Lecture Notes for Math 5630/6630 **Fall 2024**

Note 4: Bracket Methods

Tags: math.na *Date: 09/05/2024*

Disclaimer: *This lecture note is for math 5630/6630 class only.*

1 Bracket Methods

We start with a game that relates to our topic.

Alice selects a number in $\{1, 2, \cdots, 100\}$, and Bob guesses the number. Each time Alice will tell Bob if his guess is greater/smaller than the answer. Can you find a good strategy for Bob?

A naive strategy is to enumerate the numbers from the set $\{1, 2, \cdots, 100\}$, at the worst case, Bob needs 100 rounds to get the answer. It is inefficient because the information (greater or smaller) is not used.

A better strategy is to use the so-called “binary search”. For instance, if Alice chooses 42 and Bob starts with the middle number 50, he will get a response “greater” in the 1st round, eliminating about half of the numbers, leaving the pool with only $\{1, 2, \cdots, 49\}$. Then Bob chooses the middle number 25 of the remaining pool, and gets a response “smaller” in the 2nd round, eliminating 25 numbers, leaving the pool with only $\{26, 27, \cdots 49\}$, and so on.

We might notice this binary search strategy will always eliminate half of the remaining numbers if the guess is wrong.

1.1 Bisection Method

If we know $f(a) < 0$ and $f(b) > 0$, then by the intermediate value theorem, f has at least one root in $[a, b]$. To locate a root, we can first select $x_1 = \frac{a+b}{2}$, the sign of $f(x_1)$ is either the same with $f(a)$ or with $f(b)$.

1. If $f(x_1) > 0$, then there is a root in $[a, x_1]$.
2. If $f(x_1) < 0$, then there is a root in $[x_1, b]$.
3. If $f(x_1) = 0$, we find a root.

In any of the cases, we can eliminate half of the interval and on the remaining part, the signs of f at the endpoints are different. Therefore, we can **recursively** perform this algorithm until we are satisfied.

Theorem 1.1. *The nth iteration of the bisection method satisfies*

$$|x_n - x^*| \leq \frac{b-a}{2^n}.$$

This is because we can eliminate half of the interval in each iteration, and then the remaining interval's size at nth iteration is $2^{-n}(b-a)$.

Example 1.2. *If we use the bisection method to find the root $x^* = \sqrt{2} \approx 1.41421356$ of $f(x) = x^2 - 2$ on $[1,2]$, then $|x_n - \sqrt{2}| \leq 2^{-n}$, which means we can only get about 3 or 4 significant digits after 9 iterations.*

$$1.5,\ 1.25,\ 1.375,\ 1.4375,\ 1.40625,\ 1.421875,\ 1.4140625,\ 1.41796875,\ 1.416015625$$

Note: Under a floating point system (suppose evaluations of f are accurate). The iterations will not update after the machine precision is reached.

Remark 1.3. *The core idea of the bracket method is to maintain an interval whose size is getting smaller, while the evaluations at the endpoints differ in the sign. Such an interval is called a "bracket".*

1.2 Stopping Criteria

We can repeatedly perform the bisection method, but when should we stop? The most common stopping criteria are

a. $|x_{n+1} - x_n| \leq \varepsilon_{atol}$

b. $|x_{n+1} - x_n| \leq \varepsilon_{rtol}|x_n|$

c. $|f(x_n)| < \varepsilon_{ftol}$

The numbers ε_{atol}, ε_{rtol}, ε_{ftol} are called *absolute*, *relative*, *function* tolerances. Under different stopping criteria, the required number of iterations may be different.

When $|x_n|$ is tiny, the second criterion (relative error) is less robust than the first (why?). A popular combination of the first two criteria can be written as

$$|x_n - x_{n-1}| \leq \varepsilon_{tol}(1 + |x_n|).$$

Corollary 1.4. *Using the stopping criterion (a), the needed iteration number is at most*

$$n = \lceil \log_2 \frac{3(b-a)}{\varepsilon_{atol}} \rceil.$$

Because $|x_n - x_*|\le \frac{b-a}{2^n}$,

$$|x_n - x_{n-1}|\le |x_n - x^*|+|x_{n-1} - x_*|\le \frac{b-a}{2^n} + \frac{b-a}{2^{n-1}} = \frac{3(b-a)}{2^n}. \tag{0.1}$$

1.3 False Position Method

The bisection method only uses $\operatorname{sgn}(f(a))$ and $\operatorname{sgn}(f(b))$ instead of the function values. The *false position method* (*Regula falsi* in Latin) improves the bisection method by taking the function values into account. Instead of selecting the midpoint $c = \frac{a+b}{2}$, the false position method selects the point $c \in [a, b]$ that lies on the line connecting $(a, f(a))$ and $(b, f(b))$, that is

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

The false position method is also guaranteed to converge to a root if $f(a)f(b) < 0$ and the implementation is quite similar to the bisection method. It usually converges faster than the bisection method, but sometimes exceptions occur.

When f'' keeps the same sign over $[a, b]$, it is easy to show that only one side of the bracket is updating. The bracket size will never decrease to zero, which differs from the bisection method. In the following, we use an example to illustrate this, see Figure 4.1.

The function $f(x) = x^2 - 1$ over the initial bracket $[0, 3]$. The left endpoint keeps updating to the root while the right endpoint is fixed at 3.

