), then

\[
A_{k-1} = A_{h-1}B_{h-1} + A_{h-2}B_{h-2}, \\
B_{k-1} = B_{h-1}^2 + B_{h-2}^2,
\]

while if \( k = 2h \), then

\[
A_{k-1} = A_{h-1}B_h + A_{h-2}B_{h-1}, \\
B_{k-1} = B_{h-1}(B_h + B_{h-2}).
\]

H.C. Williams and P.A. Buhr [13] gave six midpoint criteria for the nearest integer continued fraction of B. Minnegerode [7] and A. Hurwitz [5]. In our paper, we give three midpoint criteria in terms of the nearest square continued fraction.

Received by the editor July 29, 2008 and, in revised form, March 15, 2009.

2000 Mathematics Subject Classification. Primary 11D09, 11Y50, 11A55, 11Y65.

Key words and phrases. Pell’s equation, nearest square continued fraction.

©2009 American Mathematical Society
Reverts to public domain 28 years from publication.

485
2. Nearest square continued fraction

This continued fraction was introduced by A.A.K. Ayyangar in 1940 and 1941 (see [2], [3]) and arose from Bhaskara’s cyclic method (circa 1150) for solving Pell’s equation (see [1]). Let \( \xi_0 = \frac{P_0 + \sqrt{D}}{Q_0} \) be a surd in standard form; i.e., \( D \) is a non-square positive integer and \( P_0, Q_0 \neq 0 \), \( \frac{P_0 - P^2}{Q_0} \) are integers, having no common factor other than 1. Then with \( c = \lfloor \xi_0 \rfloor \), the integer part of \( \xi_0 \), we can represent \( \xi_0 \) in two ways:

\[
\xi_0 = c + \frac{Q'}{P' + \sqrt{D}} = c + 1 - \frac{Q''}{P'' + \sqrt{D}},
\]

where \( \frac{P' + \sqrt{D}}{Q'} > 1 \) and \( \frac{P'' + \sqrt{D}}{Q''} > 1 \) are also standard surds. We choose the partial denominator \( \alpha_0 \) and numerator \( \epsilon_1 \) of the new continued fraction expansion as follows:

(a) \( \alpha_0 = c \) if \( |Q'| < |Q''| \), or \( |Q'| = |Q''| \) and \( Q < 0 \), \( \epsilon_1 = 1 \),

(b) \( \alpha_0 = c + 1 \) if \( |Q'| > |Q''| \), or \( |Q'| = |Q''| \) and \( Q > 0 \), \( \epsilon_1 = -1 \).

The term nearest square arises on restating (a) and (b):

(a') \( \alpha_0 = c \) if \( \frac{|P' - D|}{|Q'|} < \frac{|P'' - D|}{|Q''|} \), or \( |P' - D| = |P'' - D| \) and \( Q < 0 \),

(b') \( \alpha_0 = c + 1 \) if \( \frac{|P' - D|}{|Q'|} > \frac{|P'' - D|}{|Q''|} \), or \( |P' - D| = |P'' - D| \) and \( Q > 0 \).

Then \( \xi_0 = \alpha_0 + \frac{\epsilon_1}{\xi_1} \), where \( \lfloor \epsilon_1 \rfloor = 1, \alpha_0 \) an integer and \( \xi_1 = \frac{P' + \sqrt{D}}{Q'} > 1 \). Also \( P_1 = P' \) or \( P'' \) and \( Q_1 = Q' \) or \( Q'' \), according as \( \epsilon_1 = 1 \) or \( -1 \). We proceed similarly with \( \xi_1 \), and so on. Then

\[
(2.1) \quad \xi_n = a_n + \frac{\epsilon_{n+1}}{\xi_{n+1}} \quad \text{and} \quad \xi_0 = a_0 + \frac{\epsilon_1}{a_1} + \frac{\epsilon_2}{a_2} + \cdots.
\]

This expansion is called the nearest square continued fraction (NSCF).

Analogous relations to those for regular continued fractions hold for \( P_n, Q_n \) and \( a_n \):

\[
(2.2) \quad P_{n+1} + P_n = a_n Q_n,
\]

\[
(2.3) \quad P_{n+1}^2 + \epsilon_{n+1} Q_n Q_{n+1} = D.
\]

By Theorem I (iii) [3, p. 22], the \( |Q_n| \) successively diminish as long as \( |Q_n| > \sqrt{D} \) and so, ultimately, we have \( |Q_n| < \sqrt{D} \). When this stage is reached, the \( P_m \) and \( Q_m \) thereafter become positive and bounded, \( 0 < P_m < 2\sqrt{D}, 0 < Q_m < \sqrt{D} \) by Theorem I (iv) [3, p. 22]. This implies eventual periodicity of the complete quotients and thence the partial quotients. In particular, Theorem XII ([3 pp. 102-103]) shows that the NSCF expansion of \( \sqrt{D} \) has the form

\[
(2.4) \quad \sqrt{D} = a_0 + \frac{\epsilon_1}{a_1} + \cdots + \frac{\epsilon_k}{a_k} + \frac{\epsilon_{k+1}}{a_{k+1}} + \cdots
\]

where the asterisks denote that the period-length is \( k \) and \( \xi_p = \xi_{p+k}, \epsilon_p = \epsilon_{p+k} \) and \( a_p = a_{p+k} \) for \( p > 1 \). (It’s an easy exercise to show that \( a_0 = \lfloor \sqrt{D} \rfloor \), the nearest integer to \( \sqrt{D} \).) In [3 pp. 112-114], the finer structure of (2.4) is revealed. There are two types:

**Type I:** No complete quotient of a cycle has the form \( \frac{P + q + \sqrt{p^2 + q^2}}{p} \), where \( p > 2q > 0, \gcd(p, q) = 1 \). This type possesses the classical symmetries of
the regular continued fraction if \( k > 1 \):

\[
\begin{align*}
a_v &= a_{k-v}, \quad 1 \leq v \leq k-1, \\
Q_v &= Q_{k-v}, \quad 1 \leq v \leq k-1, \\
\epsilon_v &= \epsilon_{k+1-v}, \quad 1 \leq v \leq k, \\
P_v &= P_{k+1-v}, \quad 1 \leq v \leq k.
\end{align*}
\]

