Final Project
Advanced Root Finding Algorithm
Math 373 – Numerical Analysis I
Bohan Song
Fall 2020

1 Background

In this paper, I am solving the problem of approximating roots of functions. Specifically, Ridder's method and improved Brent's method will be used to approximate roots and the order of convergence. Ridder C. first discover his method in 1979 and inspired by false position method. Let's briefly introduce false position method. Not only does false position method rely on fix point iteration and follow similar procedure as Secant method, but also it always brackets root between two end points. False position method[6] follows following procedure:

$$x_n = x_1 - \frac{f(x_1)(x_{n-1} - x_1)}{f(x_{n-1}) - f(x_1)} \tag{1}$$

Instead of directly apply false position method directly on the function f(x) itself, Ridder's method seek to find a new function

$$h(x) = f(x)e^{\alpha x} \tag{2}$$

such that $h(x_k) = \frac{h(x_i)+h(x_j)}{2}$ and $x_k = \frac{x_i+x_j}{2}$ for the interval $[x_i, x_j]$. Note that difference between false position method and secant method is that false position method insist on initial guess x_1, which guaranteed by intermediate value theorem that root is bracketed between two end points. This method is surprisingly concise comparing to Muller's method, but have similar performance in solving real roots.

Brent's method is first discovered by Brent R.P. in 1973, which is an improvement of Dekker's method. In each step, Dekker's method decide which one of bisection method and secant method should be used to apprximate next guess point. Suppose we are given f(x), and root finding procedure is written as following[1]:

$$f(x) = \begin{cases} b_k - \frac{f(b_k)(b_k - b_{k-1})}{f(b_k) - f(b_{k-1})}, & \text{if } f(b_k) \neq \text{f}(\text{b}_{k-1}) \\ \frac{a_k+b_k}{2}, & \text{otherwise} \end{cases} \tag{3}$$

where b_k and b_{k-1} are guesses at k and k-1 iterations, a_k and b_k are two end points at k iterations. Instead of making decision between two methods, Brent's method

make choices between three method, bisection method, secant method, and inverse quadratic interpolation. Z. Zhang [7] has simplified Brent's method, then Steven A. Stage [4] corrected it. The corrected version of Brent's method will be implemented in Matlab for testing and results. This paper examine order of convergence with two examples which represents two circumstances in the section *Testing and Results*.

2 Description of the Method

Completeness Axiom - Any nonempty subset in R that is bounded above has at least one upper bound.

Definition 1 (Definition of Continuous Function)**.** *Suppose $f : D \to R$, for every $\varepsilon > 0$, there exists a $\delta > 0$ such that for all $x \in D$: if x satisfies $|x - x_0| < \delta$, f(x) satisfy $|f(x) - f(x_0)| < \varepsilon$.*

Theorem 2 (Bolzano's Theorem)**.** *Suppose f(x) is continuous on the closed interval [a,b], and suppose that f(a)f(b)¡0. Then there exists a number c in the interval [a,b],for which f(c)=0.[5]*

Proof. Suppose p = $\{x \in [a, b] | f(x) < 0\}$. Since $a < 0$, then the set p is nonempty. Since $b > 0$ and set p is bounded above by b, then by **the Completeness Axioms**, p has at least a upper bound. Let's assume α is a least upper bound. We aim to show that neither $\alpha < 0$ nor $\alpha > 0$.

Case 1 Assume $\alpha < 0$.

By definition of continuous function, for every $\varepsilon > 0$, there exists a $\delta > 0$ such that for all $x \in D$: if $|x - \alpha| < \delta$, $|f(x) - f(\alpha)| < \varepsilon$. That means there exist a δ such that $f(\alpha + \delta) < 0$. Hence, $\alpha + \delta \in p$. However, since $\alpha + \delta > \alpha$, it contradicts that α is an upper bound.

Case 2 Assume $\alpha > 0$.

