REPORTS ON COMPUTATIONAL MATHEMATICS, NO. 70/1995, THE UNIVERSITY OF IOWA

A Note on Brent's Rootfinding Method

Florian A. Potra*
Yixun Shi†

May 1995

Abstract

The behavior of the enclosing intervals produced by Brent's rootfinding method is analyzed. It is shown that for any positive tolerance there is a smooth function for which Brent's method reduces to the bisection method at each step until termination. Moreover, a cubic equation having a simple zero is constructed such that the diameters of the enclosing intervals produced by Brent's method with zero tolerance do not converge to zero.

Key Words : enclosing method, enclosing interval, simple root,linear interpolation, inverse quadratic interpolation.

AMS Subject Classification : 65H05

*Department of Mathematics, the University of Iowa, Iowa City, IA 52242. The work of this author was supported in part by NSF, Grant DMS 9305760, and by DOE, Grant DE-FG02-94ER 61855
†Dept. of Math. & Comp. Sci., Bloomsburg University, Bloomsburg, PA 17815

1 Introduction

Brent's method [5] is one of the most efficient algorithms for finding a zero of a continuous function f that changes sign at the endpoints of a given interval $[a, b]$. It produces a sequence of intervals $\{[a_n, b_n]\}$ such that

$$[a_{n+1}, b_{n+1}] \subseteq [a_n, b_n], \quad f(a_n)f(b_n) < 0, \quad n = 1, 2, \ldots \tag{1}$$

and a sequence of points $x_n \in \{a_{n+1}, b_{n+1}\}$ representing "the best current estimate" of a zero x_* of f. The algorithm terminates when

$$0.5(b_n - a_n) < \delta_n \tag{2}$$

where

$$\delta_n = 2\epsilon|u_n| + t \tag{3}$$

with

$$u_n \in \{a_n, b_n\}, \quad |f(u_n)| = \min\{|f(a_n)|, |f(b_n)|\}. \tag{4}$$

As in Peters and Wilkinson [7] and Brent [5], the quantity $2\delta_n$ represents the desired tolerance which is a combination of a relative tolerance (4ϵ) and an absolute tolerance $(2t)$. At termination the algorithm returns the approximate zero $\bar{x} = u_n$. Since $f(a_n)f(b_n) < 0$ the function f has a zero $x_* \in (a_n, b_n)$ so that

$$|\bar{x} - x_*| \le d_n \le 2\delta_n = 4\epsilon|\bar{x}| + 2t. \tag{5}$$

From [5] it follows that if

$$\delta := \min\{2\epsilon|x| + t; \ \ x \in [a, b]\} > 0 \tag{6}$$

then the algorithm terminates in at most $(1 + \log_2((b-a)/\delta))^2$ steps. This is substantially better than the worst case behavior of Dekker's rootfinding method [6] which may require $(b - a)/\delta$ steps to terminate.

For zero tolerance i.e., when $\epsilon = t = 0$, Brent's method produces theoretically an infinite sequence of intervals. In this case Brent [5] shows that $\{x_n\}$ converges to a zero x_* of f. Moreover if f' is Lipschitz continuous in a neighborhood of x_* and $f'(x_*) \neq 0$ then the R-order of convergence of $\{x_n\}$ is at least $\frac{1}{2}(1 + \sqrt{5}) = 1.618\ldots$. However the convergence of the diameters

$$d_n = b_n - a_n \quad n = 1, 2, \ldots \tag{7}$$

of the enclosing intervals has not been studied so far. Since the termination criterion is based on the diameters of the enclosing intervals such a study is well justified.

In the present paper we show that for any tolerances $\epsilon > 0$, $t > 0$, and any positive integer k there is a k-times continuously differentiable function having a simple zero x_* for which the enclosing diameters are reduced only at a linear rate in the sense that

$$d_{n+1} = \frac{1}{2} d_n \tag{8}$$

for $n = 1, 2, \ldots$ until termination. For this function Brent's method reduces to the bisection method in all steps except the first one.