For example, \( \sqrt{19} = 4 + \frac{\frac{1}{2}}{\frac{1}{3} + \frac{\frac{1}{2}}{\frac{1}{5} + \frac{1}{3} + \frac{\frac{1}{2}}{\frac{1}{5} + \frac{1}{3} + \cdots}}} \)

**Type II:** One complete quotient \( \xi_v \) in a cycle has the form \( \frac{p+q+\sqrt{p^2+q^2}}{p} \), where \( p > 2q > 0 \). In this case \( k \geq 4 \) is even and \( v = k/2 \). This type also possesses the symmetries of Type I, apart from

\[ a_{\frac{k}{2}} = 2, \quad \epsilon_{\frac{k}{2}} = -1, \quad \epsilon_{\frac{k+1}{2}} = 1, \quad a_{\frac{k}{2}-1} = a_{\frac{k+1}{2}} + 1, \quad P_{\frac{k}{2}} \neq P_{\frac{k+1}{2}} \]

and we have

\[
\sqrt{D} = a_0 + \frac{\epsilon_1}{a_1} + \cdots + \frac{\epsilon_{\frac{k}{2}-1}}{a_{\frac{k}{2}-1}} + \frac{1}{2} + \frac{1}{2 \left( a_{\frac{k}{2}-1} - 1 \right)} + \cdots + \frac{\epsilon_k}{2a_0}
\]

For example, \( \sqrt{29} = 5 + \frac{1}{3} - \frac{1}{2} + \frac{1}{2} + \frac{1}{10} \). Other examples are \( D = 53, 58, 85, 97 \).

For both types of \( D \), we have \( Q_k = 1 \). For \( P_1 = a_0, \quad P_1 = P_k \) (symmetry), \( P_1 = P_{k+1} \) (periodicity), so

\[
2a_0 = 2P_1 = P_k + P_{k+1} = a_kQ_k \quad \text{by (2.2)}
\]

\[
= 2a_0Q_k.
\]

Hence \( Q_k = 1 \) and \( \xi_k = a_0 + \sqrt{D} \). This is needed later in the proof of Lemma 1.

Similarly, the quadratic surd \( \xi_0 = (1 + \sqrt{D})/2, \quad D = 4n + 1, \) has \( a_0 = [\xi_0] \). Also \( a_k = 2a_0 - 1 = P_1 \) and

\[
2P_1 = P_k + P_{k+1} = a_kQ_k = P_1Q_k.
\]

Hence \( Q_k = 2 \) and \( \xi_k = (2a_0 - 1 + \sqrt{D})/2 \).

3. Reduced NSCF Quadratic Surds

Ayyangar [3] p. 27] gives a definition of reduced quadratic surd that is not as explicit as for regular continued fractions (see e.g. [8] p. 73]). He defines a special surd \( \xi_v \) by the inequalities

\[
Q_{v+1}^2 + \frac{1}{4}Q_v^2 \leq D, \quad Q_v^2 + \frac{1}{4}Q_{v+1}^2 \leq D,
\]

then defines a semi-reduced surd to be the successor of a special surd. Finally a reduced surd is defined to be the successor of a semi-reduced surd. He proves ([3] p. 28]) that a reduced surd is a special surd and in [3] pp. 101-102] that a quadratic surd has a purely periodic NSCF if and only if it is reduced. Examples of reduced surds that figure prominently in [3] are (i) \( \frac{p+q+\sqrt{p^2+q^2}}{p} \), where \( p > 2q > 0 \) and (ii) the successor of \( \sqrt{D} \).
4. Midpoint properties of Types I and II NSCF expansions of $\sqrt{D}$

(a) Type I: If $k = 2h$, then $P_h = P_{h+1}$.
(b) Type I: If $k = 2h + 1$, then $Q_h = Q_{h+1}$.
(c) Type II: Here $k = 2h$, $Q_{h-1}$ is even, $\epsilon_h = -1$ and $P_h = Q_h + \frac{1}{2}Q_{h-1}$. Also $P_v \neq P_{v+1}$ and $Q_v \neq Q_{v+1}$ for $1 \leq v < 2h$.

(See [3 pp. 110-114].) There are converses: Assume $k > 1$ and $1 \leq v < k$. Then

(d) $P_v = P_{v+1} \implies k = 2h$, $v = h$ and a Type I NSCF expansion.
(e) $Q_v = Q_{v+1} \implies k = 2h + 1$, $v = h$ and a Type I NSCF expansion.
(f) $Q_{v-1}$ even, $\epsilon_v = -1$ and $P_v = Q_v + \frac{1}{2}Q_{v-1} \implies k = 2h$, $v = h$ and a Type II NSCF expansion.

Proof. (d) is proved in [3 p. 111]: Suppose $P_v = P_{v+1}$. Then we know we are dealing with a Type I NSCF expansion and hence $Q_{k-v} = Q_v$. Then

$$\xi_{k-v} = \frac{P_{k-v} + \sqrt{D}}{Q_{k-v}} = \frac{P_{v+1} + \sqrt{D}}{Q_v} = \frac{P_v + \sqrt{D}}{Q_v} = \xi_v,$$

so $k - v = v$ and $k = 2v$.

(e) is similar.

