The Generation of Square Roots on a Computer with Rapid Multiplication compared with Division

By Wendy James and P. Jarratt

1. Introduction. For computers with rapid multiplication compared with division, the iteration for $A^{-1/2}$,

$$x_{n+1} = \frac{x_n}{2} (3 - Ax_n^2),$$

may provide a more economical method for generating square roots than the usual technique based on the recurrence

$$x_{n+1} = \frac{1}{2}(x_n + Ax_n^{-1}).$$

The number of iterations of (1.1) necessary to achieve a prescribed accuracy will depend on the closeness of the initial approximation used and as the range of $A$ is decreased, increasingly accurate initial approximations become possible. A way of reducing the infinite range $(0, \infty)$ of $A$ to a convenient finite interval consists in expressing $A$ in the form $a^{2b}$, $\frac{1}{4} \leq a < 1$, and using $A^{1/2} = a^{1/2b^2}$. Still more accurate initial approximations are possible if the reduced range is itself split up.

2. Initial Approximations. As our choice of form for $x_0$ is limited to polynomials, we require, for some given $r$, the coefficients $c_i$ of

$$x_0 = \sum_{i=0}^{r-1} c_i a^i,$$

where $0 < N_1 \leq a \leq N_2$. Following Kogbetliantz (1960) we may set $N_1 = b(1 - p_1)$, $0 < p_1 < 1$, $N_2 = b(1 + p_2)$, $0 < p_2 < 1$, and $a = b(1 + u)$, $-p_1 \leq u \leq p_2$, so that $u$ is within $(-1, 1)$, and approximate to $a^{-1/2}$ by a polynomial $F(u)$ having $r$ undetermined coefficients. We have $a^{-1/2} = b^{-1/2}(1 + u)^{-1/2}$

$$= \sum_{i=0}^{r} w_i u^i - F(u) = u'G(u), G(u) = \sum_{i=0}^{r} g_i u^i,$$

then the relative error $E(u)$ of the approximation is

$$E(u) = u'G(u)/a^{-1/2} = b^{1/2}(1 + u)^{1/2}u'G(u).$$

It is easy to verify that $g_{i+1}/g_i \sim -1$ rapidly as $i$ increases and hence, to good accuracy, $G(u) \sim g_0(1 + u)^{-1}$, where $g_0 = (b^{-1/2}/r!)(-\frac{1}{2})^{-1}(1 - 2\cdots 2r - 1)$. Hence

$$E(u) = Cu'(1 + u)^{-1/2}, C \text{ constant},$$

and we find that $E(u)$ has maximum modulus at $u = -p_1$, $u = p_2$. Writing $K$ as the error bound of the approximation, $K$ is minimised when $|E(-p_1)| = |E(p_2)| = K$, and from the equations

$$(1 + p_2)/(1 - p_1) = N_2/N_1, \ p_1'(1 - p_1)^{1/2} = p_2'(1 + p_2)^{1/2},\ K = C p_2'(1 + p_2)^{1/2},$$

we can determine $p_1$, $p_2$, and $K$ and hence $b$ and the coefficients of $F(u)$. However, a simple modification to the procedure will give improved approximations. Taking

Received June 18, 1964. Revised November 10, 1964.
first the case where \( F(u) \) is a polynomial of odd degree, \( K \) is a minimum when
eq \( E_i = E_{i+2} = K \) and also \( E(u) \) has constant sign in \((-1, 1)\) since \( E(0) = 0 \).
If then we write \( E_1(u) = E(u) - K/2 \), we have
\[
E_1(u) = \left( a^{-1/2} - F(u) - Ka^{-1/2}/2 \right) / a^{-1/2}
\]
and by taking \( F_1(u) = F(u) + Ka^{-1/2}/2 \sim F(u)(1 + K/2) \), \( K^2 \) being assumed
negligible, the maximum error will be reduced to \( K/2 \). If, however, \( F(u) \) is of
even degree \( r - 1 \), the possibility of improvement depends on \( r \) and the range
\((N_1, N_2)\). Now \( K \) is minimised when \( -E_1(p_1) = E_{i+2} = K \) and if we take
\( E_1(p_1) = E(u) = Ku/2p_1 \), then \( E_i(-p_1) = -K/2 \) and \( |E_i(p_1)| < K/2 \), for
\(-p_1 < u < p_2 \). Also \( E_1(u) = (a^{-1/2} - F(u) - Ka^{-1/2}/2p_1) / a^{-1/2} \) and we have
\[
F_1(u) = F(u) + Ku/2p_1 \sim F(u) \left( 1 + Ku/2p_1 \right) .
\]
However, \( F_1(u) \) is a polynomial of degree \( r \) and hence neglecting the term in \( u \)
gives \( F_2(u) = (1 + Ku/2p_1) F(u) - (K/2p_1) w_{r-1} u \). Hence
\[
|E_2(u)| \leq K/2 + Ku/2p_1 \max \left| \frac{u'}{F_2(u)} \right| = K/2 \left( 1 + \frac{w_{r-1} p_r}{p_1 F_2(p_1)} \right).
\]
Thus \( F_2(u) \) will generate more accurate approximations than \( F(u) \) provided that
\( w_{r-1} p_r/p F_2(p_2) \) is sufficiently small.

