AM225: One-dimensional root finding methods

In these notes we will consider several simple methods for finding a root of continuous scalar function $f : \mathbb{R} \to \mathbb{R}$. These notes will culminate in Ridders' method, an efficient and reliable root finding technique. In AM225, we will use these examples to compare the speed of implementations in different programming languages.

The three methods are based on when a root is initially bracketed into an interval $[a, b]$ where we assume without loss of generality that $f(a) < 0$ and $f(b) > 0$. Then by the intermediate value theorem, there exists a root of f in the interval (a, b). The three methods are tested on the function

$$f(x) = \lambda - \cos x \tag{1}$$

where $\lambda \in (-1, 1)$. The analytical root of this function is therefore $\cos^{-1} \lambda$. The root is initally bracketed between $a = 0$ and $b = \pi$.

Bisection method

The first method we will consider is the bisection method, which uses the following procedure. First define $c = (a + b)/2$ and find $f(c)$. Then there are three cases:

1. If $f(c) > 0$ then the root is within (a, c). Redefine b to c, and repeat.
2. If $f(c) < 0$ then the root is within (c, b). Redefine a to c, and repeat.
3. If $f(c) = 0$ then the root is exactly determined.

For each iteration of the above procedure, the interval containing the root is divided in half. After N iterations, the root is known to an accuracy of $(b - a)2^{-N}$. Thus reaching an accuracy of ϵ requires $\log_2 \frac{b-a}{\epsilon}$ iterations. The bisection method decreases the error by a constant factor by each iteration, which is referred to as *linear convergence*. The program `bisection.py` demonstrates the method on the test function.

The false position method

The bisection method is highly reliable, but it does not make use of any information about the actual function values—it only uses their sign. Rather than bisect based on the midpoint of the current interval, the false position method instead bisects based on the root of the linear interpolant between $(a, f(a))$ and $(b, f(b))$. Specifically, this is

$$c = a + \frac{(b - a)f(a)}{f(a) - f(b)}. \tag{2}$$

Depending on the sign of $f(c)$, the bracket is reduced to (a, c) or (c, b) as in the bisection method. While the false position method often makes better choices than bisection, there

are some cases where it performs worse. For example, consider the function

$$g(x) = e^{50x} - \frac{1}{50} \tag{3}$$

with an initial bracket of $(-1, 0)$. The true root of this function is at $\frac{\log 0.02}{50} \approx -0.07$. The function is negative and flat for most of the interval, before increasing rapidly close to $x = 0$. The assumption of approximate linearity behind Eq. 2 does not hold, and the chosen values of c approach the true root very slowly.

The convergence of the false position method is not generally known, and depends on the function considered, but is frequently far better than bisection.

Ridders' method

Ridders' method is an improvement on the false position method that turns it into an effective and reliable technique. The first step is to again consider the midpoint $c = (a + b)/2$ and find $f(c)$. Next, the unique exponential function is factored out of f in order to make the three points $(a, f(a))$, $(c, f(c))$, and $(b, f(b))$ colinear. This corresponds to finding an e^Q such that

$$f(a) - 2f(c)e^Q + f(b)e^{2Q} = 0. \tag{4}$$

Note that Eq. 4 is equivalent to enforcing that the centered-difference stencil for the second derivative on $f(a), f(c)e^Q, f(b)e^{2Q}$ is zero. This is equation is a quadratic in e^Q, and has the solution

$$e^Q = \frac{f(c) + \text{sign}[f(b)]\sqrt{f(c)^2 - f(a)f(b)}}{f(b)}. \tag{5}$$

Since $f(a)$ and $f(b)$ have opposite sign, the square root will always be real. Once this is determined, a false position step is performed on the function with the exponential factored out. Performing linear interpolation yields the false position step of

$$d = c + (c - a)\frac{\text{sign}[f(a) - f(b)]f(c)}{\sqrt{f(c)^2 - f(a)f(b)}}. \tag{6}$$

Based on the sign of $f(d)$ the bracketed region is replaced with either (a, d) if $f(d) > 0$, or (d, b) if $f(d) < 0$. However, a further optimization is possible when the sign of $f(c)$ differs from $f(d)$. In that case, the bracketed region can be reduced further to (c, d) or (d, c) depending on the ordering of c and d.