(f) Assume $Q_{v-1}$ even, $\epsilon_v = -1$ and $P_v = Q_v + \frac{1}{2}Q_{v-1}$. Then

$$D = P_v^2 + \epsilon_v Q_v Q_{v-1} = P_v^2 - Q_v Q_{v-1} = (Q_v + \frac{1}{2}Q_{v-1})^2 - Q_v Q_{v-1} = Q_v^2 + \frac{1}{4}Q_{v-1}^2 = p^2 + q^2,$$

where $p = Q_v$, $q = \frac{1}{2}Q_{v-1}$. Also $\gcd(p, q) = 1$. Next, because $\xi_v$ is reduced, it is a special surd ([3 p. 217]), so $Q_{v-1}^2 + 2Q_v^2 \leq D$. Hence

$$Q_{v-1}^2 + \frac{1}{2}Q_v^2 \leq Q_v^2 + \frac{1}{4}Q_{v-1}^2,$$

$$\frac{1}{2}Q_v^2 \leq \frac{1}{4}Q_{v-1}^2.$$

But $Q_v = Q_{v-1}$ implies $p = 2q$, so $p = 2$, $q = 1$, $D = 5$ and $\xi_v = \frac{1 + \sqrt{5}}{2}$. However this implies $k = 1$, so we deduce $p > 2q$ and $\xi_v$ has the form $\frac{p + \sqrt{D}}{2q}$, where $p > 2q > 0$. Hence we are dealing with a Type II NSCF expansion with $k = 2h$ and $v = h$. □

5. The convergents and Pell’s equation

As in [12 p. 406], we define the convergents $A_n/B_n$ by $A_{-2} = 0$, $A_{-1} = 1$, $B_{-2} = 1$, $B_{-1} = 0$ and for $i \geq -1$,

$$A_{i+1} = a_{i+1}A_i + \epsilon_{i+1}A_{i-1},$$
$$B_{i+1} = a_{i+1}B_i + \epsilon_{i+1}B_{i-1}.$$

An important property of the convergents to $\xi_0 = \frac{P_0 + \sqrt{D}}{Q_0}$ is

$$(Q_0 A_n - P_0 B_n)^2 - DB_n^2 = (-1)^{n+1}\epsilon_1\epsilon_2\cdots\epsilon_{n+1}Q_{n+1}$$

(see [12 (3.3) p. 406 and (3.5) p. 407]). For $\xi_0 = \sqrt{D}$, this reduces to

$$A_n^2 - DB_n^2 = (-1)^{n+1}\epsilon_1\epsilon_2\cdots\epsilon_{n+1}Q_{n+1}.$$
Hence, as \( Q_k = 1 \), we have

\[
A_k^2 - DB_k^2 = (-1)^k \epsilon_1 \epsilon_2 \cdots \epsilon_k. 
\]

**Remark.** Similarly, from (5.1), the convergents to \((1 + \sqrt{D})/2, D \equiv 1 \pmod{4}\)

satisfy

\[
A_n^2 - A_nB_n - \frac{(D-1)}{4} B_n^2 = (-1)^{n+1} \epsilon_1 \epsilon_2 \cdots \epsilon_n Q_{n+1}/2. 
\]

Hence, as \( Q_k = 2 \), we have

\[
A_{k-1}^2 - A_{k-1}B_{k-1} - \frac{(D-1)}{4} B_{k-1}^2 = (-1)^k \epsilon_1 \epsilon_2 \cdots \epsilon_k. 
\]

**Lemma 1.** In the NSCF expansion of \( \sqrt{D} \), with period-length \( k \), \( Q_n = 1 \) if and only if \( k \) divides \( n \).

**Proof.** We have seen that \( Q_k = 1 \). So suppose \( Q_n = 1, n \geq 1 \). Then from (2.3),

\[ P_n^2 + \epsilon_n Q_{n-1} = D. \]

Case 1. \( P_n > \sqrt{D} \). Then \( \epsilon_n = -1 \). Hence

\[ P_n^2 - D = Q_{n-1} < \sqrt{D} \]  \( (\xi_n \text{ is reduced}), \)

\[ 0 < P_n - \sqrt{D} < \frac{\sqrt{D}}{P_n + \sqrt{D}} < \frac{\sqrt{D}}{2\sqrt{D}} = \frac{1}{2}. \]

Hence \( P_n = \lfloor \sqrt{D} \rfloor \).

Case 2. \( P_n < \sqrt{D} \). Then \( \epsilon_n = 1 \). Hence

\[ Q_{n-1}^2 + 4Q_n^2 \leq D = P_n^2 + Q_{n-1} \]  \( (\xi_n \text{ is reduced}), \)

\[ (Q_{n-1} - 4)^2 \leq P_n^2, \]

\[ Q_{n-1} - 4 \leq P_n, \]

\[ Q_{n-1} \leq P_n + 1, \]

\[ D - P_n^2 = Q_{n-1} - P_n. \]

Hence \( 0 < \sqrt{D} - P_n \leq \frac{P_n}{\sqrt{D} + P_n} = \frac{P_n}{P_n} = 1 \) and again \( P_n = \lfloor \sqrt{D} \rfloor \).

Thus in both cases, \( \xi_n = a_0 + \sqrt{D} = \xi_k \) and \( k \) divides \( n \). \( \square \)

In the next section we prove that there is no smaller positive integer solution \((x, y)\) of the equation \( x^2 - Dy^2 = \pm 1 \) than \((A_{k-1}, B_{k-1})\) by showing that \( x/y \) is a convergent in the NSCF expansion of \( \sqrt{D} \).

**Remark.** Similarly, in the NSCF expansion of \((1 + \sqrt{D})/2, D \equiv 1 \pmod{4}\), with period-length \( k \), \( Q_n = 2 \) if and only if \( k \) divides \( n \).

6. **Relations between the NSCF and RCF**

In [8] pp. 147-155, Perron introduces a transformation \( t_1 \) of the following NSCF (with trivial modification when \( \lambda = 0 \)):

\[
\xi_0 = a_0 + \frac{\epsilon_1}{a_1} + \cdots + \frac{\epsilon_\lambda}{a_\lambda} + \frac{1}{a_{\lambda+1}} + \frac{\epsilon_{\lambda+2}}{a_{\lambda+2}} + \cdots ,
\]

expanding it to

\[
\xi_0 = a_0 + \frac{\epsilon_1}{a_1} + \cdots + \frac{\epsilon_\lambda}{a_\lambda - 1} + \frac{1}{a_{\lambda+1} - 1} + \frac{1}{a_{\lambda+1} - 1} + \frac{\epsilon_{\lambda+2}}{a_{\lambda+2}} + \cdots .
\]
The overall result of applying \( t_1 \) at all occurrences of \( \epsilon_\lambda = -1 \) is a transformation \( \Xi_1 \), given by the rule: Before a negative partial numerator the term \( \frac{+1}{1} \) is inserted.

