MATHEMATICS OF COMPUTATION, VOLUME 24, NUMBER 110, APRIL, 1970

Generalized Logarithmic Error and Newton's Method for the *m*th Root*

By David L. Phillips

Abstract. The problem of obtaining optimal starting values for the calculation of integer roots using Newton's method is considered. It has been shown elsewhere that if relative error is used as the measure of goodness of fit, then optimal results are not obtained when the initial approximation is a best fit. Furthermore, if the so-called logarithmic error instead of the relative error is used in the square root case, then a best initial fit is optimal for both errors. It is shown here that for each positive integer m, $m \geqq 3$, and each negative integer m, there is a certain generalized logarithmic error for which a best initial fit to the mth root is optimal. It is then shown that an optimal fit can be found by just multiplying a best relative error fit by a certain constant. Also, explicit formulas are found for the optimal initial linear fit.

Introduction. The logarithmic error $\hat{\delta}$ is defined as $\hat{\delta} = \ln(1 + \delta)$, where δ is the relative error. In [1] it is shown that if the logarithmic error is used instead of the relative error, then an initial fit to the square root that minimizes the maximum logarithmic error also minimizes the maximum logarithmic error after one or more Newton iterations. Furthermore, this best initial fit minimizes the maximum relative error after one or more Newton iterations. It is also noted that this nice property of the logarithmic error does not hold for mth roots, $m = 3, 4, 5, \ldots$, and $m = -1, -2, \ldots$. The reason that negative values of m are included here is that for negative m the Newton iteration involves only multiplication and subtraction. For machines with very slow division time, for example ILLIAC IV, it might be faster to compute $1/x$ as x^{-1} and $x^{1/m}$ as $x(x^{-1/m})^{m-1}$. It is the purpose of this note to show that for a certain generalized logarithmic error a best initial fit to the mth root minimizes the maximum relative error as well as the generalized logarithmic error after one or more Newton iterations. It will also be shown that a best generalized logarithmic fit can be obtained by simply multiplying a best relative fit by a certain constant.

Generalized Errors and Optimal Initial Fits. We will use the notation δ for the relative error and $\hat{\delta}$ for another error that can be written as $\hat{\delta} = f(\delta)$. Let y_0 be an initial approximation to the mth root of x and y_n be the nth Newton iterate. Then if δ_n is the relative error after n Newton iterations,

$$\delta_n = \frac{y_n - x^{1/m}}{x^{1/m}}, \qquad (n = 0, 1, 2, \ldots), \qquad y_{n+1} = \frac{1}{m}\left[(m - 1)y_n + \frac{x}{y_n^{m-1}}\right],$$

and

$$(1) \qquad \delta_{n+1} = \frac{1}{m}[(m - 1)(1 + \delta_n) + (1 + \delta_n)^{1-m}] - 1 \equiv g(\delta_n).$$

Received October 21, 1968, revised May 15, 1969.

AMS Subject Classifications. Primary 6520, 4140.

Key Words and Phrases. Logarithmic error, Newton's method, integer root, relative error, best fit, optimal fit.

* Work performed under the auspices of the U. S. Atomic Energy Commission.

From (1) it follows that δ_1 is not an even (or odd) function of δ_0 (except for $m = -1$). This in turn implies that a best initial fit does not in general lead to a best fit after $n \geqq 1$ Newton iterations. This is because initial errors with equal magnitude but opposite sign will not lead to errors of equal magnitude after one or more Newton iterations. However, if we can find an error $\hat{\delta} = f(\delta)$ such that $\hat{\delta}_{n+1}$ is an even function of $\hat{\delta}_n$, then initial errors of equal magnitude will lead to errors of equal magnitude after one or more Newton iterations. It will be shown in Theorem 1 below that if in addition $f(\delta)$ is monotone increasing and $f(0) = 0$, then a best initial fit yields best iterated fits after any number of Newton iterations.