More surprisingly, we give an example of a cubic polynomial having a simple zero on an interval $[a, b]$ such that Brent's method with zero tolerance $\epsilon = t = 0$ produces an infinite sequence of intervals for which $\lim_{n\to\infty} d_n = d_* > 0$. This shows that although Brent's method has very good practical behavior it does not belong to the class of so-called "enclosing methods" because the condition

$$\lim_{n\to\infty} (b_n - a_n) = 0 \tag{9}$$

is not satisfied. As noted in [3] Dekker's rootfinding method [6] does not belong to this class either.

For some efficient enclosing methods (which produce intervals satisfying (1) and (9) such that their diameters converge to zero superlinearly) we refer to the recent work of Alefeld-Potra [1, 2], Alefeld-Potra-Shi [3, 4], and Shi [8].

2 Brent's method

In this section we give a detailed description of Brent's method. Let $f : [a, b] \longrightarrow R$ be a continuous function such that

$$f(a)f(b) < 0, \tag{10}$$

and let us denote $[a_1, b_1] = [a, b]$. At each iteration Brent's algorithm calculates a point $x_n \in (a_n, b_n)$, and defines the new enclosing interval as

$$[a_{n+1}, b_{n+1}] = \begin{cases} [x_n, b_n] & \text{if } f(x_n)f(b_n) < 0 \\ [a_n, x_n] & \text{if } f(x_n)f(a_n) < 0 \end{cases} . \tag{11}$$

Thus the intervals generated by Brent's method satisfy (1). In what follows we concentrate on the computation of x_n. At the beginning of the n-th iteration one denotes by c_n "the best estimate of the root in the previous iteration" i.e.,

$$c_n = \begin{cases} u_{n-1} & \text{if } u_n = x_{n-1} \\ x_{n-1} & \text{if } u_n \neq x_{n-1} \end{cases} . \tag{12}$$

The above definition of c_n ensures the fact that $c_n \neq u_n$.

If the points a_n, b_n, c_n are not distinct then $\{c_n, u_n\} = \{a_n, b_n\}$ and Brent's method attempts to compute the new point by a secant step i.e.,

$$i_n = u_n - f[c_n, u_n]^{-1} f(u_n) , \tag{13}$$

where $f[c_n, u_n] := (f(c_n) - f(u_n))/(c_n - u_n)$ denotes the divided difference of f at the points c_n, u_n. We note that in this case we have

$$|i_n - u_n| \leq (b_n - a_n)/2 \tag{14}$$

because according to (4) $|f(u_n)| \leq |f(c_n)|$.

If the points a_n, b_n, c_n are distinct then Brent's algorithm attempts to compute the new point by inverse quadratic interpolation by defining

$$\bar{i}_n = \varphi_n(0) , \tag{15}$$

where φ_n is a quadratic function satisfying the relation

$$\varphi(f(a_n)) = a_n, \;\; \varphi(f(b_n)) = b_n, \;\; \varphi(f(c_n)) = c_n. \tag{16}$$

We note that (16) defines a unique quadratic function φ if and only if $f(a_n)$, $f(b_n)$, $f(c_n)$ are distinct. In this case the point $\bar{i}_n = \varphi_n(0)$ is accepted provided that it satisfies the following condition

$$\bar{i}_n \in (a_n, b_n), \;\; |\bar{i}_n - u_n| \leq \frac{3}{4}(b_n - a_n) , \tag{17}$$

which ensures sufficient reduction of the enclosing interval.

With the above notation the computation of x_n can be described in the following three basic steps:

<u>Step 1</u> (Computation of a point $i_n \in (a_n, b_n)$)

S1.1 If a_n, b_n, c_n are not distinct, then compute i_n by (13) and go to Step 2.

S1.2 If $f(a_n)$, $f(b_n)$, $f(c_n)$ are distinct, then compute $\bar{i}_n$ by (15). Else set $i_n = (a_n + b_n)/2$ and go to Step 2.

S1.3 If condition (17) is satisfied then set $i_n = \bar{i}_n$. Else set $i_n = (a_n + b_n)/2$.

<u>Step 2</u> (Computation of a point $b_n'' \in (a_n, b_n)$)

S2.1 If a bisection was taken at iteration $(n-1)$ then set $e_n = (b_{n-1} - a_{n-1})/2$. Else set $e_n = |i_{n-1} - u_{n-1}|$.

S2.2 If $e_n < \delta_n$ or $e_n \leq 2|i_n - u_n|$ then set $b_n'' = (a_n + b_n)/2$. Else set $b_n'' = i_n$.