A third method for obtaining best possible approximations (Eve, 1963) cons-
ists in applying a theorem of Chebyshev (Achieser, 1956). Writing \( \epsilon_0 =
(x_0 - a^{-1/2}) / a^{-1/2} \), we require the coefficients \( c_i \) in (2.1) so that in \((N_1, N_2)\)
we minimise \( \max |\epsilon_0| \). If the \( c_i \) are chosen such that \( \epsilon_0 \) vanishes at \( r \) points in \((N_1, N_2)\),
it follows that \( \epsilon_0 \) has \( r - 1 \) stationary values \( S_i \), where \( N_1 < S_1 < S_2 \cdots
< S_{r-1} < N_2 \), and by setting
\[
(2.2) \quad \epsilon_0(N_1) = -\epsilon_0(S_1) = \epsilon_0(S_2) = \cdots = (-1)^{r-1} \epsilon_0(S_{r-1}) = (-1)^r \epsilon_0(N_2),
\]
we obtain \( r \) equations from which the \( c_i \) may be calculated. The relative errors of
approximations found by this approach turn out in practice to be about a quarter
of those of Kogbetliantz's method and about half those resulting from the modified
procedure. However, the application of (2.2) leads, for \( r > 1 \), to a system of non-
linear equations and, for \( r \geq 3 \), the computational labour involved in their solution
is heavy. Consequently, since the modified Kogbetliantz method is simple to apply,
especially when \( r \) is even, a decision on which technique to adopt will depend on the
accuracy requirements of the approximation under consideration.

3. Sub-Division of \((N_1, N_2)\). If the error bound \( K \) of a particular initial ap-
proximation is too large, it may be reduced either by increasing the degree of the
approximating polynomial or by a division of the range. However, from the form of
(1.1) and the fact that its relative errors satisfy \( \epsilon_{n+1} = -\epsilon_n^2 (\epsilon_n + 3) / 2 \), it is clearly
uneconomic to use polynomials of greater degree than a cubic. Hence, in a number
of cases, we must sub-divide \((N_1, N_2)\) and it can readily be shown that if the range is
split up into \( M \) intervals, \( N_1 < d_1 \cdots < d_{M-1} < N_2 \), then for the maximum error
over the whole range to be minimised, the \( d_i \) must satisfy \( d_1 = D N_1 \),
\( d_2 = D^2 N_1 \), \cdots, \( d_{M-1} = D^{M-1} N_1 \), \( N_2 = D^M N_1 \), this result being independent of the
method of approximation used. Hence for the interval \((.25, 1)\), the sub-divisions
## Table 1

### Most Economical Approximations to $\sqrt[2]{a}$

<table>
<thead>
<tr>
<th>Accuracy in Correct Decimal Places</th>
<th>Sections $M$</th>
<th>Form of Initial Approximation</th>
<th>Iterations of (4.1) Necessary</th>
<th>Operations</th>
<th>Precomputed Constants</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>4</td>
<td>Quadratic</td>
<td>1</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>Linear with Shift</td>
<td>2</td>
<td>6 + $S$</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>Cubic</td>
<td>1</td>
<td>6</td>
<td>19</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>Linear</td>
<td>2</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>Linear</td>
<td>2</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>Quadratic with Shift (a)</td>
<td>2</td>
<td>7 + $S$</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>Quadratic with Shift (b)</td>
<td>2</td>
<td>7 + $S$</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>Quadratic</td>
<td>2</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>Cubic with Shift</td>
<td>2</td>
<td>8 + $S$</td>
<td>7</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>Quadratic</td>
<td>2</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td>16</td>
<td>2</td>
<td>Linear with Shift (a)</td>
<td>3</td>
<td>9 + $S$</td>
<td>2</td>
</tr>
<tr>
<td>18</td>
<td>2</td>
<td>Quadratic with Shift (b)</td>
<td>3</td>
<td>10 + $S$</td>
<td>5</td>
</tr>
<tr>
<td>20</td>
<td>3</td>
<td>Quadratic</td>
<td>3</td>
<td>11</td>
<td>7</td>
</tr>
<tr>
<td>20</td>
<td>2</td>
<td>Cubic with Shift</td>
<td>3</td>
<td>11 + $S$</td>
<td>7</td>
</tr>
</tbody>
</table>

