Solving Pell's equation using the nearest square continued fraction

Keith Matthews

12th June 2008

Abstract

This talk is about the nearest square continued fraction of A.A.K. Ayyangar (1941) and its use in finding the smallest positive solution of Pell's equation $x^2 - Dy^2 = \pm 1$.

Contrary to the 1944 review by D.H. Lehmer, its ”slight blemishes” are indeed compensated by the its period length being about 70% of that of the regular continued fraction of $\sqrt{D}$ and also having a 3-case mid-point criterion for solving Pell's equation.

Hugh Williams and Peter Buhr (1979) gave a 6-case midpoint criterion in terms of Hurwitz' continued fraction of the first kind. Their paper was rather complicated.

It was after studying their paper that Jim White and John Robertson came up with a 3-case midpoint criterion using the nearest square continued fraction.

Euler (1765) gave a two-case *midpoint* criterion for solving Pell's equation $x^2 - Dy^2 = \pm 1$ using the regular continued fraction (RCF) expansion of $\sqrt{D}$.

$$\sqrt{D} = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cdots}}$$

where $a_0 = \lfloor \sqrt{D} \rfloor$ and $a_i \geq 1$ for all i.
We write

$$\sqrt{D} = [a_0, a_1, a_2, \ldots].$$

The n-th convergent is defined by

$$A_n/B_n = a_0 + \cfrac{1}{a_1 + \cfrac{1}{\ddots + \cfrac{1}{a_{n-1} + \cfrac{1}{a_n}}}}$$

A_n and B_n can be computed recursively:

$$A_0 = a_0, B_0 = 1, A_1 = a_0 a_1 + 1, B_1 = a_1,$$

$$\begin{aligned} A_{i+1} &= a_{i+1} A_i + A_{i-1} \\ B_{i+1} &= a_{i+1} B_i + B_{i-1}, \end{aligned}$$

for $i \geq 1$.

The RCF for $\sqrt{D}$ is periodic with period-length k:

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

The smallest positive integer solution of $x^2 - Dy^2 = \pm 1$ is given by

$$(x, y) = (A_{k-1}, B_{k-1}),$$

where A_n/B_n is the n-th convergent to $\sqrt{D}$.

Euler observed that if $k = 2h - 1$,

$$\begin{aligned} A_{k-1} &= A_{h-1}B_{h-1} + A_{h-2}B_{h-2} \\ B_{k-1} &= B_{h-1}^2 + B_{h-2}^2, \end{aligned}$$

while if $k = 2h$,

$$\begin{aligned} A_{k-1} &= B_{h-1}(A_h + A_{h-2}) + (-1)^h \\ B_{k-1} &= B_{h-1}(B_h + B_{h-2}). \end{aligned}$$

Also if

$$\xi_n = (P_n + \sqrt{D})/Q_n = [a_n, a_{n+1}, \ldots]$$

is the n-th complete quotient of the RCF expansion of $\sqrt{D}$, then the equations

$$\begin{array}{cc} Q_h = Q_{h-1} & \text{if } k = 2h - 1, \\ P_h = P_{h+1} & \text{if } k = 2h, \end{array}$$

enable us to determine h, as $Q_v = Q_{v+1}$ (k odd) and $P_v = P_{v+1}$, (k even), $1 \leq v < k$ imply $k = h$.

The *nearest square* continued fraction (NSCF) of a quadratic surd ξ_0 was introduced by A.A.K. Ayyangar (AAK) in 1940 and based on the *cyclic* method of solving Pell's equation due to Bhaskara in 1150.