By definition of continuous function, for every $\varepsilon > 0$, there exists a $\delta > 0$ such that for all $x \in D$: if $|x - \alpha| < \delta$, $|f(x) - f(\alpha)| < \varepsilon$. That means there exist a δ such that $f(\alpha - \delta) > 0$. Note $\alpha - \delta$ is an upper bound. However, since $\alpha - \delta < \alpha$, it contradicts that α is an least upper bound.

In both cases, there is a contradiciton. Hence α must satisfies that $\alpha = 0$. Since $\alpha \in p$, then $\alpha \in [a, b]$, $f(\alpha) = 0$.

□

2.1 Ridder's Method

Suppose we are given that a continuous f(x) has only 1 root on the interval [a,b] such that $f(a)f(b) < 0$ and $a < b$. We want to apply false position method to a new

function. The idea behind false position method is to connect two end point with a straight line, and by Bolzano's Theorem [5], the line is guaranteed to intersect with x axis. Let's call intersection x_i. We are going to use $f(x_i)$ and a(one of the end point) to approximate the next intersection with x axis,as shown in Figure 1, then repeat this process until it reach halting condition.

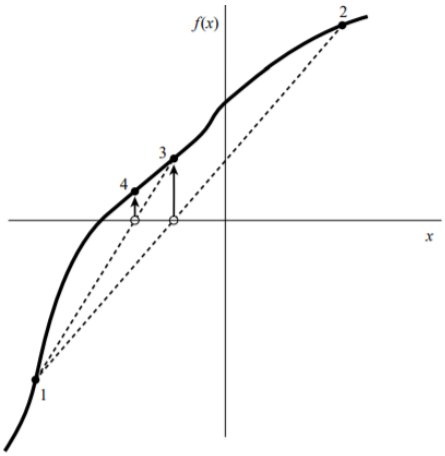

Figure 1: False Position Method

First, we want to construct a exponential function such that the function value at midpoint is exactly the average of f(a) and f(b), which can be written as following

$$h(x) = f(x)e^{\alpha x} \tag{4}$$

such that

$$h(\frac{a+b}{2}) = \frac{h(a)+h(b)}{2} \tag{5}$$

Then, we combine (4) and (5) and get an expression for α

$$\alpha = \frac{\ln\left(f(\frac{a+b}{2}) - \sqrt{f(\frac{a+b}{2})^2 - f(a)f(b)}\right) - \ln f(b)}{b-a}, \mathbf{if} f(a) > 0 \tag{6}$$

Similarly,

$$\alpha = \frac{\ln\left(f(\frac{a+b}{2}) + \sqrt{f(\frac{a+b}{2})^2 - f(a)f(b)}\right) - \ln f(b)}{b-a}, \mathbf{if} f(a) < 0 \tag{7}$$

For simplicity, let c = $\frac{a+b}{2}$ Then, we apply false position method from (1) on this new function h(x) on three initial points (a,$f(a)e^{\alpha a}$) , (c,$f(c)e^{\alpha c}$), (b,$f(b)e^{\alpha b}$), which is written as following

$$d = c + (c-a)\frac{f(c)}{\sqrt{f(c)^2 - f(a)f(b)}}\mathbf{sign}[f(a)] \tag{8}$$

where d is approximation of root at that step and an end point for next step. Surprisingly, all the terms $e^{\alpha x}$ canceled eventually, so the step(8) alone will be used to approximate root. Just like any other root finding method, halting condition is either meeting the maximum number of iteration or the interval [x_0,x_1] less than a given tolerance.

2.2 Brent's Method

The **original Brent's method**[1] discovered by Brent R.P. is really complicated with respect to decision making. Previous two approximation of roots are required to be stored for each iteration. If the function is well-behaved within the interval, method only choose between inverse quadratic interpolation and secant method. If the function is not well-behaved, an additional step is to check five conditions in which we disregard the approximation from previous two methods and use bisection method as the approximation. These five steps are not easily understood in terms of why approximation by bisection method is better.