A *best fit* over a given range, say $[a, b]$, is understood to be an approximating function (of a certain form) with an error curve which minimizes the maximum magnitude of the error. An initial fit will be called *optimal* if all of its Newton iterates are best fits. It has been shown in [2] (also [1]) that for the relative error an initial fit to $x^{1/m}$ which has the property that its first Newton iterate is a best fit is optimal.

We now prove the following:

THEOREM 1. *Let $\hat{\delta}$ be an error related to the relative error δ by the expression $\hat{\delta} = f(\delta)$, where f is a continuous monotone increasing function with $f(0) = 0$. Further, let $\hat{\delta}$ have the property that $\hat{\delta}_{n+1}$ is an even function of $\hat{\delta}_n$. Then a best initial fit to the mth root using $\hat{\delta}$ is optimal for both errors, $\hat{\delta}$ and δ.*

Proof. We will assume for convenience that $m \geqq 2$. The case $m \leqq -1$ is treated similarly.

We will first show that $\hat{\delta}_n$ and δ_n, $n = 1, 2, \ldots$ are nonnegative even functions of $\hat{\delta}_0$, monotone increasing for $\hat{\delta}_0 \geqq 0$. It follows from (1) that

$$\hat{\delta}_{n+1} = f(\delta_{n+1}) = f(g(\delta_n)) = f(g(f^{-1}(\hat{\delta}_n))) \equiv F(\hat{\delta}_n),$$

where F is an even function of $\hat{\delta}_n$, f^{-1}, the inverse function of f, is monotone increasing, and $f^{-1}(0) = 0$. Further, it follows from (1) that $g(0) = 0$ and that

$$g'(\delta_n) = \frac{d\delta_{n+1}}{d\delta_n} = \frac{m-1}{m}[1 - (1 + \delta_n)^{-m}] < 0 \quad \text{for } -1 < \delta_n < 0,$$

$$> 0 \quad \text{for } \delta_n > 0.$$

Thus, g is monotone decreasing for $\delta_n \leqq 0$ and monotone increasing for $\delta_n \geqq 0$; consequently $g \geqq 0$. It now readily follows that $F(z)$ is monotone increasing for $z \geqq 0$, and that $F(0) = 0$. Introducing the notation

$$F_1(z) \equiv F(z), \qquad F_2(z) \equiv F(F(z)), \ldots,$$

and $G_n(z) \equiv f^{-1}(F_n(z))$, we see that the $F_n(z)$ and $G_n(z)$ are nonnegative even functions, monotone increasing for $z \geqq 0$, and that

$$\hat{\delta}_n = F_n(\hat{\delta}_0), \qquad \delta_n = f^{-1}(\hat{\delta}_n) = G_n(\hat{\delta}_0), \qquad n = 1, 2, \ldots.$$

We next show that maximal initial $\hat{\delta}$ errors give maximal iterated $\hat{\delta}$ errors. Let y_0 be an initial approximation and $\hat{\delta}_{0M}$ be a maximal initial $\hat{\delta}$ error, i.e., $|\hat{\delta}_{0M}| = \max_{x\in[a,b]} |\hat{\delta}_0(x)|$. If $\hat{\delta}_{nM}$ is a maximal $\hat{\delta}$ error after n iterations,

$$\begin{aligned}\hat{\delta}_{nM} &= \max_{x\in[a,b]} F_n(\hat{\delta}_0(x)) = \max_{x\in[a,b]} F_n(|\hat{\delta}_0(x)|) = F_n\left(\max_{x\in[a,b]} |\hat{\delta}_0(x)|\right) \\ &= F_n(|\hat{\delta}_{0M}|) = F_n(\hat{\delta}_{0M}).\end{aligned}$$

Now let y_0' be another initial fit with maximal initial error $\hat{\delta}'_{0M}$ and maximal $\hat{\delta}$ error after n iterations $\hat{\delta}'_{nM}$. Then $\hat{\delta}'_{nM} = F_n(\hat{\delta}'_{0M})$. If $|\hat{\delta}'_{0M}| < |\hat{\delta}_{0M}|$, it follows that

$$\hat{\delta}'_{nM} = F_n(\hat{\delta}'_{0M}) = F_n(|\hat{\delta}'_{0M}|) < F_n(|\hat{\delta}_{0M}|) = F_n(\hat{\delta}_{0M}) = \hat{\delta}_{nM},$$

i.e., the better initial approximation gives the better approximation after n iterations. Thus a best initial $\hat{\delta}$ fit y_0 gives best iterated fits y_n.