**Table 2**

### Coefficients for Approximations Given in Table 1

<table>
<thead>
<tr>
<th>Form of Initial Approximation</th>
<th>Sections $M$</th>
<th>$\alpha = \frac{1}{(25)^{1/2} M}$</th>
<th>Accuracy</th>
<th>Range</th>
<th>$t_0$</th>
<th>$t_1$</th>
<th>$t_2$</th>
<th>$t_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear with Shift</td>
<td>2</td>
<td>.5</td>
<td>$2.526 \times 10^{-1}$</td>
<td>$\alpha, 1$</td>
<td>.89466</td>
<td>-40625</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear</td>
<td>3</td>
<td>.62996</td>
<td>$9.962 \times 10^{-2}$</td>
<td>$\alpha, 1$</td>
<td>.84275</td>
<td>-34773</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear with Shift</td>
<td>3</td>
<td>.62996</td>
<td>$1.189 \times 10^{-1}$</td>
<td>$\alpha, 1$</td>
<td>.86588</td>
<td>-375</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear</td>
<td>4</td>
<td>.70711</td>
<td>$5.014 \times 10^{-2}$</td>
<td>$\alpha, 1$</td>
<td>.81838</td>
<td>-32119</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quadratic</td>
<td>2</td>
<td>.5</td>
<td>$3.259 \times 10^{-1}$</td>
<td>$\alpha, 1$</td>
<td>1.11667</td>
<td>-1.03240</td>
<td>.417363</td>
<td></td>
</tr>
<tr>
<td>Quadratic with Shift (a)</td>
<td>2</td>
<td>.5</td>
<td>$3.954 \times 10^{-1}$</td>
<td>$\alpha, 1$</td>
<td>1.11076</td>
<td>-1.01573</td>
<td>.40625</td>
<td></td>
</tr>
<tr>
<td>Quadratic with Shift (b)</td>
<td>2</td>
<td>.5</td>
<td>$3.465 \times 10^{-2}$</td>
<td>$\alpha, 1$</td>
<td>1.11491</td>
<td>-1.02745</td>
<td>.4140625</td>
<td></td>
</tr>
<tr>
<td>Quadratic</td>
<td>3</td>
<td>.62996</td>
<td>$9.755 \times 10^{-1}$</td>
<td>$\alpha, 1$</td>
<td>1.05307</td>
<td>-87885</td>
<td>.324314</td>
<td></td>
</tr>
<tr>
<td>Quadratic</td>
<td>4</td>
<td>.70711</td>
<td>$4.099 \times 10^{-1}$</td>
<td>$\alpha, 1$</td>
<td>1.02274</td>
<td>-806890</td>
<td>.284351</td>
<td></td>
</tr>
<tr>
<td>Cubic with Shift</td>
<td>2</td>
<td>.5</td>
<td>$1.192 \times 10^{-1}$</td>
<td>$\alpha, 1$</td>
<td>1.269357</td>
<td>-1.70236</td>
<td>.360910</td>
<td>.4375</td>
</tr>
<tr>
<td>Cubic</td>
<td>4</td>
<td>.70711</td>
<td>$1.209 \times 10^{-2}$</td>
<td>$\alpha, 1$</td>
<td>1.186218</td>
<td>-1.394930</td>
<td>.9842185</td>
<td>-2755689</td>
</tr>
</tbody>
</table>
will be $\alpha^M < \alpha^{M-1} \cdots < \alpha < 1$ with $\alpha = (.25)^{1/M}$. If now we approximate in $(\alpha, 1)$ by (2.1) it is easy to prove that the corresponding approximation in $(\alpha'^{+1}, \alpha')$ is given by

$$
(3.1) \quad x_0^{(p)} = \sum_{i=0}^{r-1} C_i (1/\alpha)^{i(\alpha^{+1}/\alpha)} \alpha^i.
$$

4. Results. So that numbers remained fractional during computation, it was necessary to rewrite (1.1) in the form

$$
(4.1) \quad y_{n+1} = 2y_n (\frac{a}{4} - ay_n^2) \quad \text{with } \lim_{n \to \infty} y_n = \frac{1}{2}a^{-1/2}.
$$

Linear and quadratic approximations were then found using the Chebyshev method but for cubic approximations it was much simpler to use the modified Kogbetliantz approach. However, it was evident from our results that use of the best possible cubic would not have decreased the number of iterations of (4.1) required. In assessing the best approximation for a particular case, we have given most weight to minimising the total number of multiplications, at the same time restricting the number of intervals to a reasonable total. As in many cases final accuracies were obtained which were better than the target bounds set, it was possible to simplify $y_0$ at the expense of this excess accuracy by replacing the first multiplication in the evaluation of $y_0$ by a series of shifts. These results have been given wherever they seemed of value. In Table 1, the first entry for each specified final accuracy shows the form of approximation which offers the most rapid evaluation of square root, assuming shifting is faster than multiplication. If this is not so, these approximations may not be the best and hence where an approximation which uses shifts yields the fewest number of multiplications, both the modified and unmodified forms have been given. The “Operations” column in Table 1 gives the total number of multiplications involved in forming the approximation and carrying out the iterations together with $S$ where shifts are also needed in obtaining the initial approximation. The coefficients of the approximations described in Table 1 are given in Table 2. For unmodified approximations, only results for $(\alpha, 1)$ have been quoted since the coefficients for the other sections of the range may be readily computed from (3.1)

5. Conclusions. It is clear from an analysis of the forms of (1.1) and (1.2) that the method discussed will be superior to methods based on (1.2) if multiplication is at least four times as fast as division. This will certainly be true if multiplication but not division is done in parallel. If it is satisfied when multiplication is done serially, then a further time saving may be effected by replacing the first multiplication in the evaluation of the initial approximation by a small number of shifts.