Optimal Starting Approximations for Newton's Method

By P. H. Sterbenz and C. T. Fike

Abstract. Various writers have dealt with the subject of optimal starting approximations for square-root calculation by Newton's method. Three optimality criteria that have been used can be shown to lead to closely related approximations. This fact makes it surprisingly easy to choose a starting approximation of some prescribed form so that the maximum relative error after any number of Newton iterations is as small as possible. ■

1. Introduction. The choice of polynomial and rational starting approximations for square-root calculation by Newton's method has been the subject of various investigations (e.g., [1]–[5]). These approach the problem from several different points of view. In this paper we will show how approximations obtained from these different viewpoints are related and how some of them can be derived from others.

The problem of evaluating $\sqrt{x}$ for any $x > 0$ is easily reduced to the problem of evaluating $\sqrt{x}$ for x in some closed interval $[a, b]$ such that $0 < a < b$. Here $[a, b]$ depends on the radix of the floating-point number system of the computer to be used; typical possibilities are $[1/16, 1]$ and $[\frac{1}{2}, 2]$. The following procedure is used to compute an approximate value for $\sqrt{x}$. Using a polynomial or rational approximation $f(x)$ to $\sqrt{x}$, valid in $[a, b]$, compute a starting value $y_0 = f(x)$ and then obtain $y_1, y_2, \cdots, y_n$ by means of the relation

$$y_{k+1} = \tfrac{1}{2}(y_k + x/y_k), \qquad k = 0, 1, \cdots, n - 1.$$

Then $y_n \approx \sqrt{x}$. It is customary not to test for convergence, since the number of iterations required in practice is quite small. Instead, the number of iterations n is chosen so that y_n is a sufficiently accurate approximation to $\sqrt{x}$ for all x in $[a, b]$.

Let the relative error in the kth Newton iterate be denoted by $R_k(x)$; that is, let

$$R_k(x) = \frac{y_k - \sqrt{x}}{\sqrt{x}}.$$

Then $R_0(x)$ is the relative error of the starting approximation; that is,

$$R_0(x) = \frac{f(x) - \sqrt{x}}{\sqrt{x}}.$$

It is well known that

$$R_{k+1}(x) = \frac{R_k^2(x)}{2[1 + R_k(x)]}. \tag{1}$$

The better the approximation $f(x)$ is, the fewer iterations will be required. Thus we wish to select the coefficients of the approximation $f(x)$ in order to obtain a best fit

Received October 2, 1967, revised September 18, 1968.

in some sense. We will look for a best-fit approximation in a set V of admissible functions. Here either V is the set of all polynomials of degree $\leqq M$ or else V is the set of all rational functions $p(x)/q(x)$ where $p(x)$ and $q(x)$ are relatively prime polynomials of degree $\leqq M$ and N respectively and $q(x)$ does not vanish for $a \leqq x \leqq b$. Let W be the set of all $f(x)$ in V for which $f(x) > 0$ for $a \leqq x \leqq b$. Now $f(x) = \sqrt{a}$ is in W, and we find that for this function $|R_0(x)| < 1$ for $a \leqq x \leqq b$. For any $f(x)$ in V but not in W we would have $|R_0(x)| \geqq 1$. Therefore, in searching for a best-fit approximation to $\sqrt{x}$ we restrict our attention to W.

2. Optimality Criteria. In trying to select the best approximation from W, three criteria have been used. First, we can try to minimize the maximum of $|R_0(x)|$ for $a \leqq x \leqq b$. This leads to the best-fit approximation to $\sqrt{x}$ in the relative-error sense for the interval $[a, b]$, a type of approximation treated extensively in the literature (e.g., [2] and [4]). Another criterion, for which it may be easier to find the optimal approximation by analytical methods, is to minimize the maximum of $|\delta(x)|$, where