A similar argument with δ_n replacing $\hat{\delta}_n$ $(n = 1, 2, \ldots)$ and G_n replacing F_n shows that each iterate y_n of a best initial $\hat{\delta}$ fit y_0 is a best δ fit. This completes the proof of the theorem.

COROLLARY 1. *A best initial fit to* $1/x$ *using the relative error is optimal.*

Proof. In this case $(m = -1)$ it follows from (1) that $\delta_{n+1} = -\delta_n^2$. Thus we simply let $\hat{\delta} = \delta$ in the theorem.

Generalized Logarithmic Error. We now seek an error $\hat{\delta} = f(\delta)$ satisfying the conditions of Theorem 1. In particular we are looking for an error $\hat{\delta} = f(\delta)$ such that $\hat{\delta}_{n+1} = f(\delta_{n+1})$ is an even function of $\hat{\delta}_n$ (as it turns out we find f implicitly through the inverse function f^{-1}). As in Theorem 1, we can write $\hat{\delta}_{n+1}$ as

$$\hat{\delta}_{n+1} = f(\delta_{n+1}) = f(g(\delta_n)) = f(g(f^{-1}(\hat{\delta}_n))) \equiv F(\hat{\delta}_n).$$

Notice that $F(\hat{\delta}_n)$ will be an even function of $\hat{\delta}_n$ if $g(f^{-1}(\hat{\delta}_n))$ is an even function of $\hat{\delta}_n$, i.e., if

$$\frac{1}{m}[(m-1)(1+f^{-1}(\hat{\delta}_n)) + (1+f^{-1}(\hat{\delta}_n))^{1-m}]$$
$$= \frac{1}{m}[(m-1)(1+f^{-1}(-\hat{\delta}_n)) + (1+f^{-1}(-\hat{\delta}_n))^{1-m}],$$

or

$$(2) \qquad \begin{aligned}(m-1)[1+f^{-1}(\hat{\delta}_n) - (1+f^{-1}(-\hat{\delta}_n))]\\ = (1+f^{-1}(-\hat{\delta}_n))^{1-m} - (1+f^{-1}(\hat{\delta}_n))^{1-m}.\end{aligned}$$

In order to find a function f such that (2) is satisfied, we find it convenient to put

$$(3) \qquad 1 + f^{-1}(\hat{\delta}_n) = \exp[r(\hat{\delta}_n) + s(\hat{\delta}_n)],$$

where r is an arbitrary even function and s is an arbitrary odd function. Substituting (3) into (2) we get

$$(m-1)e^r(e^s - e^{-s}) = e^{r(1-m)}(e^{(m-1)s} - e^{-(m-1)s}),$$

or

$$e^{mr} = \sinh(m-1)s/((m-1)\sinh s).$$

Then, since $\delta_n = f^{-1}(\hat{\delta}_n)$,

$$(1+\delta_n)^m = e^{mr}e^{ms} = e^{ms}\sinh(m-1)s/((m-1)\sinh s),$$

with $s(\hat{\delta}_n)$ an arbitrary odd function. Thus there are an infinite number of solutions. An obvious choice for s is the simplest of all odd functions, $s = \hat{\delta}_n$. We now define the error $\hat{\delta}$ by means of the equation

$$(1 + \delta)^m = e^{m\hat{\delta}} \sinh(m - 1)\hat{\delta}/((m - 1)\sinh \hat{\delta}),$$