The NSCF is a *half-regular* continued fraction. ie.

$$\xi_0 = a_0 + \cfrac{\epsilon_1}{a_1 + \cfrac{\epsilon_2}{a_2 + \cdots}}$$

where the a_i are integers and

$$a_i \geq 1, \epsilon_i = \pm 1 \text{ and } a_i + \epsilon_{i+1} \geq 1 \text{ if } i \geq 1.$$

We write the continued fraction as

$$\xi_0 = a_0 + \frac{\epsilon_1|}{|a_1} + \frac{\epsilon_2|}{|a_2} + \cdots$$

Let $\xi_0 = \frac{P+\sqrt{D}}{Q}$ be a quadratic surd in *standard form*.
ie. D is a non-square positive integer and $P, Q \neq 0, \frac{D-P^2}{Q}$ are integers, having no common factor other than 1.

Then with $a = \lfloor \xi_0 \rfloor$, the integer part of ξ_0, we can represent ξ_0 in one of two forms (*positive* or *negative* representations)

$$\frac{P+\sqrt{D}}{Q} = a + \frac{Q'}{P'+\sqrt{D}} = a + 1 - \frac{Q''}{P''+\sqrt{D}},$$

where $\frac{P'+\sqrt{D}}{Q'} > 1$ and $\frac{P''+\sqrt{D}}{Q''} > 1$ are also standard surds.

These equations imply

$$\begin{array}{llll} (1) & P' = aQ - P; & (2) & P^{''} = (a+1)Q - P, \\ (3) & P'^2 = R - QQ'; & (4) & {P^{''}}^2 = R + QQ^{''}. \end{array}$$

These in turn imply

$$(5) \quad P^{''} - P' = Q \text{ and } (6) \quad P^{''} + P' = Q' + Q^{''}.$$

We get P' and Q' from (1) and (3), then $P^{''}$ and $Q^{''}$ from (5) and (6), respectively.

AAK chose the partial denominator a_0 and numerator ϵ_1 of the new continued fraction development as follows:

(a) $a_0 = a$ if $|Q'| < |Q^{''}|$, or $|Q'| = |Q^{''}|$ and $Q < 0$,

(b) $a_0 = a + 1$ if $|Q'| > |Q^{''}|$, or $|Q'| = |Q^{''}|$ and $Q > 0$.

Also

$$\epsilon_1 = 1 \qquad \text{and } \xi_1 = \tfrac{P' + \sqrt{D}}{Q'} \qquad \text{in case (a),}$$

$$\epsilon_1 = -1 \qquad \text{and } \xi_1 = \tfrac{P^{''} + \sqrt{D}}{Q^{''}} \qquad \text{in case (b).}$$

Then $\xi_0 = a_0 + \frac{\epsilon_1}{\xi_1}$ and

$\epsilon_1 = \pm 1, a_0$ is an integer and $\xi_1 = \frac{P_1 + \sqrt{D}}{Q_1} > 1$.

We proceed similarly with ξ_1 and so on:

$$\xi_n = a_n + \frac{\epsilon_{n+1}}{\xi_{n+1}}$$

and

$$\xi_0 = a_0 + \frac{\epsilon_1|}{|a_1} + \frac{\epsilon_2|}{|a_2} + \cdots$$

ξ_{n+1} is called the *successor* of ξ_n.

Relations analogous to those for regular continued fractions hold for P_n, Q_n and $a_n, n \geq 0$:

$$\begin{aligned} P_{n+1} + P_n &= a_n Q_n \\ P_{n+1}^2 + \epsilon_{n+1} Q_n Q_{n+1} &= D. \end{aligned}$$

The $|Q_n|$ successively diminish as long as $|Q_n| > \sqrt{D}$ and so eventually, we have $|Q_n| < \sqrt{D}$. When this stage is reached, $0 < P_i < 2\sqrt{D}$ and $0 < Q_i < \sqrt{D}$ for $i \geq n+1$.

This implies eventual periodicity of the complete quotients and hence the partial quotients.

AAK defines ξ_v to be a *special* surd if

$$Q_{v-1}^2 + \tfrac{1}{4}Q_v^2 \le D, \quad Q_v^2 + \tfrac{1}{4}Q_{v-1}^2 \le D.$$

A *semi-reduced* surd is the successor of a special surd.
A *reduced* surd to the successor of a semi-reduced surd.