$$\delta(x) = \log \frac{f(x)}{\sqrt{x}} . \tag{2}$$

Here

$$\delta(x) = \log [1 + R_0(x)] . \tag{3}$$

Several approximations to $\sqrt{x}$ that are optimal in this sense also appear in the literature (e.g., [1] and [3]). A third approach is to note that we are going to use a fixed number of Newton iterations and always going to take y_n as an approximation to $\sqrt{x}$. Therefore, we can try to minimize the maximum of $|R_n(x)|$ for the last iteration. It is easy to show (see [5]) that the $f(x)$ that minimizes the maximum of $|R_1(x)|$ also minimizes the maximum of $|R_n(x)|$ for every $n \geqq 1$. The three alternatives are thus to select the $f(x)$ in W that minimizes the absolute value of either $R_0(x)$, $R_1(x)$, or $\delta(x)$.

Minimizing $|R_1(x)|$ (and consequently also $|R_n(x)|$) would appear to be what we would like to do in practice, but this would also appear to be the most difficult of the three alternatives to deal with analytically. Thus Moursund [5] in his results concerning the existence of an $f(x)$ optimal by this criterion resorted to a generalization of the classical Chebyshev approximation theory. Fortunately, as we will show below, there are surprisingly simple relationships among the three optimization criteria which make it possible to avoid such difficulties.

3. Relationships Among Optimization Criteria. It follows from (3) that $R_0(x) = e^{\delta(x)} - 1$. Therefore, we have

$$R_1(x) = [e^{\delta(x)} - 1]^2/2e^{\delta(x)} ,$$

which simplifies to

$$R_1(x) = \cosh \delta(x) - 1 . \tag{4}$$

Since $\cosh x$ is an even function and is monotone for all $x > 0$, this yields

$$\max_{[a,b]} |R_1(x)| = \cosh \left(\max_{[a,b]} |\delta(x)| \right) - 1 .$$

Therefore, if we minimize the maximum of $|\delta(x)|$, we have also minimized the maximum of $|R_1(x)|$. Thus, we have

THEOREM 1.* *A function $f(x)$ in W minimizes the maximum of $|\delta(x)|$ if and only if it minimizes the maximum of $|R_1(x)|$.*

We now turn to the relationship between $\delta(x)$ and $R_0(x)$. We note that for any α and any $f(x)$ in W,

$$\frac{\alpha f(x) - \sqrt{x}}{\sqrt{x}} = \alpha e^{\delta(x)} - 1 , \tag{5}$$

and, if $\alpha > 0$,

$$\log \frac{\alpha f(x)}{\sqrt{x}} = \log [\alpha(1 + R_0(x))] . \tag{6}$$

LEMMA 1. *For $f(x)$ in W, let the minimum and maximum of $\delta(x)$ for $a \leqq x \leqq b$ be denoted by λ_1 and λ_2 respectively. Let λ be the larger of $|\lambda_1|$, $|\lambda_2|$; and let*

$$\mu = \max_{[a,b]} \left| \frac{f(x)/\cosh \lambda - \sqrt{x}}{\sqrt{x}} \right| .$$

Then $\mu = \tanh \lambda$.

Proof. From (5),

$$\frac{f(x)/\cosh \lambda - \sqrt{x}}{\sqrt{x}} = \frac{1}{\cosh \lambda} e^{\delta(x)} - 1 ,$$

so the minimum and maximum of this expression are μ_1 and μ_2 respectively, where

$$\mu_i = \frac{1}{\cosh \lambda} e^{\lambda_i} - 1 .$$

Now either $\lambda = \lambda_2$ or $\lambda = -\lambda_1$, so

$$\mu = \max_{[a,b]} \left| \frac{e^{\pm\lambda}}{\cosh \lambda} - 1 \right| .$$

Since

$$e^{\pm\lambda}/\cosh \lambda - 1 = \pm \tanh \lambda ,$$

the lemma follows.