Also each \( a_v \) is replaced by

- (a) \( a_v \) if \( \epsilon_v = +1 \), \( \epsilon_{v+1} = +1 \),
- (b) \( a_v - 1 \) if \( \epsilon_v = +1 \), \( \epsilon_{v+1} = -1 \), or \( \epsilon_v = -1 \), \( \epsilon_{v+1} = +1 \),
- (c) \( a_v - 2 \) if \( \epsilon_v = -1 \), \( \epsilon_{v+1} = -1 \).

Here \( \epsilon_0 = +1 \).

The partial quotients corresponding to an NSCF reduced quadratic surd are greater than 1 (\([3, \text{p. 29}]\)). So in view of Lemma 2 below and (b) and (c) above, \( \Xi_1 \) will convert the NSCF expansion of \( \sqrt{D} \) into an RCF expansion.

**Lemma 2.** Suppose \( \xi_v \) and \( \xi_{v+1} \) are NSCF reduced quadratic surds. Then if \( \epsilon_v = -1 \) and \( \epsilon_{v+1} = -1 \), we have \( a_v \geq 3 \).

**Proof.** Assume \( \xi_v \) and \( \xi_{v+1} \) are reduced. Then from [3, p. 27], we have

\[
P_{v+1} \geq Q_v + \frac{1}{2} Q_{v+1},
\]

\[
P_v \geq Q_v + \frac{1}{2} Q_{v-1}.
\]

Then \([6.3]\) and \([6.4]\) give

\[a_v Q_v = P_{v+1} + P_v \geq 2Q_v + \frac{1}{2} Q_{v+1} + \frac{1}{2} Q_{v-1}.
\]

Hence \( a_v Q_v > 2Q_v \), as \( Q_{v+1} > 0 \) and \( Q_{v-1} > 0 \). Hence \( a_v > 2 \). \( \square \)

**Lemma 3.** The period length of the RCF expansion of \( \sqrt{D} \) is \( k + r \), where \( r \) is the number of \( \epsilon_v = -1 \) occurring in the period partial numerators \( \epsilon_1, \ldots, \epsilon_k \) of the NSCF expansion of \( \sqrt{D} \).

**Proof.** If \( r = 0 \), there is nothing to prove. So we assume \( r > 0 \). According to \([8, \text{Satz 5.9, p. 152}]\), under \( \Xi_1 \),

(i) \( \epsilon_{v+1} = -1 \) gives rise to RCF convergents

\[A_{m-1}'B_{m-1}' = (A_v - A_{v-1})/(B_v - B_{v-1}), \quad A_m' / B_m' = A_v / B_v \]

and RCF complete quotients

\[\frac{P_m' + \sqrt{D}}{Q_m'} = \xi_{v+1}/(\xi_{v+1} - 1), \quad \frac{P_{m+1}' + \sqrt{D}}{Q_{m+1}'} = \xi_{v+1} - 1.\]

(ii) \( \epsilon_{v+1} = 1 \) gives rise to the RCF convergent \( A_v / B_v \) and the RCF complete quotient \( \xi_{v+1} \).

Consequently the NSCF complete quotients \( \xi_1, \ldots, \xi_k \) will give rise to an RCF period \( \xi_1', \ldots, \xi_{k+r} \) of complete quotients. We prove that this is a least period.

This will follow by showing that \( \xi_i/(\xi_i - 1) = a + \sqrt{D} \) is impossible, for \( a + \sqrt{D} \) is RCF-reduced and hence \( a = [\sqrt{D}] \). We can assume \( D > 3 \). Then \( a > 1 \) and

\[
\xi_i = \frac{a + \sqrt{D}}{a - 1 + \sqrt{D}} = 1 + \frac{1}{a - 1 + \sqrt{D}} = 2 - \frac{a - 2 + \sqrt{D}}{a - 1 + \sqrt{D}}
\]

\[
= 2 - \frac{D - (a - 2)^2}{D - (a - 1)(a - 2) + \sqrt{D}}.
\]

Then \( Q_{i+1}' = D - (a - 2)^2 = D - a^2 + 4(a - 1) > 1 = Q_{i+1}' \).
Hence (5.5) is the NSCF expansion of $\xi_i$. But partial denominators of such a reduced surd are at least 2 (see [3, Corollary 4, p. 29]), so we have a contradiction. So under $\mathcal{T}_1$, the NSCF complete quotients $\xi_1, \ldots, \xi_k$ will produce a period of RCF complete quotients $\xi'_m = \frac{P'_m + \sqrt{D}}{Q'_m}, 1 \leq m \leq k + r$, where $Q'_m > 1$ if $1 \leq m < k + r$ and $Q'_{k+r} = 1$. Consequently this is a least period of the RCF expansion of $\sqrt{D}$. □

**Lemma 4.** If $x^2 - Dy^2 = \pm 1$, $x, y > 0$, then $x/y$ is an NSCF convergent to $\sqrt{D}$.

**Proof.** Since $x/y$ is an RCF convergent $A'_{m-1}/B'_{m-1}$ to $\sqrt{D}$, so

$$A'^2_{m-1} - DB'^2_{m-1} = (-1)^m Q'_m = (-1)^m.$$

If $x/y$ is not an NSCF convergent of $\sqrt{D}$, it has the form $(A_n - A_{n-1})/(B_n - B_{n-1})$, where $\epsilon_{n+1} = -1$. However this would imply $\xi_{n+1}/(\xi_{n+1} - 1) = \frac{P'_n + \sqrt{D}}{Q'_n}$ and we have seen that this is impossible. Hence $x/y$ is an NSCF convergent to $\sqrt{D}$. □

**Remark.** The diophantine equation $x^2 - xy - \frac{(D-1)}{4}y^2 = \pm 1$, $D \equiv 1 \pmod{4}$ is also of interest. We can similarly show that if $D \geq 13$ and $x > 0, y > 0$, then $x/y$ is an NSCF convergent to $(1 + \sqrt{D})/2$.