Properties:

1. A semi-reduced surd is a special surd.
2. A quadratic surd has a purely periodic NSCF expansion if and only if it is reduced.
3. If ξ_v is reduced, then $P_v > 0, Q_v > 0$ and $a_v \ge 2$.

Examples: (i) $\frac{p+q+\sqrt{p^2+q^2}}{p}, p > 2q > 0, \gcd(p,q) = 1,$
(ii) the successor of $\sqrt{D}$.

The NSCF development of $\sqrt{D}$ has the form

$$\sqrt{D} = a_0 + \frac{\epsilon_1|}{|\underset{*}{a_1}} + \cdots \frac{\epsilon_k|}{|\underset{*}{2a_0}}, \tag{1}$$

where the asterisks denote a period of length k and $\xi_p = \xi_{p+k}$, $\epsilon_p = \epsilon_{p+k}$, $a_p = a_{p+k}$ for $p \geq 1$.

Note: a_0 is the nearest integer to $\sqrt{D}$.

There are two types of NSCF expansions of $\sqrt{D}$:

(I) No complete quotient of the 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 $k > 1$:

$$\begin{array}{rcrl} a_v & = & a_{k-v} & (1 \leq v \leq k-1) \\ Q_v & = & Q_{k-v} & (1 \leq v \leq k-1) \\ \epsilon_v & = & \epsilon_{k+1-v} & (1 \leq v \leq k) \\ P_v & = & P_{k+1-v} & (1 \leq v \leq k). \end{array}$$

Note: If $k = 2h+1$, then $Q_h = Q_{h+1}$.
Conversely $Q_v = Q_{v+1}, 1 \leq v < k$ implies $v = h$.

If $k = 2h$, then $P_h = P_{h+1}$.
Conversely $P_v = P_{v+1}, 1 \leq v < k$ implies $v = h$.

Examples.

$\sqrt{73} = 9 - \frac{1|}{|\underset{*}{2}} + \frac{1|}{|5} + \frac{1|}{|5} + \frac{1|}{|2} - \frac{1|}{|\underset{*}{18}}$. (odd period)

$\sqrt{19} = 4 + \frac{1|}{|\underset{*}{3}} - \frac{1|}{|5} - \frac{1|}{3} + \frac{1|}{|\underset{*}{8}}$. (even period)

$\sqrt{n^2+1} = n + \frac{1|}{|\underset{*}{2n}} \quad (n \geq 1), \quad \sqrt{n^2-1} = n - \frac{1|}{|\underset{*}{2n}} \quad (n > 1).$

(II) There is one complete quotient ξ_h in the cycle of the form $\frac{p+q+\sqrt{p^2+q^2}}{p}$, where $p > 2q > 0, \gcd(p,q) = 1$. In this case $k \geq 4$ is even and $h = k/2$. This type also possesses the symmetries of Type I, apart from a central set of three unsymmetrical terms:

$$a_{\frac{k}{2}} = 2, \epsilon_{\frac{k}{2}} = -1, \epsilon_{\frac{k}{2}+1} = 1, a_{\frac{k}{2}-1} = a_{\frac{k}{2}+1} + 1.$$

$$\sqrt{D} = a_0 + \frac{\epsilon_1|}{|\underset{*}{a_1}} + \cdots + \frac{\epsilon_{\frac{k}{2}-1}|}{|a_{\frac{k}{2}-1}} - \frac{1|}{|2} + \frac{1|}{|a_{\frac{k}{2}-1} - 1} + \cdots + \frac{\epsilon_k|}{|2\underset{*}{a_0}}.$$

For example $\sqrt{29} = 5 + \frac{1|}{|\underset{*}{3}} - \frac{1|}{|2} + \frac{1|}{2} + \frac{1|}{|\underset{*}{10}}$.

Other examples are $53, 58, 85, 97$.

E_n is the number of $D < 10^n$ of Type I with even period.

O_n is the number of $D < 10^n$ of Type I with odd period.

