JOURNAL OF APPROXIMATION THEORY 3, 156–163 (1970)

Optimal Starting Approximations for Newton's Method

G. D. TAYLOR*

Michigan State University, East Lansing, Michigan 48823

Communicated by John R. Rice

Received May 25, 1969

1. INTRODUCTION

Recently, R. F. King and D. L. Phillips [4] and P. H. Sterbenz and C. T. Fike [7] have found (independently) that the best starting approximation for the Newton–Raphson calculation of $\sqrt{x}$, [5], and the best logarithmic approximation to $\sqrt{x}$ are the same.

In establishing this result, they have shown that the starting approximation suggested by W. J. Cody [1] for calculating double precision square roots on the CDC-3600 is the best possible choice. Also, the problem of calculating these best starting approximations is now reduced to a standard Remes algorithm, at worst. In our discussion of this problem we shall follow the write-up of the second paper.

In this particular paper [7], Sterbenz and Fike discussed three optimality criteria that have been used for starting approximations for the calculation of square-roots by Newton's method. Using a polynomial or rational approximation $y_0(x) = y(x)$ to $\sqrt{x}$, valid in $[a, b]$ $(0 < a < b)$, we let $y_1, \ldots, y_n$ be defined by

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

y_n is the final approximation to $\sqrt{x}$. The three approaches studied in [7] are:

(i) Find the unique $\tilde{y}$ which minimizes

$$\max_{x\in[a,b]} \left| \frac{y(x) - \sqrt{x}}{\sqrt{x}} \right|,$$

as y ranges over the class of polynomials of degree less than or equal to some $m > 0$, or over a usual class of rational approximants. Then set $y_0(x) = \tilde{y}(x)$. (See [2] and [3].)

* Supported in part by NSF Grant GP-7624.

(ii) Find that $y^*(x) = y_0(x)$ which minimizes

$$\max_{x\in[a,b]} \left| \frac{y_n(x) - \sqrt{x}}{\sqrt{x}} \right|.$$

This approach was studied in [5] and it was shown that there exists a unique solution to this problem and it is the same as the solution to the problem: find $y_0(x)$ minimizing

$$\max_{x\in[a,b]} \left| \frac{y_1(x) - \sqrt{x}}{\sqrt{x}} \right| \qquad \text{(i.e., case when } n = 1\text{)}.$$

(iii) The third approach is to minimize

$$\max_{x\in[a,b]} \left| \ln \frac{y(x)}{\sqrt{x}} \right|,$$

and take the minimizing function as y_0. This approach has been used by various authors (see for example [1]) and has the advantage that analytical methods may be applied to optimize this expression.

In [7], Sterbenz and Fike showed that the (unique) solutions to (ii) and (iii) are the same and that this solution is a multiple of the solution to (i). In this paper we shall show that a somewhat similar situation prevails for a much wider class of problems.

Actually, this second result of [7] holds in general. That is, for any $f \in C[a, b]$, $f > O$ and any Haar subspace (or a usual class of rational approximants) W, it is easily seen that the best relative approximation $\tilde{y}$ to f from W, satisfying

$$\max_{x\in[a,b]} \left| \frac{f(x) - \tilde{y}(x)}{f(x)} \right| = \inf_{y\in W} \max_{x\in[a,b]} \left| \frac{f(x) - y(x)}{f(x)} \right| = \lambda,$$

is such that

$$\max_{x\in[a,b]} \left| \ln \frac{\beta\tilde{y}(x)}{f(x)} \right| = \inf_{y\in W} \max_{y\in[a,b]} \left| \ln \frac{y(x)}{f(x)} \right|$$

where $\beta = (1 - \lambda^2)^{-1/2}$. Thus, the best logarithmic approximation to f is always equal to $(1 - \lambda^2)^{-1/2}$ times the best relative approximation to f. This can easily be seen from the fact that if $y \in W$, $y > O$ on $[a, b]$ then $E_1(x) = f(x) - y(x)/f(x)$ attains its maximum (minimum) in $[a, b]$ at precisely those points where $E_2(x) = \ln(y(x)/f(x)$ attains its maximum (minimum).