<u>Step 3</u> (Computation of x_n)

If $|b_n'' - u_n| > \delta_n$, then set $x_n = b_n''$. Else set

$$x_n = \begin{cases} u_n + \delta_n & \text{if } u_n = a_n \\ u_n - \delta_n & \text{if } u_n = b_n. \end{cases}$$

It is easily seen that if the quantity δ given by (6) is strictly positive then the diameters d_n decrease at least at a Q-linear rate of $1 - \delta/(b-a)$.

Theorem 2.1 *If $\delta > 0$ then the sequence of diameters $d_n = b_n - a_n$ produced by Brent's method satisfy*

$$d_{n+1} \leq (1 - \delta/(b-a))d_n \tag{18}$$

for $n = 1, 2. \ldots$ until termination.

While it appears that the Q-linear rate of $1-\delta/(b-a)$ cannot be improved in general, Brent has proved that his method converges faster than indicated by relation (18). Indeed he has shown that at least once in every

$$m(\delta) = 2\log_2((b-a)/\delta) \tag{19}$$

steps the diameter d_n is reduced by half, which would correspond to an R-linear rate of

$$[0.5(1-\delta/(b-a))^{m(\delta)-1}]^{1/m(\delta)} \,. \tag{20}$$

This is better than the rate of $(1-\delta/(b-a))$, but both rates tend to 1 as δ approaches zero so that the convergence to zero of the diameters produced by Brent's method with $\epsilon = t = 0$ is not guaranteed. In fact as shown in Section 4 there is a simple example for which $\lim_{n\to\infty} d_n = d_* > 0$.

3 Each step may be a bisection step

In this section, we show that for any positive integer k and any positive tolerances ϵ and t, there exists a function f and an interval $[a, b]$ such that

(i) f is k times continuously differentiable on $[a, b]$;

(ii) f has a simple root x_* in $[a, b]$;

(iii) f' is Lipschitz continuous on $[a, b]$;

(iv) in solving $f(x) = 0$ with the initial interval $[a, b]$, Brent's algorithm takes bisections in every iteration except for the initial step. That is, it produces enclosing intervals that satisfy

$$b_{n+1} - a_{n+1} = \frac{1}{2}(b_n - a_n)$$

for all $n \geq 2$ until termination.

The example is described in what follows.

Example 1 For any given positive integer k and tolerances ϵ and t, we define the function

$$f(x) = \begin{cases} \frac{(x-2)^{k+1}}{\alpha^{k+1}} - \frac{5}{11}, & 2 \leq x \leq 2+\alpha \\ -\frac{5}{11}, & \frac{1}{6}(1-5\alpha) \leq x < 2 \end{cases} \tag{21}$$

with $\alpha = 8\epsilon + 2t$, where ϵ and t are the tolerances required by Brent's method. Let

$$a = \frac{1}{6}(1-5\alpha), \quad b = 2+\alpha \tag{22}$$

It is easily seen that the function f and the interval $[a, b]$ satisfy the conditions (i), (ii) and (iii) above.

Theorem 3.1 *For any $\epsilon > 0$ and $t > 0$ consider the function (21) and the starting interval (22). Then Brent's method applied to this problem will reduce to the bisection method for all steps except the first one.*

Proof:

Initial Iteration: $u_1 = a_1$ since

$$|f(a_1)| = \frac{5}{11} < \frac{6}{11} = |f(b_1)|.$$

Hence for the initial iteration we have $c_1 = b_1$ (see [5]) so that linear interpolation at a_1 and b_1 is used. Thus $i_1 = 1$ with $|i_1 - u_1| = \frac{5}{11}(b_1 - a_1)$. It is then easy to see that x_1 is equal to 1 and hence $[a_2, b_2] = [1, 2 + \alpha]$. Furthermore, $u_2 = a_2$ because $|f(a_2)| = \frac{5}{11} < |f(b_2)|$.