F_n is the number of $D < 10^n$ of Type II.

N_n is the number of $D < 10^n$.

n	E_n	O_n	F_n	N_n
2	60	25	5	90
3	762	165	42	969
4	8252	1266	382	9900
5	85856	10465	3363	99684
6	878243	90533	30224	999000

Note: $P_v \neq P_{v+1}, 1 \leq v < k$.

$\epsilon_h = -1$, Q_{h-1} is even and $P_h = Q_h + \frac{1}{2}Q_{h-1}$
(observed by John Robertson and Jim White).

Conversely if $\epsilon_v = -1$, Q_{v-1} is even and
$P_v = Q_v + \frac{1}{2}Q_{v-1}, 1 \leq v < k$, then D is of Type II and $v = h$.

For both types I and II, we have $Q_k = 1$. For

$$\begin{aligned}\sqrt{D} &= a_0 + \frac{\epsilon_1 Q_1}{P_1 + \sqrt{D}} \\ &= a_0 + \frac{\epsilon_1 Q_1 (P_1 - \sqrt{D})}{P_1^2 - D} \\ &= a_0 - P_1 + \sqrt{D}.\end{aligned}$$

Hence $P_1 = a_0$. Then

$$\begin{aligned} P_1 &= P_k \text{ (symmetry)} \\ P_1 &= P_{k+1} \text{ (periodicity)} \\ 2a_0 = 2P_1 &= P_k + P_{k+1} \\ &= a_k Q_k \\ &= 2a_0 Q_k.\end{aligned}$$

Note: $\xi_k = \frac{P_k + \sqrt{D}}{Q_k} = a_0 + \sqrt{D}$.

A classical result for a half-regular expansion of $\xi_0 = \frac{P_0+\sqrt{D}}{Q_0}$ is

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

In the special case $\xi_0 = \sqrt{D}$, where $Q_0 = 1 = Q_k$, we have

$$A_{k-1}^2 - DB_{k-1}^2 = (-1)^k\epsilon_1\epsilon_2\cdots\epsilon_k.$$

Also by periodicity, $Q_n = 1$ if k divides n.

Conversely, suppose $Q_n = 1, n \geq 1$.

Then $\xi_n = P_n + \sqrt{D}$.

We prove $P_n = [\sqrt{D}] = a_0$, the nearest integer to $\sqrt{D}$.

Then $\xi_n = a_0 + \sqrt{D} = \xi_k$ and k divides n.

We start with $P_n^2 + \epsilon_n Q_{n-1} Q_n = D$, noting that $Q_{n-1} > 0, Q_n > 0$.

Case 1. $P_n > \sqrt{D}$. Then $\epsilon_n = -1$.

$$P_n^2 - D = Q_{n-1} < \sqrt{D} \text{ (ξ_n is reduced)}$$
$$0 < P_n - \sqrt{D} < \frac{\sqrt{D}}{P_n + \sqrt{D}} < \frac{\sqrt{D}}{2\sqrt{D}} = \frac{1}{2}.$$

Hence $P_n = [\sqrt{D}]$.

Case 2. $P_n < \sqrt{D}$. Then $\epsilon_n = 1$.

$$\begin{aligned}
Q_{n-1}^2 + \tfrac{1}{4}Q_n^2 &\leq D = P_n^2 + Q_{n-1} \text{ (ξ_n is reduced)} \\
(Q_{n-1} - \tfrac{1}{2})^2 &\leq P_n^2 \\
Q_{n-1} - \tfrac{1}{2} &\leq P_n \\
Q_{n-1} &\leq P_n + \tfrac{1}{2} \\
D - P_n^2 = Q_{n-1} &\leq P_n \\
0 < \sqrt{D} - P_n &\leq \frac{P_n}{\sqrt{D} + P_n} < \frac{P_n}{2P_n} = \frac{1}{2}.
\end{aligned}$$

Again $P_n = [\sqrt{D}]$.