In [6], the problem of finding optimal starting functions for calculating

x^α, $\alpha \in (0, 1)$, on $[a, b]$, $0 < a < b$, was studied. In that paper it was shown that there exists a unique $y^*(x)$ that minimizes

$$\varphi(y) = \max_{x\in[a,b]} \left| \frac{y_n(x) - x^\alpha}{x^\alpha} \right|, \tag{1}$$

where $y_n(x)$ is defined by ($\beta = 1/\alpha$)

$$y_k(x) = \frac{(\beta - 1)\, y_{k-1}^{\beta}(x) + x}{\beta y_{k-1}^{\beta-1}(x)}, \qquad k = 1, 2, \ldots, n,$$
$$y_0(x) = y(x). \tag{2}$$

It was further shown that the same y^* minimizes (1) for all $n \geqslant 1$ and that this solution is uniquely determined by a finite set of points

$$x_1 < x_2 < \cdots < x_{m+2}$$

for which

$$\frac{y_1(x_i) - x_i^\alpha}{x_i^\alpha} = \max_{x\in[a,b]} \left| \frac{y_1(x) - x^\alpha}{x^\alpha} \right| \tag{3}$$

and

$$\operatorname{sgn}(y^*(x_i) - x_i^\alpha) = (-1)^{i+1} \operatorname{sgn}(y^*(x_1) - x_1^\alpha), \qquad i = 1, \ldots, m + 2,$$

where the number, $m + 1$, of alternations is the same as in the standard uniform approximation problem with the same class of approximants.

We shall show that y^*, the solution to (1), is a positive multiple of $\tilde{y}$, the unique function minimizing

$$\max_{x\in[a,b]} \left| \frac{y(x) - x^\alpha}{x^\alpha} \right|. \tag{4}$$

Also, we shall show that y^* is a positive multiple of $\hat{y}$ the unique function which minimizes

$$\max_{x\in[a,b]} \left| \ln \frac{y(x)}{x^\alpha} \right|. \tag{5}$$

Finally we show that $\hat{y} = y^*$ if and only if $\alpha = \frac{1}{2}$.

In doing this, we shall have a greatly simplifed method of finding the y^* minimizing (1) over that proposed in [6]. Our method of proof is quite different from that used in [7].

2. Main Results

Let

$$R_0(x) = \frac{y(x) - x^\alpha}{x^\alpha},$$

$$R_1(x) = \frac{y_1(x) - x^\alpha}{x^\alpha},$$

where $y(x)$ is any fixed positive approximant and $y_1(x)$ is given by (2). Note that $R_1(x) \geqslant 0$ for all x, equality holding at z if and only if $y_1(z) = z^\alpha$. Using (2), we see that ($\beta = 1/\alpha$)

$$R_1(x) = \frac{(\beta - 1)(R_0(x) + 1)^\beta - \beta(R_0(x) + 1)^{\beta-1} + 1}{\beta(R_0(x) + 1)^{\beta-1}}. \tag{6}$$

Letting

$$\delta(x) = \ln \frac{y(x)}{x^\alpha} = \ln[1 + R_0(x)],$$

we have

$$R_1(x) = \frac{(\beta - 1)\, e^{\beta\delta(x)} + 1}{\beta e^{(\beta-1)\delta(x)}} - 1. \tag{7}$$

Now let $x_1 < x_2 < \cdots < x_{m+2}$ be a set of characterizing extremal points for $\tilde{y}$, the minimizing solution of (4), that is,

$$\left| \frac{\tilde{y}(x_i) - {x_i}^\alpha}{{x_i}^\alpha} \right| = \max_{x \in [a,b]} \left| \frac{\tilde{y}(x) - x^\alpha}{x^\alpha} \right|$$

and

$$\operatorname{sgn}(\tilde{y}(x_i) - {x_i}^\alpha) = (-1)^{i+1} \operatorname{sgn}(\tilde{y}(x_1) - {x_1}^\alpha), \qquad i = 1,\ldots, m + 2.$$

Next, let γ be a number satisfying

$$\min_{i=1,2} \left(\frac{{x_i}^\alpha}{\tilde{y}(x_i)} \right) < \gamma < \max_{i=1,2} \left(\frac{{x_i}^\alpha}{\tilde{y}(x_i)} \right). \tag{8}$$

Set

$$R_{0,\gamma}(x) = \frac{\gamma\tilde{y}(x) - x^\alpha}{x^\alpha}, \tag{9}$$

$$R_{1,\gamma}(x) = \frac{y_1(x) - x^\alpha}{x^\alpha}, \tag{10}$$

where

$$y_1(x) = \frac{(\beta - 1)[\gamma\tilde{y}(x)]^\beta + x}{\beta[\gamma\tilde{y}(x)]^{\beta-1}}.$$

We claim that there exists a unique γ^* in the range (8), for which

$$R_{1\gamma^*}(x_1) = R_{1\gamma^*}(x_2). \tag{11}$$

Using the fact that

$$\varphi(t) = \frac{(\beta - 1)\, e^{\beta t} + 1}{\beta e^{(\beta-1)t}} - 1 \tag{12}$$

is nonnegative, vanishes at 0, strictly decreases for $t < 0$, and strictly increases for $t > 0$, we obtain that $R_{1\gamma}(x_1)$ is a continuous function of $\gamma(\gamma > 0)$ which takes on the value 0 when $\gamma = x_1^\alpha/\tilde{y}(x_1)$ and increases strictly as γ moves away from the value $x_1^\alpha/\tilde{y}(x_1)$. Likewise, $R_{1\gamma}(x_2)$ is a continuous function of $\gamma(\gamma > 0)$ which takes on the value 0 at $x_2^\alpha/\tilde{y}(x_2)$ and increases strictly as γ moves away from this value, implying the existence and uniqueness of γ^*.