Second Iteration: $c_2 = u_1 = a_1$, $u_2 = a_2 = 1$, and $b_2 = 2 + \alpha$. Since c_2, a_2, and b_2 are distinct, the linear interpolation is skipped and the inverse quadratic interpolation is tried. However, since $f(c_2) = f(a_2)$ this inverse quadratic interpolation is not well–defined. Therefore the algorithm takes

$$i_2 = \frac{a_2 + b_2}{2},$$

which implies that $b_2^{''} = i_2 = \frac{a_2+b_2}{2}$. It is obvious that $b_2^{''} < 2$. Hence

$$\begin{aligned} |b_2^{''} - u_2| &= |b_2^{''} - a_2| \\ &= |b_2^{''} - b_2| \\ &> \alpha = 8\epsilon + 2t > 2\epsilon|u_2| + t. \end{aligned}$$

Thus, the algorithm takes $x_2 = b_2^{''}$ and therefore the second iteration is a bisection step.

Similarly, a bisection is taken in the third iteration. In general, if the bisection step is taken in all the following iterations through the $(n-1)$–th iteration, and if $b_n = 2 + \alpha$, $a_n < 2$, and $\frac{a_n+b_n}{2} < 2$, then a bisection will also be taken in the n–th iteration, i.e. $x_n = \frac{a_n+b_n}{2}$. This happens because

$$|\frac{a_n + b_n}{2} - u_n| = |\frac{a_n + b_n}{2} - b_n| > \alpha > 2\epsilon|u_n| + t$$

where u_n is equal to a_n in this case. This way, the bisection is used in every iteration until we reach the point, say the N–th iteration, when $b_N = 2+\alpha$, $a_N < 2$, and $\frac{a_N+b_N}{2} \geq 2$. At this step we have again $b''_N = \frac{a_N+b_N}{2}$. Since

$$\begin{aligned} |b''_N - u_N| &= |b''_N - a_N| \\ &= |\frac{b_N - a_N}{2}| \\ &> \frac{\alpha}{2} = 4\epsilon + t > 2\epsilon|u_N| + t, \end{aligned}$$

the algorithm defines $x_N = b''_N = \frac{a_N+b_N}{2}$. Hence once again bisection is used. The next enclosing interval $[a_{N+1}, b_{N+1}]$ is now either $[a_N, \frac{a_N+b_N}{2}]$ or $[\frac{a_N+b_N}{2}, b_N]$. It is easily seen that $f(2+(\frac{10}{11})^{\frac{1}{k+1}}\alpha) = \frac{5}{11}$ and that $|f(x)| < \frac{5}{11}$ for all $x \in (2, 2+(\frac{10}{11})^{\frac{1}{k+1}}\alpha)$. It is also obvious that

$$2 \leq \frac{a_N+b_N}{2} < 2 + \frac{1}{2}\alpha < 2 + (\frac{10}{11})^{\frac{1}{k+1}}\alpha.$$

Accordingly we have

$$|f(\frac{a_N+b_N}{2})| < \frac{5}{11} \qquad \text{if} \ \ [a_{N+1}, b_{N+1}] = [a_N, \frac{a_N+b_N}{2}]$$

and

$$|f(\frac{a_N+b_N}{2})| \leq \frac{5}{11} \qquad \text{if} \ \ [a_{N+1}, b_{N+1}] = [\frac{a_N+b_N}{2}, b_N].$$

In either case $u_{N+1} = \frac{a_N+b_N}{2}$. Therefore

$$\frac{b_{N+1} - a_{N+1}}{2} \leq \frac{\alpha}{2} = 4\epsilon + t \leq 2\epsilon|u_{N+1}| + t$$

and the algorithm terminates at this iteration. It cannot terminate at any previous iteration, because for any $n \leq N$,

$$\frac{b_n - a_n}{2} > \frac{\alpha}{2} = 4\epsilon + t > 2\epsilon|a_n| + t = 2\epsilon|u_n| + t.$$

4 Diameters may not converge to zero

Next, we consider the situation when $\epsilon = t = 0$ and the effect of rounding errors is ignored. In this case, if the function f is continuously differentiable on $[a, b]$, and if x_* is a simple root, then Brent's algorithm will asymptotically stop doing bisections as mentioned by Brent [5]. Therefore, asymptotically each iteration of Brent's algorithm will take either a linear interpolation or an inverse quadratic interpolation. Consequently the sequence $\{x_n\}$ converges to x_* with the R-order of convergence at least $\frac{1}{2}(1+\sqrt{5}) = 1.618\ldots$. However, the following example shows that in this case the diameters $(b_n - a_n)$ may not converge to zero.