The convergents $A_{kt-1}/B_{kt-1}, t \geq 1$, in fact give *all* positive integer solutions of Pell's equation $x^2 - Dy^2 = \pm 1$.

For if $x^2 - Dy^2 = \pm 1$, $x > 0, y > 0$, we can prove that x/y is an NSCF convergent to $\sqrt{D}$, as follows.

It is certainly an RCF convergent.

We now introduce a transformation $\mathfrak{T}_1$ of Perron, which converts a half-regular continued fraction to an RCF:

To get the RCF partial quotients:

Before a negative partial numerator, insert the term $\frac{+1|}{|1}$.
Replace each $a_n, n \geq 0$ by:

(a) a_n if ϵ_n= $+1$, ϵ_{n+1}= $+1$,

(b) $a_n - 1$ if ϵ_n= $+1$, ϵ_{n+1}= -1, or ϵ_n= -1, ϵ_{n+1}= $+1$,

(c) $a_n - 2$ if ϵ_n= -1, ϵ_{n+1}= -1.

Here $\epsilon_0 = 1$.

Note: If ξ_v and ξ_{v+1} are NSCF reduced quadratic surds and $\epsilon_v = -1$ and $\epsilon_{v+1} = -1$, then $a_v \geq 3$.

Hence $\mathfrak{T}_1$ produces a "genuine" RCF, ie. with no zero partial quotients.

For $n \geq 0$,

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

$$A'_{m-1}/B'_{m-1} = (A_n - A_{n-1})/(B_n - B_{n-1}), \quad A'_m/B'_m = A_n/B_n$$

and RCF complete quotients

$$\frac{P'_m + \sqrt{D}}{Q'_m} = \xi_{n+1}/(\xi_{n+1} - 1), \quad \frac{P'_{m+1} + \sqrt{D}}{Q'_{m+1}} = \xi_{n+1} - 1.$$

(ii) $\epsilon_{n+1} = 1$ gives rise to RCF convergent A_n/B_n and RCF complete quotient ξ_{n+1}.

It is not difficult to show that x/y does not have the form $(A_n - A_{n-1})/(B_n - B_{n-1})$ and hence x/y must also be an NSCF convergent.

Remark. Arguing along these lines shows that the period length of the RCF expansion of $\sqrt{D}$ is $k + r$, where r is the number of $\epsilon_n = -1$ occurring in the period partial numerators $\epsilon_1, \ldots, \epsilon_k$ of the NSCF expansion of $\sqrt{D}$.

Example. $D = 97$. The NSCF expansion of $\sqrt{97}$ is of type II, with period-length 6. There are five $\epsilon_i = -1$ in the period range $1 \leq i \leq 6$ and the period-length of the RCF expansion is 11.

j	i	ξ_i	ξ'_j	ϵ_i	a_i	a'_j	A_i/B_i	A'_j/B'_j
0	0	$\frac{0+\sqrt{97}}{1}$	$\frac{0+\sqrt{97}}{1}$	1	10	9	10/1	9/1
1			$\frac{9+\sqrt{97}}{16}$			1		10/1
2	1	$\frac{10+\sqrt{97}}{3}$	$\frac{7+\sqrt{97}}{3}$	-1	7	5	69/7	59/6
3			$\frac{8+\sqrt{97}}{11}$			1		69/7
4	2	$\frac{11+\sqrt{97}}{8}$	$\frac{3+\sqrt{97}}{8}$	-1	3	1	197/20	128/13
5			$\frac{5+\sqrt{97}}{9}$			1		197/20
6	3	$\frac{13+\sqrt{97}}{9}$	$\frac{4+\sqrt{97}}{9}$	-1	2	1	325/33	325/33
7	4	$\frac{5+\sqrt{97}}{8}$	$\frac{5+\sqrt{97}}{8}$	1	2	1	847/86	522/53
8			$\frac{3+\sqrt{97}}{11}$			1		847/86
9	5	$\frac{11+\sqrt{97}}{3}$	$\frac{8+\sqrt{97}}{3}$	-1	7	5	5604/569	4757/483
10			$\frac{7+\sqrt{97}}{16}$			1		5604/569
11	6	$\frac{10+\sqrt{97}}{1}$	$\frac{9+\sqrt{97}}{1}$	-1	20	18	111233/11294	105629/10725
12			$\frac{9+\sqrt{97}}{16}$			1		111233/11294
13	7	$\frac{10+\sqrt{97}}{3}$	$\frac{7+\sqrt{97}}{3}$	-1	7	5	773027/78489	661794/67195