7. **Midpoint criteria for determining $A_{k-1}$ and $B_{k-1}$**

Exactly one of the following will apply for any $D > 0$, not a square:

**P-test:** For some $h$, $1 \leq h < k$, $P_h = P_{h+1}$, in which case $k = 2h$ and

\[(7.1) \quad A_{k-1} = A_h B_{h-1} + \epsilon_h A_{h-1} B_{h-2},\]
\[(7.2) \quad B_{k-1} = B_{h-1}(B_h + \epsilon_h B_{h-2}).\]

In this case $A^2_{k-1} - DB^2_{k-1} = 1$.

**Q-test:** For some $h$, $0 \leq h < k$, $Q_h = Q_{h+1}$, in which case $k = 2h + 1$ and

\[(7.3) \quad A_{k-1} = A_h B_h + \epsilon_{h+1} A_{h-1} B_{h-1},\]
\[(7.4) \quad B_{k-1} = B^2_h + \epsilon_{h+1} B^2_{h-1}.\]

In this case $A^2_{k-1} - DB^2_{k-1} = -\epsilon_{h+1}$.

**PQ-test:** For some $h$, $1 \leq h < k$, $Q_{h-1}$ is even, $P_h = Q_h + \frac{1}{2}Q_{h-1}$ and $\epsilon_h = -1$, in which case $k = 2h$ and

\[(7.5) \quad A_{k-1} = A_h B_{h-1} - B_{h-2}(A_{h-1} - A_{h-2}),\]
\[(7.6) \quad B_{k-1} = 2B^2_{h-1} - B_h B_{h-2}.\]

In this case $A^2_{k-1} - DB^2_{k-1} = -1$.

Before we prove these statements, we restate the symmetry properties of the partial numerators and denominators of the NSCF expansion of $\sqrt{D}$ in the following form, for use in Lemma 5 below:

1. If $k = 2h + 1$ and $1 \leq t \leq h$, then

\[(7.7) \quad \epsilon_{h+1+t} = \epsilon_{h+1-t},\]
\[(7.8) \quad a_{h+t} = a_{h+1-t}.\]
(2) If \( k = 2h \) and Type I with \( 1 \leq t \leq h \) or Type II with \( 3 \leq t \leq h \), then

\[
\begin{align*}
\epsilon_{h+t} &= \epsilon_{h-t+1}, \\
a_{h+t-1} &= a_{h-t+1}.
\end{align*}
\]

Lemma 5. \( k = 2h + 1, h \geq 1 \). Then for Type I and \( 0 \leq t \leq h \), we have

\[
\begin{align*}
A_{2h} &= A_{h+1}B_{h-t} + \epsilon_{h+1+t}A_{h+t-1}B_{h-t-1}, \\
B_{2h} &= B_{h+1}B_{h-t} + \epsilon_{h+1+t}B_{h+t-1}B_{h-t-1}.
\end{align*}
\]

(ii) Let \( k = 2h, h \geq 1 \). Then for Type I and \( 0 \leq t \leq h \), or Type II with \( h \geq 2 \) and \( 2 \leq t \leq h \), we have

\[
\begin{align*}
A_{2h-1} &= A_{h+t-1}B_{h-t} + \epsilon_{h+t}A_{h+t-2}B_{h-t-1}, \\
B_{2h-1} &= B_{h+t-1}B_{h-t} + \epsilon_{h+t}B_{h+t-2}B_{h-t-1}.
\end{align*}
\]

Proof. We prove (7.11) by induction on \( t, h \geq 0 \). Let

\[
f(t) = A_{h+t}B_{h-t} + \epsilon_{h+1+t}A_{h+t-1}B_{h-t-1}.
\]

We show \( f(h) = A_{2h} \) and \( f(t) = f(t-1) \) if \( h \geq 1 \). First note that (7.11) holds when \( t = h \). For then

\[
f(h) = A_{2h}B_0 + \epsilon_{2h+1}A_{2h-1}B_{-1} = A_{2h}.
\]

Next

\[
f(t) = A_{h+t}B_{h-t} + \epsilon_{h+1+t}A_{h+t-1}B_{h-t-1} \\
= (a_{h+t}A_{h+t-1} + \epsilon_{h+t}A_{h+t-2}) + \epsilon_{h+1+t}A_{h+t-1}B_{h-t-1} \\
= A_{h+t-1}(a_{h+1}B_{h-t} + \epsilon_{h+t}B_{h-t-1}) + \epsilon_{h+t}A_{h+t-2}B_{h-t} \\
= A_{h+t-1}(a_{h+1}B_{h-t} + \epsilon_{h+1+t}B_{h-t-1}) + \epsilon_{h+t}A_{h+t-2}B_{h-t} \\
= A_{h+t-1}B_{h+t-1}A_{h+t-2}B_{h-t} = f(t-1).
\]

Similarly for equation (7.12).

Equations (7.13) and (7.14) are proved similarly using equations (7.9) and (7.10), noting that for Type II, we can assume \( h \geq 3 \), for \( h = 2 \), equations (7.13) and (7.14) are trivially true. \( \square \)

The \( P \)-test: Substituting \( t = 0 \) in (7.13) gives

\[
A_{2h-1} = A_{h-1}B_h + \epsilon_h A_{h-2}B_{h-1} = A_{h-1}(a_hB_{h-1} + \epsilon_h B_{h-2}) + (A_h - a_h A_{h-1})B_{h-1} = A_hB_{h-1} + \epsilon_h A_{h-1}B_{h-2},
\]

which is the first equation of the \( P \)-test. Substituting \( t = 0 \) in (7.14) gives

\[
B_{2h-1} = B_{h-1}B_h + \epsilon_h B_{h-2}B_{h-1},
\]

which is the second equation of the \( P \)-test.