Ridders' method gives *quadratic convergence* at each iteration, whereby the error is squared at each step. This is similar to Newton's method. However, note that two function evaluations are required per step (to evaluate $f(c)$ and $f(d)$) and thus to compare to the other methods it is fairer to view it as order $\sqrt{2}$ convergence instead of order 2.

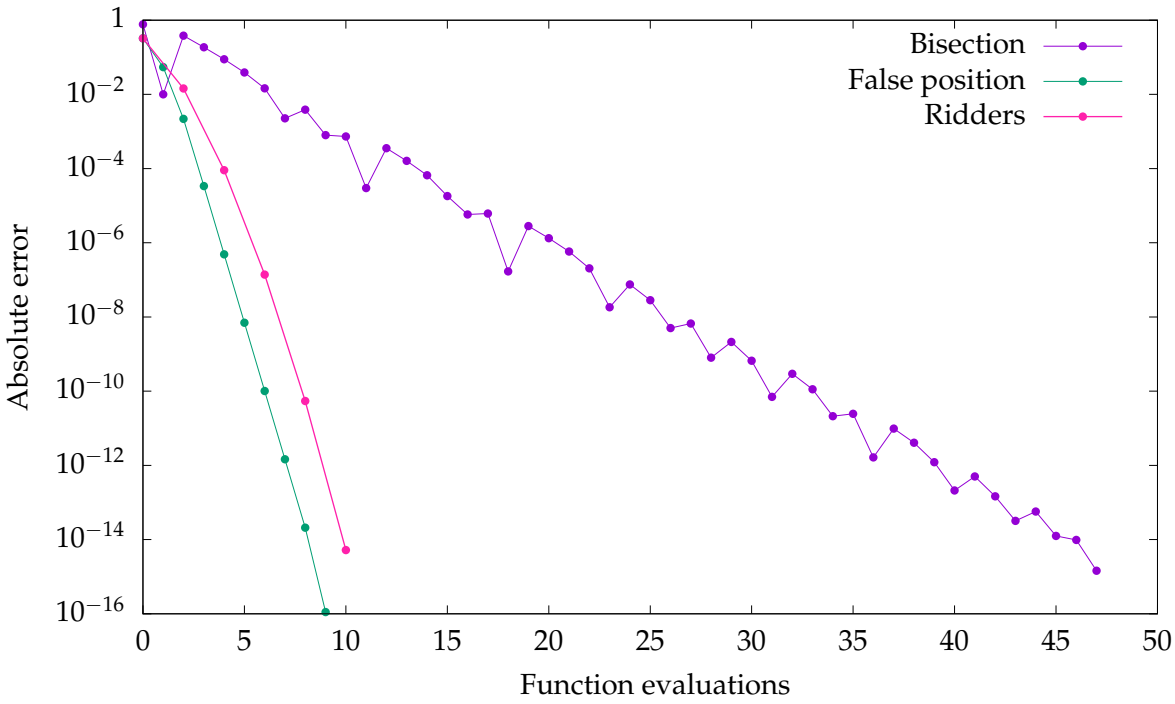

Figure 1: Convergence rates for the three different methods considered.

Numerical comparison

Figure 1 shows the convergence of the three different methods on the test problem with $\lambda = 0.7$. The graph confirms the convergence rates discussed in the previous sections. The bisection method exhibits slow, linear convergence, whereas the false position method and Ridders' method converge more rapidly. For this particular test the false position method does slightly better in terms of accuracy versus the number of function evaluations. However Ridders' method is more generally applicable and reliable.