Exactly one of the following P, Q and PQ tests 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

$$\begin{aligned} A_{k-1} &= A_h B_{h-1} + \epsilon_h A_{h-1} B_{h-2} \\ B_{k-1} &= B_{h-1}(B_h + \epsilon_h B_{h-2}). \end{aligned}$$

In this case $A_{k-1}^2 - DB_{k-1}^2 = 1$.

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

$$\begin{aligned} A_{k-1} &= A_h B_h + \epsilon_{h+1} A_{h-1} B_{h-1} \\ B_{k-1} &= B_h^2 + \epsilon_{h+1} B_{h-1}^2. \end{aligned}$$

In this case $A_{k-1}^2 - DB_{k-1}^2 = -\epsilon_{h+1}$.

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

$$\begin{aligned} A_{k-1} &= A_h B_{h-1} - B_{h-2}(A_{h-1} - A_{h-2}) \\ B_{k-1} &= 2B_{h-1}^2 - B_h B_{h-2}. \end{aligned}$$

In this case $A_{k-1}^2 - DB_{k-1}^2 = -1$.

The formulae for A_{k-1} and B_{k-1} depend on the following *conservation* identities which are proved using "downward" induction on t:

(i) Let $k = 2h + 1$. Then for Type I and $0 \leq t \leq h$, we have

$$A_{2h} = A_{h+t}B_{h-t} + \epsilon_{h+1+t}A_{h+t-1}B_{h-t-1}$$
$$B_{2h} = B_{h+t}B_{h-t} + \epsilon_{h+1+t}B_{h+t-1}B_{h-t-1}$$

(ii) Let $k = 2h$. 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

$$A_{2h-1} = A_{h+t-1}B_{h-t} + \epsilon_{h+t}A_{h+t-2}B_{h-t-1}$$
$$B_{2h-1} = B_{h+t-1}B_{h-t} + \epsilon_{h+t}B_{h+t-2}B_{h-t-1}$$

Let $\pi(D)$ and $p(D)$ respectively denote the periods of the NSCF and RCF expansions of $\sqrt{D}$, where D is not a perfect square and let

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

n	$\Pi(n)$	$p(n)$	$\Pi(n)/P(n)$
1000000	152198657	219245100	.6941941
2000000	417839927	601858071	.6942499
3000000	755029499	1087529823	.6942609
4000000	1149044240	1655081352	.6942524
5000000	1592110649	2293328944	.6942356
6000000	2078609220	2994112273	.6942322
7000000	2604125007	3751067951	.6942356
8000000	3165696279	4559939520	.6942408
9000000	3760639205	5416886128	.6942437
10000000	4387213325	6319390242	.6942463

There are grounds for believing that

$$\Pi(n)/P(n) \to \frac{\log\left(\frac{1+\sqrt{5}}{2}\right)}{\log 2} = .6942419\cdots$$

For D with a long RCF period, we expect $\pi(D)/p(D)$ to be near this value.

For example, $D = 26437680473689$, Daniel Shanks (1974) and quoted by William Adams (1979).

$$p(D) = 18331889,\ \pi(D) = 12726394,\ \pi(D)/p(D) = .6942216\cdots$$

This D obeys the PQ-test.

AAK's paper and a LaTeX version are available at

http://www.numbertheory.org/continued_fractions.html

BCMATH versions of NSCF and some other continued fraction algorithms are available at

http://www.numbertheory.org/php/CFRAC.html