The \( Q \)-test: If \( k = 1 \), then equations (7.3) and (7.4) are trivially true. So we can assume \( k > 1 \). Then substituting \( t = 0 \) in (7.11) gives

\[
A_{2h} = A_h B_h + \epsilon_{h+1} A_{h-1} B_{h-1},
\]

which is the first equation in the \( Q \)-test. Substituting \( t = 0 \) in (7.12) gives

\[
B_{2h} = B_h B_h + \epsilon_{h+1} B_{h-1} B_{h-1},
\]

which is the second equation of the \( Q \)-test.
The **PQ-test**: We take \( t = 2 \) in equations (7.13) and (7.14) to get

\[
A_{2h-1} = A_{h+1}B_{h-2} + \epsilon_{h+2}A_{h}B_{h-3},
\]

(7.15)

\[
B_{2h-1} = B_{h+1}B_{h-2} + \epsilon_{h+2}B_{h}B_{h-3}.
\]

(7.16)

We also have

\[
\epsilon_{h} = -1, \epsilon_{h+1} = 1, a_{h+1} = a_{h-1} - 1, a_{h} = 2, \epsilon_{h+2} = \epsilon_{h-1}.
\]

(7.17)

Also \( B_{h-1} = a_{h-1}B_{h-2} + \epsilon_{h-1}B_{h-3} \). Hence (7.15) gives

\[
A_{2h-1} = (a_{h+1}A_{h} + \epsilon_{h+1}A_{h-1})B_{h-2} + \epsilon_{h-1}A_{h}B_{h-3}
\]

\[
= (a_{h+1}A_{h} + \epsilon_{h+1}A_{h-1})B_{h-2} + (B_{h-1} - a_{h-1})A_{h}
\]

\[
= (a_{h+1} - a_{h})A_{h}B_{h-2} + A_{h-1}B_{h-2} + B_{h-1}A_{h}
\]

\[
= -A_{h}B_{h-2} + A_{h-1}B_{h-2} + B_{h-1}A_{h}
\]

(7.18)

But \( A_{h} = a_{h}A_{h-1} + \epsilon_{h}A_{h-2} = 2A_{h-1} - A_{h-2} \). Hence

\[
A_{h} - A_{h-1} = A_{h-1} - A_{h-2}
\]

and (7.18) gives

\[
A_{2h-1} = A_{h}B_{h-1} - (A_{h-1} - A_{h-2})B_{h-2},
\]

which is the first equation of the PQ-test. Finally, (7.16) gives

\[
B_{2h-1} = (a_{h+1}B_{h} + \epsilon_{h+1}B_{h-1})B_{h-2} + \epsilon_{h-1}B_{h}B_{h-3}
\]

\[
= (a_{h+1}B_{h} + \epsilon_{h+1}B_{h-1})B_{h-2} + B_{h}(B_{h-1} - a_{h-1}B_{h-2})
\]

\[
= (a_{h+1} - a_{h})B_{h}B_{h-2} + B_{h-1}(\epsilon_{h+1}B_{h-2} + B_{h})
\]

(7.19)

But \( B_{h} = a_{h}B_{h-1} + \epsilon_{h}B_{h-2} = 2B_{h-1} - B_{h-2} \) by (7.17). Hence

\[
B_{h-2} + B_{h} = 2B_{h-1}.
\]

(7.20)

Then (7.19) and (7.20) give

\[
B_{2h-1} = -B_{h}B_{h-2} + 2B_{h-1}^2,
\]

which is the second equation of the PQ-test.

We now verify the third equation of each of the three tests. Recall equation (5.3):

\[
A_{k-1}^2 - DB_{k-1}^2 = (-1)^{k} \epsilon_1 \cdots \epsilon_k.
\]

(7.21)

Hence if \( k = 2h \) and \( D \) is of Type I, then (7.21) and the symmetries (7.9) and (7.10) give

\[
A_{k-1}^2 - DB_{k-1}^2 = 1,
\]

which corresponds to the third equation of the \( P \)-test.

Assume \( k = 2h + 1 \). If \( k = 1 \), then (a) \( D = t^2 + 1, t \geq 1 \) or (b) \( D = t^2 - 1, t > 1 \). In both cases, \( A_0 = 1, B_0 = 1 \), while in case (a), \( \epsilon_1 = 1 \) and case (b), \( \epsilon_1 = -1 \) and we see that the third equation of the \( Q \)-test is satisfied.

If \( k > 1 \), equation (7.21) and the symmetries (7.7) and (7.8) give

\[
A_{k-1}^2 - DB_{k-1}^2 = -\epsilon_{h+1},
\]

which again is the third equation of the \( Q \)-test.
Finally if $k = 2h$ and the continued fraction is of Type II, then $A_{2k-1}^2 - DB_{2k-1}^2 = -1,$ which corresponds to the third equation of the $PQ$-test.