Next, we observe that

$$R_{1\gamma^*}(x_{i+1}) = R_{1\gamma^*}(x_i), \qquad i = 1, 2, \ldots, m + 1,$$

and

$$\operatorname{sgn}(\gamma^*\tilde{y}(x_i) - x_i^\alpha) = \operatorname{sgn}(\tilde{y}(x_i) - x_i^\alpha), \qquad i = 1, \ldots, m + 2.$$

The first equality follows from

$$\frac{\tilde{y}(x_i) - x_i^\alpha}{x_i^\alpha} = \frac{\tilde{y}(x_{i+2}) - x_{i+2}^\alpha}{x_{i+2}^\alpha}, \qquad i = 1, \ldots, m$$

and (11). The second equality follows from the restriction (8) on γ. Also, if $z \in [a, b]$, then

$$\min_{i=1,2}\left(\frac{\tilde{y}(x_i)}{x_i^\alpha}\right) \leqslant \frac{\tilde{y}(z)}{z^\alpha} \leqslant \max_{i=1,2}\left(\frac{\tilde{y}(x_i)}{x_i^\alpha}\right),$$

implying

$$\min_{i=1,2}(R_{0\gamma^*}(x_i)) \leqslant R_{0\gamma^*}(z) \leqslant \max_{i=1,2}(R_{0\gamma^*}(x_i)),$$

or

$$\min_{i=1,2}(\delta_{\gamma^*}(x_i)) \leqslant \delta_{\gamma^*}(z) \leqslant \max_{i=1,2}(\delta_{\gamma^*}(x_i)),$$

where

$$\delta_{\gamma^*}(x) = \ln[1 + R_{0\gamma^*}(x)]. \tag{13}$$

Thus,

$$\varphi(\delta_{\gamma^*}(z)) \leqslant \max_{i=1,2} \{\varphi(\delta_{\gamma^*}(x_i))\}.$$

But

$$\varphi(\delta_{\gamma^*}(x_1)) = \varphi(\delta_{\gamma^*}(x_2)) = R_{1\gamma^*}(x_1),$$

so that

$$0 \leqslant R_{1\gamma^*}(z) \leqslant R_{1\gamma^*}(x_1).$$

Thus, by the theory developed in [6], see (3), it follows that $\gamma^*\tilde{y}(x)$ is the unique function minimizing (1).

Now suppose $\gamma > 0$ is outside the interval (8). Then by a reasoning similar to that used above, it can be shown that

$$R_{1\gamma}(x_1) \neq R_{1\gamma}(x_2).$$

Using this fact and equating $R_{1\gamma^*}(x_1)$ and $R_{1\gamma^*}(x_2)$, we find that

$$\begin{aligned} \gamma^* &= \left[\frac{x_1{}^{\alpha}x_2\,\tilde{y}^{\beta-1}(x_1) - x_1x_2{}^{\alpha}\tilde{y}^{\beta-1}(x_2)}{(\beta-1)[\tilde{y}(x_1)\,\tilde{y}(x_2)]^{\beta-1}[x_2{}^{\alpha}\tilde{y}(x_1) - x_1{}^{\alpha}\tilde{y}(x_2)]}\right]^{\alpha} \\ &= \left[\frac{(1+\lambda)^{\beta-1} - (1-\lambda)^{\beta-1}}{2(\beta-1)\,\lambda(1-\lambda^2)^{\beta-1}}\right]^{a}, \end{aligned} \tag{14}$$

where $\lambda = \| (\tilde{y}(x) - x^{\alpha})/x^{\alpha} \|$.

Combining, we have

THEOREM 1. *Let $\alpha \in (0, 1)$, $\beta = 1/\alpha$ and $0 < a < b$. Then there exists a unique polynomial or rational approximant $y^*(x)$, minimizing*

$$\varphi(y) = \max_{x\in[a,b]} \left| \frac{y_n(x) - x^{\alpha}}{x^{\alpha}} \right|,$$

where

$$y_k(x) = \frac{(\beta-1)\,y_{k-1}^{\beta}(x) + x}{\beta y_{k-1}^{\beta-1}(x)}, \qquad k = 1, 2, \ldots, n,$$

$$y_0(x) = y(x),$$

and where $y(x)$ varies over the class of approximants. Moreover, $y^*(x) = \gamma^*\tilde{y}(x)$, where $\tilde{y}(x)$ is the unique approximant minimizing

$$\max_{x\in[a,b]} \left| \frac{y(x) - x^{\alpha}}{x^{\alpha}} \right|$$

and γ^* is given by (14).