Improved Brent's method[4] merge bisection method and secant method together where bisection method determines which side the root falls in and operate secant method on that half interval. Improved Brent's method avoid making complicated decisions on bisection method, while it makes sure approximation of root is always bracketed between two end point. Here is what the decision looks like. Suppose f(x) is a continuous function that has a root between a and b. let $c = \frac{a+b}{2}$, and d_i be an approximation at step i. The procedure is indicated as following.

If $f(a) \neq f(c)$ and $f(c) \neq f(b)$ **choose inverse quadratic method**

$$d_i = \frac{af(b)f(c)}{(f(a)-f(b))(f(a)-f(c))} + \frac{bf(a)f(c)}{(f(b)-f(a))(f(b)-f(c))} + \frac{cf(b)f(a)}{(f(c)-f(a))(f(c)-f(b))} \tag{9}$$

then we check if the approximation d_i is in [a,b]. If not, we disregard the approximation and approximate with **bisection method** and **secant method**, which is shown in following.

$$d_i = \begin{cases} c - \frac{f(c)(c-a)}{f(c)-f(a)}, & \text{if } f(c)f(a) < 0 \\ b - \frac{f(b)(b-c)}{f(b)-f(c)}, & \text{if } f(c)f(b) < 0 \end{cases} \tag{10}$$

If it happens that $f(a) = f(b)$ or $f(b) = f(c)$, we simply use formula above in (10). Now, we try to make sure that four points a, c, d_i, b always satisfies $a \leqslant c \leqslant d_i \leqslant b$. However, above procedure guarantee d_i is between [a,b], but not d_i less than s.

Hence we swap d_i and c if $d_i < c$. Among three intervals [a,c] , [c,d_i],and [d_i,b], we choose the one that the product of function values at end points is less than 0.

Both secant method and inverse quadratic interpolation are proved to be superlinear convergence. The precise order of convergence of **secant method** is the positive real solution of $(\varphi)^2 - (\varphi) - 1 = 0$ [2], whereas the precise order of convergence of **inverse quadratic interpolation** is the positive solution of $(\varphi)^3 - (\varphi)^2 - (\varphi) - 1 = 0$. Therefore, the best case happens when Brent's method continuously choose inverse quadratic interpolation, which is close to quadratic convergence. The worst case occurs when Brent's method continuously choose bisection method and secant method. In this case, the order of convergence should be between 1 and $\sqrt{2}$.

3 Code Implementation

Ridder method first operate formula from (6) or (7) to find h(x), then it apply formula from (8) to approximate root; Brent's method will use formula from (9) and (10) to find the root. We set tolerance to be 10^{-10} and maximum number of iteration to be 15. The halting conditions are either the length of interval less than tolerance or reaching maximum number of iterations.

- `secant.m`: Outputs an approximation of root using one round of secant method.
- `inversequadratic.m`: Outputs an approximation of root using one round of inverse quadratic interpolation.
- `brent.m`: Outputs an approximation of root, number of iterations, and error estimation with respect to true root(approximation from Matlab built-in function fzero).
- `ridder.m`: Outputs an approximation of root, number of iterations, and error estimation with respect to true root(approximation from Matlab built-in funciton fzero).

The file `main.m` executes ridder.m and brent.m to display examples outlined in next section.

4 Testing and Results

We will perform tests for Brent's method and Ridder's method. We want to see which one converges faster in terms of order of convergence and numbers of interations.

4.1 Example 1: `f(x) = sin(1/x)`

The function $sin(\frac{1}{x})$ oscillates quickly as x goes to 0. We choose a interval between 0.2 and 0.4 with the properties that $f(0.2) < 0$ and $f(0.4) > 0$. Hence, the function is guaranteed to have at least one root by Bolzano's theorem. We set tolerance to be 10^{-10} and maximum number of iteration to be 15. Error estimation is calucated by comparing to estimation by Matlab fzero function. Since halting condition is $|x_0 - x_1| < 10^{-10}$ where x0 and x1 are endpoints, then the approximation might actually reach better accuracy than the tolerance. Here are two tables of results.