Remark. If $\xi_0 = (1 + \sqrt{D})/2, D \equiv 1 \pmod{4},$ we also have $P, Q$ and $PQ$ tests, with the Pell equations replaced, as follows: If $k$ is the period-length, then

\[
P : A_{2k-1}^2 - A_{k-1}B_{k-1} - \frac{(D+1)}{2}B_{k-1}^2 = 1;
\]

\[
Q : A_{2k-1}^2 - A_{k-1}B_{k-1} - \frac{(D+1)}{2}B_{k-1}^2 = -\epsilon_{k+1}, k = 2h + 1;
\]

\[
PQ : A_{2k-1}^2 - A_{k-1}B_{k-1} - \frac{(D+1)}{2}B_{k-1}^2 = -1.
\]

The above proofs go through; except if $k = 1$, then (a) $D = t^2 + 4, t \geq 3, t$ odd or (b) $D = t^2 - 4, t \geq 3, t$ odd. In both cases, $A_0 = (t + 1)/2, B_0 = 1,$ while in case (a), $\epsilon_1 = 1$ and case (b), $\epsilon_1 = -1.$

Remark. The reduced period $\pi$ of Williams and Buhr ([13, p. 373]) can be shown to be equal to $k,$ the NSCF period-length of $\sqrt{D}.$ From [13, p. 374], the following hold:

(a) Conditions 1 and 2 satisfy the $P$-test for $\sqrt{D};$

(b) Conditions 3, 4 and 5 satisfy the $Q$-test for $\sqrt{D};$

(c) Condition 6 satisfies the $PQ$-test for $\sqrt{D}.$

8. Numerical results

In Table 1 we give the frequency of occurrence of each of three criteria for the NSCF expansion of $\sqrt{D}$ for non-square $D \leq M.$

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|}
\hline
$M$ & $P$-test & $Q$-test & $PQ$-test & Total \\
\hline
100 & 60 & 25 & 5 & 90 \\
1000 & 762 & 165 & 42 & 969 \\
10000 & 8252 & 1266 & 382 & 9900 \\
100000 & 85856 & 10465 & 3363 & 99684 \\
1000000 & 878243 & 90533 & 30224 & 999000 \\
10000000 & 8915623 & 805295 & 275920 & 9996838 \\
\hline
\end{tabular}
\caption{Frequency of $P, Q$ and $PQ$ criteria for $\sqrt{D}, D \leq M.$}
\end{table}

In Table 2, $D = 97$ and the NSCF expansion of $\sqrt{97}$ is of type II, with period length 6. There are 5 negative $\epsilon_i$’s in the period range $1 \leq i \leq 6$ and the period length of the RCF expansion of $\sqrt{97}$ is 11.

In Table 3, we compare $\pi(D)$ and $p(D),$ the respective periods of the NSCF and RCF expansions of $\sqrt{D},$ where $D$ is not a perfect square. We let

$$
\Pi(n) = \sum_{D \leq n} \pi(D), \quad P(n) = \sum_{D \leq n} p(D).
$$

Then it appears that $\Pi(n)/P(n) \to \tau = \log_2 \left( 1 + \sqrt{5} \right) = 0.6942419 \ldots$

The limiting behaviour in Table 3 was also observed for the nearest integer continued fraction by Williams and Buhr [13, p. 377] and Riesel [10, p. 260].
(8.2)

\[
X/Y = b_0 + \frac{\epsilon_1}{b_1} + \cdots + \frac{\epsilon_k}{[b_{k\pi(D)}^{-1}]^{-1}}
\]

By theorems of Heilbronn and Rieger (see [9, p. 159]), for \(D\) with a long RCF period, we expect the ratio \(\pi(D)/p(D)\) of the lengths of these finite continued fractions to approximate \(\tau\). For example, with \(D = 26437680473689\) (an example...
of Daniel Shanks [11] with a long RCF period) we have \( p(D) = 18331889, \pi(D) = 12726394, \pi(D)/p(D) = .6942216 \cdot \cdot \cdot \).

9. Computational Tests

We conclude with a comparison of the computational performance of continued fraction algorithms and consider the question of which of the three CF algorithms (RCF, NICF, NSCF) is the more computationally efficient for solving the Pell equation for any given value of \( D \). All tests were performed on a Sun Sparcv9 processor (750MHz). The programs were written in C and used the GMP (Version 4.2.4) multiple precision arithmetic library for convergent calculations.

Two versions of each algorithm were tested, a “standard” version and a “quotient-optimised” version. In both versions, we employ some fairly obvious optimisations such as developing only one convergent sequence \( B_n \), then solving directly for the corresponding \( A_n \) once only at the conclusion of the main loop. This typically halves the amount of processing that would be required if we had developed both sequences.

In the “standard” programs, the calculation of each \( B_n = a_nB_{n-1} \pm B_{n-2} \) is performed in two steps, a multiplication giving \( a_nB_{n-1} \) followed by the addition/subtraction of \( B_{n-2} \). The quotient-optimised versions use two distinct timesaving optimisations to this process. The first improvement is to introduce special handling of partial quotient values 1, 2 and 4. These can benefit from special handling, and also occur with sufficient frequency to make this well worthwhile.

In the RCF, for example, a partial quotient value of 1 occurs with average frequency 41.5%. Computing the new convergent thus requires only the addition of the previous two values, avoiding an unnecessary multiplication. In all three algorithms, partial quotient values of 2 and 4 can benefit from using the special GMP function for multiplication by powers of 2. This function uses shift instructions to perform the operation, and these are usually faster than normal multiplication.

For all other quotient values, improvement over the standard version is also obtained, by using a GMP function that gives a combined multiply-and-add (or subtract) operation, allowing the convergent calculation to be performed as a single step.

The first set of test results involves “short-period” tests. We processed all values of \( D \) in the range \([10^6(n-1), 10^6n]\) for \( n = 1 \) to 6. Table 4 lists the results obtained using the standard convergent method. Times are given in seconds, and for NICF and NSCF, the times relative to RCF are also given.

<table>
<thead>
<tr>
<th>( n )</th>
<th>RCF</th>
<th>NICF ratio</th>
<th>NSCF ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>81</td>
<td>64</td>
<td>.790</td>
</tr>
<tr>
<td>2</td>
<td>147</td>
<td>118</td>
<td>.803</td>
</tr>
<tr>
<td>3</td>
<td>200</td>
<td>160</td>
<td>.800</td>
</tr>
<tr>
<td>4</td>
<td>248</td>
<td>196</td>
<td>.790</td>
</tr>
<tr>
<td>5</td>
<td>291</td>
<td>230</td>
<td>.790</td>
</tr>
<tr>
<td>6</td>
<td>334</td>
<td>258</td>
<td>.772</td>
</tr>
</tbody>
</table>

We also ran a set of “long-period” tests. Here we processed specific values \( D_n \) of \( D \) with substantially long period lengths. The specific values and corresponding
period lengths are listed in Table 5. Note that the period length ratios in each case are all very close to the expected average of 0.694.