By the same methods, we prove

THEOREM 2. *Let* $\alpha \in (0, 1)$ *and* $0 < a < b$. *Then there is a unique polynomial or rational approximant* $\hat{y}(x)$, *minimizing*

$$\Phi(y) = \max_{x\in[a,b]} \left| \ln \frac{y(x)}{x^{\alpha}} \right| \tag{15}$$

as $y(x)$ *varies over the positive approximants. We have*

$$\hat{y}(x) = \hat{\gamma}\tilde{y}(x), \tag{16}$$

where

$$\hat{\gamma} = \left[\frac{(x_1x_2)^{\alpha}}{\tilde{y}(x_1)\,\tilde{y}(x_2)}\right]^{1/2} = \left(\frac{1}{1-\lambda^2}\right)^{1/2} \tag{17}$$

and $\lambda, \tilde{y}(x), x_1$ *and* x_2 *are as described above. Furthermore,* $\hat{\gamma} = \gamma^*$ *if and only if* $\alpha = \frac{1}{2}$.

Proof of Theorem 2: In proving this Theorem, we shall not use the general fact that the best logarithmic approximation to a given positive function $f \in C[a, b]$ is $(1 - \lambda^2)^{-1/2}$ times the best relative approximation to f, where λ is the relative error. Instead, we note that the existence and uniqueness of $\hat{y}$, minimizing (15), follow by the usual arguments. (16) and (17) follow by exactly the same methods used to prove Theorem 1. As to the last statement of Theorem 2, we note that equality, when $\alpha = \frac{1}{2}$, was shown in [7]. To show that equality cannot occur otherwise, we simply must show that if φ [of (12)] satisfies

$$\varphi(t) = \varphi(-t), \tag{18}$$

for some $t > 0$, then $\beta = 2$. The equality (18) may be simplified to

$$(\beta - 1)\sinh t - \sinh(\beta - 1)t = 0,$$

for which we wish to show that $\beta = 2$ is the only solution larger than 1. Looking at

$$\psi(\beta) = (\beta - 1)\sinh t - \sinh(\beta - 1)t,$$

we see that $\psi(1) = \psi(2) = 0$ and

$$\psi''(\beta) = -t^2 \sinh(\beta - 1)t.$$

Since $\psi''(\beta) < 0$, for $\beta > 1$, we have by Rolle's Theorem, that ψ vanishes at 1 and 2, and nowhere else in $[1, \infty)$.

SUMMARY

This work greatly simplifies finding the optimal starting function y^*, minimizing (1). To date, the method of calculating y^* had consisted of a modified Remes algorithm in which one had to solve a nonlinear system of equations by means of Newton's method of higher order. Using the above results, we can calculate $y^*(x)$ by calculating the best relative approximation $\tilde{y}$ to x^α and multiplying by a constant depending upon the relative error λ.

$$\lambda = \left| \frac{\tilde{y}(x) - x^\alpha}{x^\alpha} \right|.$$

Added in proof: This problem has also been solved by D. L. Phillips. See D. L. Phillips, Generalized logarithmic error and Newton's method for the m-th root, *Math. Comp.*, **24** (1970).

REFERENCES

1. W. J. CODY, Double-precision square root for the CDC-3600, *Comm. ACM* **7** (1964), 715–718.
2. J. EVE, Starting approximations for the iterative calculation of square roots, *Comput. J.* **6** (1963), 274–276.
3. C. T. FIKE, Starting approximations for square-root calculation on IBM System/360, *Comm. ACM* **9** (1966), 297–299.
4. R. F. KING AND D. L. PHILLIPS, The logarithmic error and Newton's method for the square root, *Comm. ACM* **12** (1969), 87–88.
5. D. G. MOURSUND, Optimal starting values for Newton–Raphson calculation of $\sqrt{x}$, *Comm. ACM* **10** (1967), 430–432.
6. D. G. MOURSUND AND G. D. TAYLOR, Optimal starting values for the Newton–Raphson calculation of inverses of certain functions, *SIAM J. Number. Anal.* **5** (1968), 138–150.
7. P. H. STERBENZ AND C. T. FIKE, Optimal starting approximations for Newton's Method, *Math. Comp.* **23** (1969), 313–318.