Iterations	Ridder's error	Brent's error
1	0.00608573	0.00609183
2	8.98573e-06	1.72234e-05
3	2.04001e-09	8.22551e-09
4	1.113e-13	9.87266e-13
5	1.11022e-16	1.11022e-16
6	1.66533e-16	1.11022e-16
7	NA	1.66533e-16

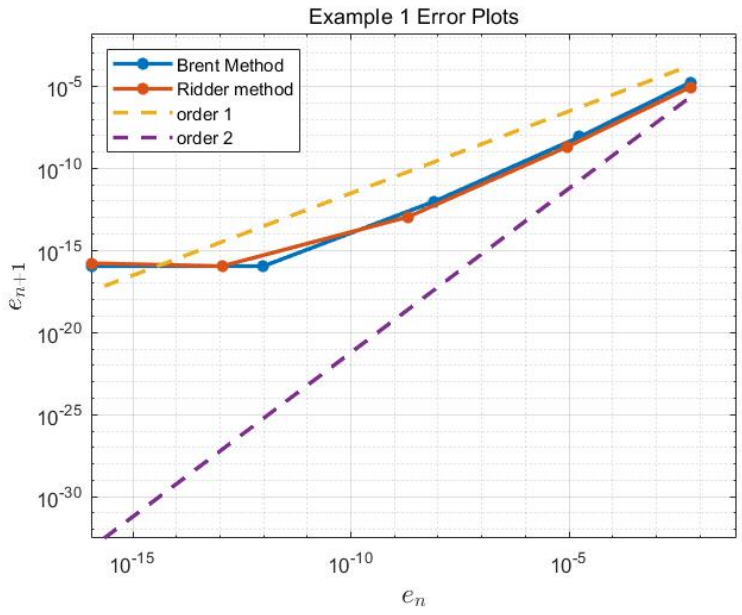

Figure 2: Error Plot

As you can see in Figure 1, Brent's error generally line up with Ridder's error. Both methods reach 12 decimal places in 4 iterations. Estimated order of convergence for Ridder's method is 0.703961; estimated order of convergence for Brent's method is 1.00722. This does not mean that Brent's converge faster. Ridder's

estimated order of convergence is smaller because estimated root in fifth iteration is really Small and Matlab causes a round off error. For the same reason, the last term in vector of order of convergence is Inf, so there are only 5 points plotted for Brent's method.

4.2 Example 2: `f(x) = 1/sin(x) - x`

We set interval $[x_0,x_1]$to be [1/2, 2], and $f(1/2)f(2) < 0$, which is a requisite for both Brent's method and Ridder's method. Error estimation is compared to approximation by Matlab fzero function. We set tolerance to be 10^{-10} and maximum number of iteration to be 15. As staetd previously, since halting condition is $|x_0 - x_1| < 10^{-10}$ where x0 and x1 are endpoints, then the approximation might actually reach better accuracy than the tolerance. The following two graph are the results.

Iterations	Ridder's error	Brent's error
1	0.0142904	0.027422
2	0.00176697	4.03616e-05
3	3.68665e-08	1.14787e-08
4	1.39888e-14	8.06466e-13
5	0	0