LEMMA 2. *For $f(x)$ in W let the minimum and maximum of $R_0(x)$ for $a \leqq x \leqq b$ be denoted by μ_1 and μ_2 respectively. Let μ be the larger of $|\mu_1|$, $|\mu_2|$; and suppose that $\mu < 1$. Let*

$$g(x) = \log \frac{f(x)}{(1 - \mu^2)^{1/2}\sqrt{x}} ,$$

and let the minimum and maximum of $g(x)$ for $a \leqq x \leqq b$ be denoted by λ_1 and λ_2 respectively. Let λ be the larger of $|\lambda_1|$, $|\lambda_2|$. Then $\lambda = \operatorname{arctanh} \mu$. Morecver, if $\mu_1 = -\mu_2$, then $\lambda_1 = -\lambda_2$.

Proof. From (6),

* We are indebted to the referee for pointing out that this result was also discovered by King and Phillips and will appear in [6].

$$g(x) = \log \frac{1 + R_0(x)}{(1 - \mu^2)^{1/2}},$$

and therefore

$$\lambda_i = \log \frac{1 + \mu_i}{(1 - \mu^2)^{1/2}}.$$

Then

$$\lambda_1 \geqq \log \frac{1 - \mu}{(1 - \mu^2)^{1/2}}, \tag{7}$$

$$\lambda_2 \leqq \log \frac{1 + \mu}{(1 - \mu^2)^{1/2}}. \tag{8}$$

Now either $\mu = -\mu_1$ or $\mu = \mu_2$, so equality holds in at least one of (7) or (8). Since

$$\log \frac{1 \pm \mu}{(1 - \mu^2)^{1/2}} = \pm \log \left(\frac{1 + \mu}{1 - \mu}\right)^{1/2} = \pm \operatorname{arctanh} \mu,$$

we have $\lambda = \operatorname{arctanh} \mu$. If $\mu_2 = -\mu_1$, then equality holds in both (7) and (8), and $\lambda_1 = -\lambda_2$.

Now let $f^*(x)$ be the best-fit approximation to $\sqrt{x}$ in W in the sense of relative error; that is, let $f^*(x)$ be that function in W which minimizes $\max_{[a,b]} |R_0(x)|$. From the theory of best-fit approximations (see, e.g., [7]), we know that $f^*(x)$ exists, is unique, and is characterized by the fact that it yields an equal-ripple $R_0(x)$ with a sufficient number of extreme points. Let

$$\mu_* = \max_{[a,b]} \left| \frac{f^*(x) - \sqrt{x}}{\sqrt{x}} \right|.$$

Since, as we have observed, there is an $f(x)$ in W for which $|R_0(x)| < 1$ for $a \leqq x \leqq b$, it follows that $\mu_* < 1$. Now let

$$\bar{f}(x) = \frac{f^*(x)}{(1 - \mu_*^2)^{1/2}}$$

and

$$\bar{\lambda} = \max_{[a,b]} \left| \log \frac{\bar{f}(x)}{\sqrt{x}} \right|.$$

By Lemma 2,

$$\bar{\lambda} = \operatorname{arctanh} \mu_*. \tag{9}$$

Now, by Lemma 1,

$$\max_{[a,b]} \left| \frac{\bar{f}(x)/\cosh \bar{\lambda} - \sqrt{x}}{\sqrt{x}} \right| = \tanh \bar{\lambda} = \mu_*.$$

If $f(x)$ is any function in W, and if $\lambda = \max_{[a,b]} |\delta(x)|$, then by Lemma 1

$$\max_{[a,b]} \left| \frac{f(x)/\cosh \lambda - \sqrt{x}}{\sqrt{x}} \right| = \tanh \lambda.$$

