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 $A^{-1/2}$,

$$x_{n+1} = \frac{x_n}{2}(3 - Ax_n^2), \qquad (1.1)$$
$$A^{1/2} = A \lim_{n\to\infty} x_n ,$$

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

$$x_{n+1} = \tfrac{1}{2}(x_n + Ax_n^{-1}). \qquad (1.2)$$

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 $a2^{2b}$, $\frac{1}{4} \leqq a < 1$, and using $A^{1/2} = a^{1/2}2^b$. 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

$$x_0 = \sum_{i=0}^{r-1} c_i a^i, \qquad (2.1)$$

where $0 < N_1 \leqq a \leqq N_2$. 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 $a = b(1 + u)$, $-p_1 \leqq u \leqq p_2$, so that u is within $(-1, 1)$, and approximate to $a^{-1/2}$ by a polynomial $F(u)$ having r undetermined coefficients. We have $a^{-1/2} = b^{-1/2}(1 + u)^{-1/2} = \sum_{i=0}^{\infty} w_i u^i$, where $w_i = (b^{-1/2}/i!)(-\frac{1}{2})^{i-1}(1\cdot 2 \cdots 2i - 1)$ and if now we let $\sum_{i=0}^{\infty} w_i u^i - F(u) = u^r G(u)$, $G(u) = \sum_{i=0}^{\infty} g_i u^i$, then the relative error $E(u)$ of the approximation is $E(u) = u^r G(u)/a^{-1/2} = b^{1/2}(1 + u)^{1/2}u^r G(u)$. It is easy to verify that $g_{i+1}/g_i \to -1$ rapidly as i increases and hence, to good accuracy, $G(u) \sim g_0(1 + u)^{-1}$, where $g_0 = (b^{-1/2}/r!)(-\frac{1}{2})^{r-1}(1\cdot 2 \cdots 2r - 1)$. Hence $E(u) = Cu^r(1 + u)^{-1/2}$, C 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

$$(1 + p_2)/(1 - p_1) = N_2/N_1,\ p_1^r/(1 - p_1)^{1/2} = p_2^r/(1 + p_2)^{1/2},\ K = Cp_2^r/(1 + p_2)^{1/2},$$

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 $E(-p_1) = E(p_2) = K$ 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

$$E_1(u) = (a^{-1/2} - F(u) - Ka^{-1/2}/2)/a^{-1/2}$$

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 $-E(-p_1) = E(p_2) = K$ and if we take $E_1(u) = E(u) - Ku/2p_1$, then $E_1(-p_1) = -K/2$ and $| E_1(u)| < K/2$, for $-p_1 < u \leqq p_2$. Also $E_1(u) = (a^{-1/2} - F(u) - Ka^{-1/2}/2p_1)/a^{-1/2}$ and we have

$$F_1(u) = F(u) + \frac{Ku}{2p_1} a^{-1/2} \sim F(u)\left(1 + \frac{Ku}{2p_1}\right).$$

However, $F_1(u)$ is a polynomial of degree r and hence neglecting the term in u gives $F_2(u) = (1 + Ku/2p_1)F(u) - (K/2p_1)w_{r-1}u^r$. Hence

$$| E_2(u) | \leqq \frac{K}{2} + \frac{Kw_{r-1}}{2p_1} \operatorname*{Max}_{-p_1 \leqq u \leqq p_2} \left| \frac{u^r}{F_2(u)} \right| = \frac{K}{2}\left(1 + \frac{w_{r-1}}{p_1} \frac{p_2^r}{F_2(p_2)}\right).$$

Thus $F_2(u)$ will generate more accurate approximations than $F(u)$ provided that $w_{r-1}p_2^r/p_1F_2(p_2)$ is sufficiently small.

A third method for obtaining best possible approximations (Eve, 1963) consists 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 ϵ_0 vanishes at r points in (N_1, N_2), it follows that ϵ_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 nonlinear equations and, for $r \geqq 3$, 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 approximation 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 $d_1 = DN_1$, $d_2 = D^2N_1, \cdots, d_{M-1} = D^{M-1}N_1$, $N_2 = D^MN_1$, 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 $1/2\sqrt{a}$

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

TABLE 2

Coefficients for Approximations Given in Table 1

Form of Initial Approximation	Sections M	$\alpha = (.25)^{1/M}$	Accuracy	Range	t_0	t_1	t_2	t_3
Linear with Shift	2	.5	$.2526 \times 10^{-1}$	$\alpha, 1$	.89486	−.40625		
				α^2, α	1.2560	−1.125		
Linear	3	.62996	$.9962 \times 10^{-2}$	$\alpha, 1$	.84275	−.34773		
Linear with Shift	3	.62996	$.1889 \times 10^{-1}$	$\alpha, 1$	.86555	−.375		
				α^2, α	1.0576	−.6875		
				α^3, α^2	1.3325	−1.375		
Linear	4	.70711	$.5614 \times 10^{-2}$	$\alpha, 1$	.81838	−.32119		
Quadratic	2	.5	$.3259 \times 10^{-2}$	$\alpha, 1$	1.11667	−1.03240	.417363	
Quadratic with Shift (a)	2	.5	$.3954 \times 10^{-2}$	$\alpha, 1$	1.11076	−1.01573	.40625	
				α^2, α	1.58107	−2.93059	2.375	
Quadratic with Shift (b)	2	.5	$.3465 \times 10^{-2}$	$\alpha, 1$	1.11491	−1.02745	.4140625	
				α^2, α	1.58107	−2.93059	2.375	
Quadratic	3	.62996	$.9755 \times 10^{-3}$	$\alpha, 1$	1.05307	−.876895	.324314	
Quadratic	4	.70711	$.4099 \times 10^{-3}$	$\alpha, 1$	1.02274	−.806890	.284351	
Cubic with Shift	2	.5	$.1921 \times 10^{-2}$	$\alpha, 1$	1.269357	−1.702386	1.369610	−.4375
				α^2, α	1.797562	−4.835662	7.804216	−5.0
Cubic	4	.70711	$.1209 \times 10^{-3}$	$\alpha, 1$	1.186218	−1.394930	.9842185	−.2755689

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 (α^{j+1}, α^j) is given by

$$x_0^{(j)} = \sum_{i=0}^{r-1} C_i(1/\alpha)^{j((2i+1)/2)}\alpha^i. \tag{3.1}$$

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

$$y_{n+1} = 2y_n(\tfrac{3}{4} - ay_n^2) \qquad \text{with } \lim_{n\to\infty} y_n = \tfrac{1}{2}a^{-1/2}. \tag{4.1}$$

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.

Department of Mathematics
Bradford Institute of Technology
Bradford 7, England

1. N. I. ACHIESER, *Theory of Approximation* (English translation), Ungar, New York, 1956. MR 20 #1872.
2. J. EVE, "Starting approximation for the iterative calculation of square roots," *Comput. J.*, v. 6, 1963, p. 274.
3. E. G. KOGBETLIANTZ, "Generation of elementary functions," *Mathematical Methods for Digital Computers*, Wiley, New York, 1960, pp. 7–35. MR 22 #8681.
4. A. RALSTON & H. S. WILF, (Eds.), *Mathematical Methods for Digital Computers*, Wiley, New York, 1960. MR 22 #8680.