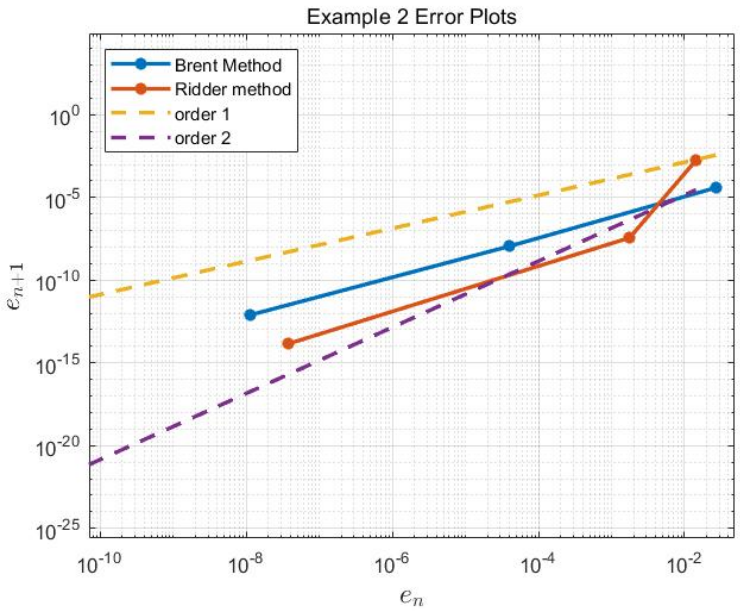

Figure 3: Error Plot

In 5th iterations, both method reach a great accuracy such that the error is so close to fzero approximation that Matlab is only able to output a zero instead. In

this example, it seems that Brent's method converge more consistently comparing to Ridder's method. Ridder's method only improve 1 decimal place between first and second iterations, but third iteration is able to reach 7 decimal place of accuracy. **However**, with respect to order of convergence, Ridder's method does a better job. The estimated order of convergence for Brent's method is 1.17124; the estimated order of convergence for Ridder's method is 1.3718. In addition, Ridder's method reach a more decimal place than Brent's method at 4th step.

4.3 Discussion

In two examples that we show above, **Improved Brent's method** and **Ridder's method** both converge superlinearly, which means that order of convergence is between 1 and 2. However, when dealing with highly oscillating functions(example 1), both method barely converge superlinearly, whose order of convergence is practically 1. Improved Brent's inverse quadratic interpolation converge linearly because oscillating functions require interpolation of higher degree of polynomials and quadratic interpolation causes a large error. Ridder's false position method seem to work better on a interval where $f'(x)$ is small. In **example 1**, function oscillate really quick around 0, and some slope approach infinity during oscillation, which might cause linear convergence of Ridder's method. **Example 2** shows that in some situations, Ridder's method is not as consistent as Brent's method. This is because Brent's method has secant method and bisection method as a backup plan when inverse quadratic interpolation fails. In general, both method performs better than linear convergence when tackling various functions. Improved Brent's method tend to be consistent whereas Ridder's method tend to converge a little faster at cost of inconsistency. Neither of two methods perform sufficiently better than the other.

References

[1] Brent's method. (2020). Retrieved November 22, 2020, from https://en.wikipedia.org/wiki/Brent%27s_method.

[2] Pedro Díez, A note on the convergence of the secant method for simple and multiple roots, *Applied Mathematics Letters*, Volume 16, Issue 8, 2003, Pages 1211-1215, ISSN 0893-9659 https://doi.org/10.1016/S0893-9659(03)90119-4.

[3] Ridder's method. (2020). Retrieved November 22, 2020, from https://en.wikipedia.org/wiki/Ridders%27_method.

[4] Steven A. Stage, Comments on An Improvement to the Brent's Method, Volume (4) of International Journal of Exerimental Algorithms, 2013, http://www.cscjournals.org/manuscript/Journals/IJEA/Volume4/Issue1/IJEA-33.pdf .

[5] The Intermediate Value Theorem , Milefoot Mathematics, from http://www.milefoot.com/math/calculus/limits/IntValueTheorem13.htm.

[6] Weisstein, Eric W. "Method of False Position." From MathWorld–A Wolfram Web Resource. https://mathworld.wolfram.com/MethodofFalsePosition.html

[7] Z. Zhang, An Improvement to the Brent's Method, IJEA, vol. 2, pp. 21-26, May 31, 2011 http://www.cscjournals.org/manuscript/Journals/IJEA/Volume2/Issue1/IJEA-7.pdf.