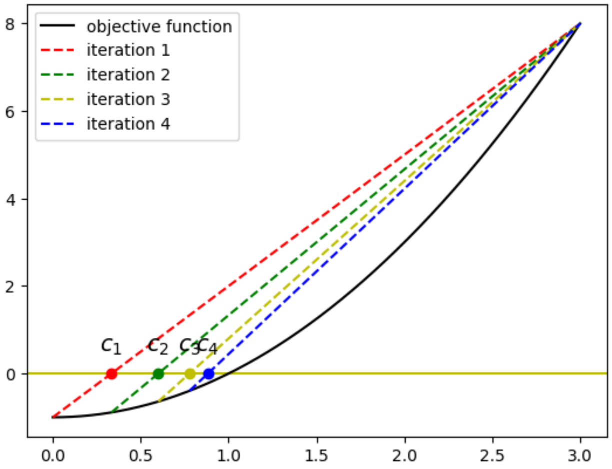

Figure 4.1: False Position Method

2 Order of Convergence

The order of convergence quantifies how fast the sequence approximates the limiting value.

Definition 2.1. *The order of convergence of a sequence $\{x_n\}$ is $p > 0$ if*

$$\lim_{n\to\infty} \frac{|x_{n+1} - x^*|}{|x_n - x^*|^p} = \rho > 0.$$

The constant ρ is the rate of convergence. If $p = 1$ and $\rho < 1$ the sequence is said to have linear convergence. If $p = 2$, the sequence is said to have quadratic convergence.

If the limit does **not** exist while the upper bound exists for sufficiently large n:

$$\frac{|x_{n+1} - x^*|}{|x_n - x^*|^p} \le \rho,$$

then the order of convergence is *at least* p and the rate of convergence is *at most* ρ. Most textbooks do not distinguish this from Definition 2.1.

Example 2.2. *In the bisection method, it is known that*

$$|x_n - x^*| \le \frac{b-a}{2^n},$$

which means the error decays at least exponentially fast. However, it does not imply linear convergence. For instance, if we take

$$x^* = (0.101001000100001\cdots)_2$$

where we insert k zeros between the k-th nonzero digit and the $(k+1)$-th nonzero digit. Using the bisection method to approach this number, we will find

$$\limsup_{n\to\infty} \frac{|x_{n+1} - x^*|}{|x_n - x^*|} = \infty.$$

In practice, the limit or even upper bound may not exist for $\frac{|x_{n+1}-x^*|}{|x_n-x^*|^p}$, it is possible to consider the convergence rate in a weaker sense. For instance, suppose for any sufficiently large n that the inequality

$$\lim_{k\to\infty} \sqrt[k]{\frac{|x_{n+k} - x^*|}{|x_n - x^*|^{p^k}}} = \rho$$

holds for certain $p > 0$ and $\rho > 0$, then the sequence has a *mean* order of convergence is p and a *mean* convergence rate ρ. It means that, on average, each iteration contributes an order of p and a rate of ρ.

Theorem 2.3. *The* ***mean*** *convergence order of the bisection method is (at least)* 1.

Without loss of generality, we may assume the initial bracket is on $[0, 1]$. Let the root $x^* = 0.b_1 b_2 \cdots$ be the binary representation, then the sequence of bisection method can be written as

$$x_n = 0.b_1 b_2 \cdots b_{n-1} 1,$$

where b_i is the i-th bit of the binary representation. The error at the n-th iteration is

$$|x_n - x^*| = |2^{-n} - \sum_{j \geq n} 2^{-j} b_j| = \begin{cases} \sum_{j>n} 2^{-j} b_j & \text{if } b_n = 1 \\ \sum_{j>n} 2^{-j}(1 - b_j) & \text{if } b_n = 0 \end{cases}$$

For each n such that $b_n = 1$ (similar argument holds for $b_n = 0$), we can find $s \in \mathbb{N}$ such that $b_{s+n} = 1$, otherwise we arrive at the exact solution.

$$\frac{|x_{n+k} - x^*|}{|x_n - x^*|} = \frac{\sum_{j>n+k} 2^{-j} b_j}{\sum_{j>n} 2^{-j} b_j} \leq \frac{2^{-(n+k)}}{2^{-(n+s)}} = 2^{s-k}$$

Therefore, the geometric mean of the convergence rate is bounded by $\frac{1}{2}$.

$$\rho = \lim_{k \to \infty} \sqrt[k]{\frac{|x_{n+k} - x^*|}{|x_n - x^*|}} \leq \lim_{k \to \infty} 2^{s/k} \frac{1}{2} = \frac{1}{2}.$$

A common technique to study the order of convergence is to use **asymptotic analysis** with the Taylor expansion. Let us use the false position method as an example.

Theorem 2.4. *The false position method converges linearly.*

Assume $f(x) \in C^2[a, b]$ and x^* be a simple root and f'' has a fixed sign on $[a, b]$. It is known that one of the bracket's endpoints is fixed (e.g. b is fixed). Then

$$x_{n+1} = x_n - f(x_n) \frac{b - x_n}{f(b) - f(x_n)}.$$

Since $f(x_n) = f(x^*) + f'(x^*)(x_n - x^*) + \frac{f''(\zeta)}{2}(x_n - x_*)^2$ and $f(x^*) = 0$, we get

$$\begin{aligned} x_{n+1} - x^* &= x_n - x^* - (x_n - x_*) \left(f'(x^*) + \frac{f''(\zeta)}{2}(x_n - x^*) \right) \frac{b - x_n}{f(b) - f(x_n)} \\ &= (x_n - x^*) \left[1 - \left(f'(x^*) + \frac{f''(\zeta)}{2}(x_n - x^*) \right) \frac{b - x_n}{f(b) - f(x_n)} \right] \\ &\approx (x_n - x^*) \left[1 - f'(x^*) \frac{b - x^*}{f(b)} \right]. \end{aligned}$$