or

$$\begin{aligned} \delta = f^{-1}(\hat{\delta}) &= e^{\hat{\delta}}\left(\frac{\sinh(m - 1)\hat{\delta}}{(m - 1)\sinh \hat{\delta}}\right)^{1/m} - 1 \\ &= \left(\frac{e^{2(m-1)\hat{\delta}} + e^{2(m-2)\hat{\delta}} + \cdots + e^{2\hat{\delta}}}{m - 1}\right)^{1/m} - 1, \qquad m > 0 \\ &= \left(\frac{e^{2m\hat{\delta}} + e^{2(m+1)\hat{\delta}} + \cdots + 1}{1 - m}\right)^{1/m} - 1, \qquad m < 0. \end{aligned} \tag{4}$$

Equation (4) implicitly defines $\hat{\delta} = f(\delta)$ and shows that f is a monotone increasing function and that $f(0) = 0$. For $m = 2$, $\hat{\delta} = f(\delta)$ reduces to the logarithmic error. Thus we will call $\hat{\delta} = f(\delta)$ the generalized logarithmic error. Notice that for small $\hat{\delta}$ we have $\hat{\delta} \approx \ln(1 + \delta) \approx \delta$ so that the relative, logarithmic, and generalized logarithmic errors are essentially the same for sufficiently small δ. For $m = 3$ we can easily express $\hat{\delta}$ in terms of δ. The result in this case is

$$\hat{\delta} = \tfrac{1}{2}\ln[((1 + 8(1 + \delta)^3)^{1/2} - 1)/2].$$

The generalized logarithmic error satisfies the conditions of Theorem 1. We thus have the following:

COROLLARY 2. *A best initial fit to the mth root using the generalized logarithmic error is optimal for both the generalized logarithmic and relative errors.*

Best Fits Using Errors of the Form $\hat{\delta} = f(\delta)$. When $f(\delta)$ is a continuous monotone increasing function with $f(0) = 0$, best fits using the error $\hat{\delta} = f(\delta)$ are related to best fits using the relative error in a simple way. To be more precise we have the following:

THEOREM 2. *Let $F(x)$ be a bounded continuous function for which either $F(x) > 0$ or $F(x) < 0$ holds in $[a, b]$, and $\mathscr{F}$ be a class of bounded continuous functions such that for real k, $kz(x) \in \mathscr{F}$ whenever $z \in \mathscr{F}$. Let $y \in \mathscr{F}$ be a best fit to $F(x)$ using the relative error δ. Further, let $\sigma = \max_{x\in[a,b]} \delta(x)$ be the maximum relative error of y. Then $cy(x)$ is a best fit to $F(x)$ for the error $\hat{\delta} = f(\delta)$, where $c > 0$ is the constant satisfying the equation*

$$f(c(1 + \sigma) - 1) = -f(c(1 - \sigma) - 1),$$

and f is a continuous monotone increasing function with $f(0) = 0$.

Proof. First we introduce the notation δ_z and $\hat{\delta}_z$ to denote the relative and $\hat{\delta}$ errors, respectively, of the approximation $z(x)$. It follows from the definition of relative error that for any constant k,

$$\delta_{kz} = k(1 + \delta_z) - 1.$$

It easily follows that for a best fit, y, using the relative error,

$$-\min_{x\in[a,b]} \delta_y = \max_{x\in[a,b]} \delta_y \equiv \sigma.$$

For otherwise $(1 - \varepsilon)y(x)$, for sufficiently small ε of the appropriate sign, would give a better fit.