Example 2 Let

$$f(x) = x + \frac{1}{3}x^3 - \frac{61}{120\sqrt{5}} \tag{23}$$

and let

$$[a, b] = [\frac{1}{10\sqrt{5}}, \frac{9}{10\sqrt{5}}]. \tag{24}$$

Obviously f is continuously differentiable and $x_* = \frac{1}{2\sqrt{5}}$ is a simple root in $[\frac{1}{10\sqrt{5}}, \frac{9}{10\sqrt{5}}]$. f is strictly increasing and convex on this interval because $f'(x) = 1 + x^2$ and $f"(x) = 2x$. As mentioned before, according to Brent [5] there is an N_0 such that for all $n \geq N_0$ either a linear interpolation or an inverse quadratic interpolation is taken in the n–th iteration.

At the n–th iteration for $n \geq N_0$, let a_n, b_n, c_n, u_n be as described in Section 2. According to the definition of u_n and c_n there are only four possibilities:

(A_n) $u_n = a_n$ and $c_n = b_n$;
(B_n) $u_n = a_n$ and $c_n < a_n$;
(C_n) $u_n = b_n$ and $c_n = a_n$;
(D_n) $u_n = b_n$ and $c_n > b_n$.

In the proof of the next theorem we show that the states A_n and C_n are transient while B_n and D_n are absorbing.

Theorem 4.1 *Let $\{[a_n, b_n]\}$ be the sequence of enclosing intervals produced by applying Brent's method with $\epsilon = t = 0$ to the function (23) and starting* *interval (24). Then there exists a positive integer N_1 such that either*

$$a_n < a_{n+1} < \cdots < x_* < b_* = \cdots = b_{n+1} = b_n, \forall n > N_1$$

or

$$a_n = a_{n+1} = \cdots = a_* < x_* < \cdots < b_{n+1} < b_n, \forall n > N_1.$$

Proof: We prove first that it is sufficient to consider B_n and D_n. It is easily seen that $A_n \Rightarrow B_{n+1}$. Indeed, if A_n holds then the secant method is used, and because f is strictly increasing and convex we must have B_{n+1}. If C_n holds then the secant method is used, and therefore $C_n \Rightarrow (A_{n+1}$ or $C_{n+1})$. Because the secant method is monotonely convergent for convex and increasing functions there is $p \geq 1$ such that A_{n+p} holds. Thus B_{n+p+1} holds. We will prove now that $B_n \Rightarrow B_{n+1}$. Let us suppose that B_n holds. Since a_n, b_n, c_n are distinct, Brent's method uses inverse quadratic interpolation. Hence

$$\begin{aligned} x_n &= a_n - f(a_n)f^{-1}[f(a_n), f(b_n)] \\ &\quad + f(a_n)f(b_n)f^{-1}[f(a_n), f(b_n), f(c_n)] \ , \end{aligned} \tag{25}$$

where

$$f^{-1}[f(a_n), f(b_n)] = \frac{b_n - a_n}{f(b_n) - f(a_n)}$$

, and

$$f^{-1}[f(a_n), f(b_n), f(c_n)] = \frac{f^{-1}[f(b_n), f(c_n)] - f^{-1}[f(a_n), f(b_n)]}{f(c_n) - f(a_n)}.$$

We notice that $x_n > a_n$, because otherwise the inverse quadratic interpolation polynomial would have two local extremes, which is impossible. On the other hand, let

$$\begin{aligned} P_1(y) &= f^{-1}(f(a_n)) + (y - f(a_n))f^{-1}[f(a_n), f(b_n)] \\ &= a_n + (y - f(a_n))f^{-1}[f(a_n), f(b_n)] \end{aligned}$$

be the linear interpolation of $f^{-1}(y)$ at $f(a_n)$ and $f(b_n)$. Then

$$\begin{aligned} x_* - P_1(0) &= f^{-1}(0) - P_1(0) \\ &= (0 - f(a_n))(0 - f(b_n))f^{-1}[f(a_n), f(b_n), 0] \\ &= f(a_n)f(b_n)f^{-1}[f(a_n), f(b_n), f(x_*)]. \end{aligned} \tag{26}$$