Table 5. Long-period \( D_n \) examples.

<table>
<thead>
<tr>
<th>( n )</th>
<th>( D_n )</th>
<th>RCF period-length</th>
<th>NSCF-NICF period-length</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( 10^{14} + 3 )</td>
<td>625024</td>
<td>433550</td>
</tr>
<tr>
<td>2</td>
<td>( 10^{14} + 7 )</td>
<td>869844</td>
<td>604092</td>
</tr>
<tr>
<td>3</td>
<td>( 10^{12} + 24 )</td>
<td>1005170</td>
<td>697848</td>
</tr>
<tr>
<td>4</td>
<td>( 10^{12} + 189 )</td>
<td>2064689</td>
<td>1433367</td>
</tr>
<tr>
<td>5</td>
<td>( 10^{12} + 294 )</td>
<td>2963566</td>
<td>2057350</td>
</tr>
</tbody>
</table>

Table 6 lists the solution times for each \( D_n \) in seconds, using the standard convergent method, with the ratios of times for NICF and NSCF relative to the corresponding RCF times.

Table 6. Long-period times for \( D_n \) (standard method).

<table>
<thead>
<tr>
<th>( n )</th>
<th>RCF</th>
<th>NICF</th>
<th>ratio</th>
<th>NSCF</th>
<th>ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>63</td>
<td>43</td>
<td>.683</td>
<td>43</td>
<td>.683</td>
</tr>
<tr>
<td>2</td>
<td>133</td>
<td>93</td>
<td>.699</td>
<td>93</td>
<td>.699</td>
</tr>
<tr>
<td>3</td>
<td>181</td>
<td>126</td>
<td>.696</td>
<td>126</td>
<td>.696</td>
</tr>
<tr>
<td>4</td>
<td>850</td>
<td>592</td>
<td>.696</td>
<td>589</td>
<td>.693</td>
</tr>
<tr>
<td>5</td>
<td>1797</td>
<td>1251</td>
<td>.696</td>
<td>1246</td>
<td>.693</td>
</tr>
</tbody>
</table>

With the standard method of convergent calculation, our main observations are these:

(a) there is no significant difference between NICF and NSCF;
(b) as period lengths increase, the relative performance of NSCF and NICF against RCF is increasingly close to the corresponding ratio of period lengths.

These results generally conform with expectations: as period lengths increase, the computational cost is increasingly dominated by the cost of calculating the convergents, with both NSCF and NICF performing exactly the same number of convergent calculation steps.

A different trend becomes evident, however, when the same tests are performed using the quotient-optimisation method. Tables 7 and 8 show the corresponding results using this method.

All three CF algorithms benefit substantially from quotient optimisation, but it is RCF that benefits the most. At short period lengths it now performs just as well as NICF or NSCF, and as period lengths increase it becomes noticeably faster. This can be explained by examining the average relative frequencies of quotient occurrences for the values in question.

Quotient values of 2 or 4 occur around 23% of the time for RCF, and 33% for NICF and NSCF. What tips the balance in favour of RCF is the 41.5% frequency of quotient value 1, which never occurs at all with NICF or NSCF. The optimisation for this particular case is also the one that is most beneficial, as it avoids multiplication altogether.
Table 7. Short-period times for $[10^6(n - 1), 10^6n]$ (quotient-optimised).

<table>
<thead>
<tr>
<th>n</th>
<th>RCF</th>
<th>NICF</th>
<th>ratio</th>
<th>NSCF</th>
<th>ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>56</td>
<td>54</td>
<td>.964</td>
<td>54</td>
<td>.964</td>
</tr>
<tr>
<td>2</td>
<td>102</td>
<td>101</td>
<td>.990</td>
<td>102</td>
<td>1.000</td>
</tr>
<tr>
<td>3</td>
<td>136</td>
<td>135</td>
<td>.993</td>
<td>135</td>
<td>.993</td>
</tr>
<tr>
<td>4</td>
<td>165</td>
<td>165</td>
<td>1.000</td>
<td>165</td>
<td>1.000</td>
</tr>
<tr>
<td>5</td>
<td>193</td>
<td>195</td>
<td>1.010</td>
<td>196</td>
<td>1.016</td>
</tr>
<tr>
<td>6</td>
<td>218</td>
<td>221</td>
<td>1.014</td>
<td>224</td>
<td>1.028</td>
</tr>
</tbody>
</table>

Table 8. Long-period times for $D_n$ (quotient-optimised).

<table>
<thead>
<tr>
<th>n</th>
<th>RCF</th>
<th>NICF</th>
<th>ratio</th>
<th>NSCF</th>
<th>ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>31</td>
<td>35</td>
<td>1.129</td>
<td>35</td>
<td>.964</td>
</tr>
<tr>
<td>2</td>
<td>65</td>
<td>72</td>
<td>1.108</td>
<td>73</td>
<td>1.123</td>
</tr>
<tr>
<td>3</td>
<td>90</td>
<td>100</td>
<td>1.111</td>
<td>100</td>
<td>1.111</td>
</tr>
<tr>
<td>4</td>
<td>398</td>
<td>468</td>
<td>1.176</td>
<td>467</td>
<td>1.173</td>
</tr>
<tr>
<td>5</td>
<td>880</td>
<td>986</td>
<td>1.120</td>
<td>987</td>
<td>1.122</td>
</tr>
</tbody>
</table>

Acknowledgement

We are grateful to Alan Offer for improving the continued fraction alignment and helping with the computations.

References

8. O. Perron, *Die Lehre von den Kettenbrüchen*, Band 1, Teubner, 1954. MR0064172 (16,239c)