Now assume that for the $\hat{\delta}$ error there exists a better fit, $g(x)$, to $F(x)$ than $cy(x)$. Since f is monotone increasing we have

$$\max_x f(\delta_{cy}) = f\left(\max_x \delta_{cy}\right) = f\left(c\left(1 + \max_x \delta_y\right) - 1\right)$$
$$= f\big(c(1 + \sigma) - 1\big),$$

and similarly

$$\min_x f(\delta_{cy}) = f\big(c(1 - \sigma) - 1\big).$$

But $f\big(c(1 + \sigma) - 1\big) = -f\big(c(1 - \sigma) - 1\big)$ so that

$$f\big(c(1 - \sigma) - 1\big) = \min_x f(\delta_{cy}) < \hat{\delta}_g \equiv f(\delta_g) < \max_x f(\delta_{cy}) = f\big(c(1 + \sigma) - 1\big).$$

It then follows that

$$c(1 - \sigma) - 1 < \delta_g < c(1 + \sigma) - 1, \quad \text{or} \quad -\sigma < \frac{1}{c}(1 + \delta_g) - 1 \equiv \delta_{g/c} < \sigma,$$

which contradicts y being a best fit to $F(x)$ for the relative error. Hence the assumption that there exists a better fit than $cy(x)$ to $F(x)$ for the $\hat{\delta}$ error has led to a contradiction.

COROLLARY 3. *For the generalized logarithmic error*

$$c = \frac{1}{(1 - \sigma^2)^{1/2}}$$
$$(5) \qquad \times \left(\frac{\left(\frac{1 + \sigma}{1 - \sigma}\right)^{(m-2)/2} + \left(\frac{1 + \sigma}{1 - \sigma}\right)^{(m-3)/2}\left(\frac{1 - \sigma}{1 + \sigma}\right) + \cdots + \left(\frac{1 - \sigma}{1 + \sigma}\right)^{(m-2)/2}}{(m - 1)} \right)^{1/m}$$
$$= \frac{1}{1 + \sigma}\left(\frac{\left(\frac{1 + \sigma}{1 - \sigma}\right)^{(m-1)} + \left(\frac{1 + \sigma}{1 - \sigma}\right)^{(m-2)} + \cdots + \left(\frac{1 + \sigma}{1 - \sigma}\right)}{m - 1} \right)^{1/m}.$$

Proof. Put $\hat{\sigma} = f\big(c(1 + \sigma) - 1\big)$. Then $-\hat{\sigma} = f\big(c(1 - \sigma) - 1\big)$, $f^{-1}(\hat{\sigma}) = c(1 + \sigma) - 1$, and $f^{-1}(-\hat{\sigma}) = c(1 - \sigma) - 1$. It follows from (4) that

$$c(1 + \sigma) - 1 = e^{\hat{\sigma}}\left(\frac{\sinh(m - 1)\hat{\sigma}}{(m - 1)\sinh \hat{\sigma}}\right)^{1/m} - 1,$$

and

$$c(1 - \sigma) - 1 = e^{-\hat{\sigma}}\left(\frac{\sinh(m - 1)\hat{\sigma}}{(m - 1)\sinh \hat{\sigma}}\right)^{1/m} - 1,$$

so that

$$e^{2\hat{\sigma}} = \frac{1 + \sigma}{1 - \sigma}, \qquad \hat{\sigma} = \frac{1}{2}\ln\left(\frac{1 + \sigma}{1 - \sigma}\right).$$

Thus

$$c = \frac{e^{\hat{\sigma}}}{1 + \sigma}\left(\frac{\sinh(m - 1)\hat{\sigma}}{(m - 1)\sinh \hat{\sigma}}\right)^{1/m}$$

$$= \frac{1}{(1 - \sigma^2)^{1/2}}\left(\frac{\left(\dfrac{1 + \sigma}{1 - \sigma}\right)^{(m-1)/2} - \left(\dfrac{1 - \sigma}{1 + \sigma}\right)^{(m-1)/2}}{(m - 1)\left(\left(\dfrac{1 + \sigma}{1 - \sigma}\right)^{1/2} - \left(\dfrac{1 - \sigma}{1 + \sigma}\right)^{1/2}\right)}\right)^{1/m},$$

from which (5) follows.

Combining the results of Theorem 2 and Corollaries 2 and 3 we can state the following:

THEOREM 3. *An optimal initial approximation to $x^{1/m}$ using either the generalized logarithm or relative error is obtained by multiplying a best initial approximation using the relative error by the constant c given by Eq.* (5).