The above relation implies

$$\begin{aligned} x_* &= P_1(0) + f(a_n)f(b_n)f^{-1}[f(a_n), f(b_n), f(x_*)] \\ &= a_n - f(a_n)f^{-1}[f(a_n), f(b_n)] \\ &\quad + f(a_n)f(b_n)f^{-1}[f(a_n), f(b_n), f(x_*)]. \end{aligned} \tag{27}$$

Therefore,

$$\begin{aligned} x_* - x_n &= f(a_n)f(b_n)(f^{-1}[f(a_n), f(b_n), f(x_*)] - f^{-1}[f(a_n), f(b_n), f(c_n)]) \\ &= f(a_n)f(b_n)(f(x_*) - f(c_n))f^{-1}[f(a_n), f(b_n), f(c_n), f(x_*)] \\ &= f(a_n)f(b_n)(f(x_*) - f(c_n))\frac{[f^{-1}(\eta)]^{(3)}}{6}\ , \end{aligned} \tag{28}$$

for some $\eta \in f([c_n, b_n])$. Since

$$[f^{-1}(\eta)]^{(3)} = \frac{-f^{(3)}(\xi)f'(\xi) + 3[f"(\xi)]^2}{[f'(\xi)]^5} = \frac{-2 + 10\xi^2}{(1+\xi^2)^5} \tag{29}$$

for $\xi = f^{-1}(\eta) \in [c_n, b_n] \subseteq [\frac{1}{10\sqrt{5}}, \frac{9}{10\sqrt{5}}]$, we must have $[f^{-1}(\eta)]^{(3)} < 0$. This implies that $x_* - x_n > 0$ because $f(a_n)f(b_n) < 0$ and $f(x_*) - f(c_n) > 0$.

Hence we deduce that $a_n < x_n < x_*$. This indicates that at the next iteration $u_{n+1} = x_n = a_{n+1}$, $c_{n+1} = a_n < a_{n+1}$, and $b_{n+1} = b_n$. Thus B_{n+1} holds and therefore we can prove by induction that

$$b_n = b_{n+1} = b_{n+2} = \dots$$

and

$$a_n < a_{n+1} < a_{n+2} < \dots < x_*.$$

A similar argument shows that $D_n \Rightarrow D_{n+1}$. Hence inverse quadratic interpolation is used and $x_n < b_n$ must hold. Properties (25), (27), (28), and (29) remain true with $\eta \in f([a_n, c_n])$ and $\xi \in [a_n, c_n]$. We deduce again that $[f^{-1}(\eta)]^{(3)} < 0$ and $f(a_n)f(b_n) < 0$. But now we have $x_* - x_n < 0$, because in this case $f(x_*) - f(c_n) < 0$. Hence we obtain

$$x_* < x_n < b_n.$$

Thus the same situation will happen in the next iteration, which proves our claim.

References

[1] G. Alefeld and F. A. Potra. On two higher order enclosing methods of J.W. Schmidt. *ZAMM*, 68(8):331–337, 1988.

[2] G. Alefeld and F. A. Potra. Some efficient methods for enclosing simple zeroes of nonlinear equations. *BIT*, 32(2):334–344, 1992.

[3] G. Alefeld an F. A. Potra and Y. Shi. On enclosing simple roots of nonlinear equations. *Mathematics of Computation*, 61(204):733–744, 1993.

[4] G. Alefeld an F. A. Potra and Y. Shi. Enclosing zeros of continuous functions. *ACM Transactions on Mathematical Software (TOMS)*, 1995. to appear.

[5] R. P. Brent. *Algorithms for minimization without derivatives.* Prentice-Hall, Englewood Cliffs, New Jersey, 1972.

[6] T. J. Dekker. Finding a zero by means of successive linear interpolation. In B. Dejon and P.Henrici, editors, *Constructive Aspects of the Fundamental Theorem of Algebra.* Wiley Interscience, 1969.

[7] G. Peters and J. H. Wilkinson. Eigenvalues of $Ax = \lambda Bx$ with band symmetric A and B. *Comp. J.*, 12:398–404, 1969.

[8] Y. Shi. Improving efficiency indices in enclosing a root of an equation. Preprint. Submitted for publication.