But $f(x)/\cosh \lambda$ is in W, and therefore

$$\tanh \lambda \geqq \mu_* . \tag{10}$$

Then

$$\lambda \geqq \bar{\lambda} . \tag{11}$$

Therefore $\bar{f}(x)$ minimizes the maximum of $|\delta(x)|$. To prove the uniqueness of $\bar{f}(x)$, we note that if $f(x)$ is any function in W with $\lambda = \bar{\lambda}$, then by the uniqueness of $f^*(x)$ we have $f(x)/\cosh \lambda = f^*(x)$ and

$$f(x) = f^*(x) \cosh \operatorname{arctanh} \mu_* = \frac{f^*(x)}{(1 - \mu_*^2)^{1/2}} .$$

Thus $\bar{f}(x)$ is the unique polynomial in W which minimizes the maximum of $|\delta(x)|$. Moreover, since $f^*(x)$ yields an equal-ripple $R_0(x)$ and since x yields an extreme value for $f^*(x)$ if and only if it yields an extreme value for $\bar{f}(x)$, it follows from Lemma 2 that $\bar{f}(x)$ yields an equal-ripple $\delta(x)$ with the same number of extreme points that $R_0(x)$ has. The fundamental relationship between $\bar{f}(x)$ and $f^*(x)$ is given by

$$\bar{f}(x) = f^*(x)/(1 - \mu_*^2)^{1/2} , \tag{12}$$

$$f^*(x) = \bar{f}(x)/\cosh \bar{\lambda} . \tag{13}$$

Thus we have proved

THEOREM 2. *There is a unique function $\bar{f}(x)$ in W which minimizes the maximum of $|\delta(x)|$ for $a \leqq x \leqq b$. It is characterized by the fact that it yields an equal-ripple $\delta(x)$, and it is related to the function $f^*(x)$ in W which minimizes the maximum of $|R_0(x)|$ by* (12) *and* (13).

4. Conclusion. The significance of the above results can be illustrated in the following way. Consider the problem of approximating $\sqrt{x}$ in the interval [1/16, 1] by means of a rational function expressible in the form $A + B/(x + C)$. The coefficients of the rational function $f_1(x)$ that makes the maximum of $|\delta(x)|$ a minimum can easily be obtained with the aid of formulas given by Maehly (posthumously in the appendix of [3]). (It may be noted that Maehly showed by analytical methods how to get *exact* values for the coefficients, not just approximate numerical values.) Numerical values for the coefficients of the rational function $f_2(x)$ that makes the maximum of $|R_0(x)|$ a minimum were given by Fike [4]; these results were obtained with Remez' method. Numerical values for the coefficients of the rational function $f_3(x)$ that makes the maximum of $|R_1(x)|$ a minimum were also given by Fike [8]; these results were obtained by an *ad hoc* method similar to Remez' method. It is now clear from the results stated in this paper that $f_1(x)$ and $f_3(x)$ are the same and that $f_2(x)$ is merely a constant multiple of them.

IBM Systems Research Institute
New York, New York 10017

1. H. J. MAEHLY, *Approximations for the* CDC 1604, Control Data Corp., 1960.
2. J. EVE, "Starting approximations for the iterative calculation of square roots," *Comput. J.*, v. 6, 1963, pp. 274–276.
3. W. J. CODY, "Double-precision square root for the CDC-3600," *Comm. ACM*, v. 7, 1964, pp. 715–718.
4. C. T. FIKE, "Starting approximations for square-root calculation on IBM System/360," *Comm. ACM*, v. 9, 1966, pp. 297–299.
5. D. G. MOURSUND, "Optimal starting values for Newton-Raphson calculation of $\sqrt{x}$," *Comm. ACM*, v. 10, 1967, pp. 430–432.
6. R. F. KING & D. L. PHILLIPS, "The logarithmic error and Newton's method for the square root," *Comm. ACM*, v. 12, 1969, pp. 87–88.
7. N. I. ACHIESER, *Theory of Approximation*, OGIZ, Moscow, 1947; English transl., Ungar, New York, 1956. MR **10**, 33; MR **20** #1872.
8. C. T. FIKE, "Letter to the editor," *Comm. ACM*, v. 10, 1967, pp. 683–684.