Example. We now use the results of Theorem 3 to find the optimal linear fit $y_0 = A + Bx$ to $x^{1/m}$ on the interval $[a, b]$. We first find the best initial fit, $\bar{y}_0 = \bar{A} + \bar{B}x$, for the relative error. It is easy to show that the best fit satisfies the conditions

$$\delta(a) = \delta(b) = -\delta(\bar{x})$$

where

$$\left.\frac{d\delta(x)}{dx}\right|_{x=\bar{x}} = 0.$$

Solving these equations is straightforward and yields

$$\bar{A} = \frac{2a^{1/m}}{(1 + ak) + \dfrac{m}{m - 1}(ak(m - 1))^{1/m}}, \qquad \bar{B} = k\bar{A},$$

where

(6) $$k \equiv \frac{b^{1/m} - a^{1/m}}{ba^{1/m} - ab^{1/m}}.$$

Also we get

$$1 - \sigma = \frac{\bar{A}m}{m - 1}((m - 1)k)^{1/m}, \qquad 1 + \sigma = \bar{A}\frac{1 + ak}{a^{1/m}}.$$

Thus the optimal linear fit is $y_0 = A + Bx$ where

(7) $$A = \frac{(a/(m - 1))^{1/m}}{(1 + ak)}\left(\left(\frac{(1 + ak)(m - 1)}{m((m - 1)ak)^{1/m}}\right)^{m-1} + \cdots + \frac{(1 + ak)(m - 1)}{m((m - 1)ak)^{1/m}}\right)^{1/m},$$

(8) $B = kA.$

For $m = 2$ the fit reduces to

$$y_0 = \frac{(ab)^{1/2} + x}{(ab)^{1/8}(2(\sqrt{a} + \sqrt{b}))^{1/2}}$$

which agrees with the results already found in [1].

Table 1 gives the values of A and B for several values of m and typical intervals $[a, b]$.

TABLE 1

	A	B
$m = 2$ $[\frac{1}{4}, 1]$	.3432945240	.6865890480
$m = 3$ $[\frac{1}{8}, 1]$	.4541610792	.6055481056
$m = 4$ $[\frac{1}{16}, 1]$	.5083290509	.5809474868
$m = 5$ $[\frac{1}{32}, 1]$	.5411774362	.5772559320
$m = 6$ $[\frac{1}{64}, 1]$	.5644226063	.5826297871
$m = -1$ $[\frac{1}{2}, 1]$	2.823529412	−1.882352941
$m = -2$ $[\frac{1}{4}, 1]$	2.130151160	−1.217229234
$m = -3$ $[\frac{1}{8}, 1]$	1.898387403	−1.012473282
$m = -4$ $[\frac{1}{16}, 1]$	1.778282355	−.9178231511
$m = -5$ $[\frac{1}{32}, 1]$	1.700553087	−.8637729968

Added in Proof. When $m = 2$ Theorem 3 reduces to a result essentially the same as one given in a recent paper of Sterbenz and Fike [3]. Also, in a forthcoming paper [4] Taylor gives a different (and independent) proof that the optimal initial approximation to $x^{1/m}$ is a constant multiple of the best initial relative error approximation.

Argonne National Laboratory
Argonne, Illinois 60439

1. RICHARD F. KING & DAVID L. PHILLIPS, "The logarithmic error and Newton's method for the square root," *Comm. ACM*, v. 12, 1969, pp. 87–88.

2. D. G. MOURSUND & G. D. TAYLOR, "Optimal starting values for the Newton-Raphson calculation of inverses of certain functions," *SIAM J. Numer. Anal.*, v. 5, 1968, pp. 138–150. MR 37 #1074.

3. P. H. STERBENZ & C. T. FIKE, "Optimal starting approximations for Newton's method," *Math. Comp.*, v. 23, 1969, pp. 313–318.

4. G. D. TAYLOR, "Optimal starting approximations for Newton's method," *J. Approximation Theory*. (To appear.)