Jim Lambers
MAT 460/560
Fall Semester 2009-10
Lecture 8 Notes

These notes correspond to Section 2.1 in the text.

Nonlinear Equations

To this point, we have only considered the solution of linear equations. We now explore the much more difficult problem of solving nonlinear equations of the form

$$f(\mathbf{x}) = \mathbf{0},$$

where $f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^m$ can be any known function. A solution $\mathbf{x}$ of such a nonlinear equation is called a *root* of the equation, as well as a *zero* of the function f.

Existence and Uniqueness

For simplicity, we assume that the function $f : \mathbb{R}^n \to \mathbb{R}^m$ is continuous on the domain under consideration. Then, each equation $f_i(\mathbf{x}) = 0$, $i = 1, \ldots, m$, defines a hypersurface in $\mathbb{R}^m$. The solution of $f(\mathbf{x}) = 0$ is the intersection of these hypersurfaces, if the intersection is not empty. It is not hard to see that there can be a unique solution, infinitely many solutions, or no solution at all.

For a general equation $f(\mathbf{x}) = \mathbf{0}$, it is not possible to characterize the conditions under which a solution exists or is unique. However, in some situations, it is possible to determine existence analytically. For example, in one dimension, the Intermediate Value Theorem implies that if a continuous function $f(x)$ satisfies $f(a) \leq 0$ and $f(b) \geq 0$ where $a < b$, then $f(x) = 0$ for some $x \in (a, b)$.

Similarly, it can be concluded that $f(x) = 0$ for some $x \in (a, b)$ if the function $(x - z)f(x) \geq 0$ for $x = a$ and $x = b$, where $z \in (a, b)$. This condition can be generalized to higher dimensions. If $S \subset \mathbb{R}^n$ is an open, bounded set, and $(\mathbf{x} - \mathbf{z})^T f(\mathbf{x}) \geq 0$ for all $\mathbf{x}$ on the boundary of S and for some $\mathbf{z} \in S$, then $f(\mathbf{x}) = \mathbf{0}$ for some $\mathbf{x} \in S$. Unfortunately, checking this condition can be difficult in practice.

One useful result from calculus that can be used to establish existence and, in some sense, uniqueness of a solution is the *Inverse Function Theorem*, which states that if the Jacobian of f is nonsingular at a point $\mathbf{x}_0$, then f is invertible near $\mathbf{x}_0$ and the equation $f(\mathbf{x}) = \mathbf{y}$ has a unique solution for all $\mathbf{y}$ near $f(\mathbf{x}_0)$.

If the Jacobian of f at a point $\mathbf{x}_0$ is singular, then f is said to be *degenerate* at $\mathbf{x}_0$. Suppose that $\mathbf{x}_0$ is a solution of $f(\mathbf{x}) = \mathbf{0}$. Then, in one dimension, degeneracy means $f'(x_0) = 0$, and we say that x_0 is a *double root* of $f(x)$. Similarly, if $f^{(j)}(x_0) = 0$ for $j = 0, \ldots, m - 1$, then x_0 is a root of multiplicity m. We will see that degeneracy can cause difficulties when trying to solve nonlinear equations.

Sensitivity

Recall that the absolute condition number of a function $f(x)$ is approximated by $|f'(x)|$. In solving a nonlinear equation in one dimension, we are trying to solve an inverse problem, where the forward problem is the evaluation of f at $x = 0$. It follows that the condition number for solving $f(x) = 0$ is approximately $1/|f'(x_0)|$, where x_0 is the solution. This discussion can be generalized to higher dimensions, where the condition number is measured using the norm of the Jacobian.

The Bisection Method

Suppose that $f(x)$ is a continuous function that changes sign on the interval $[a, b]$. Then, by the Intermediate Value Theorem, $f(x) = 0$ for some $x \in [a, b]$. How can we find the solution, knowing that it lies in this interval?

The method of *bisection* attempts to reduce the size of the interval in which a solution is known to exist. Suppose that we evaluate $f(m)$, where $m = (a + b)/2$. If $f(m) = 0$, then we are done. Otherwise, f must change sign on the interval $[a, m]$ or $[m, b]$, since $f(a)$ and $f(b)$ have different signs. Therefore, we can cut the size of our search space in half, and continue this process until the interval of interest is sufficiently small, in which case we must be close to a solution. The following algorithm implements this approach.

Algorithm (Bisection) Let f be a continuous function on the interval $[a, b]$ that changes sign on (a, b). The following algorithm computes an approximation p^* to a number p in (a, b) such that $f(p) = 0$.

```
for $j = 1, 2, \ldots$ do
    $p_j = (a + b)/2$
    if $f(p_j) = 0$ or $b - a$ is sufficiently small then
        $p^* = p_j$
        return $p^*$
    end
    if $f(a)f(p_j) < 0$ then
        $b = p_j$
    else
        $a = p_j$
    end
end
```

At the beginning, it is known that (a, b) contains a solution. During each iteration, this algorithm updates the interval (a, b) by checking whether f changes sign in the first half (a, p_j), or in the second half (p_j, b). Once the correct half is found, the interval (a, b) is set equal to that half. Therefore, at the beginning of *each* iteration, it is known that the current interval (a, b) contains a solution.

The test $f(a)f(p_j) < 0$ is used to determine whether f changes sign in the interval (a, p_j) or (p_j, b). This test is more efficient than checking whether $f(a)$ is positive and $f(p_j)$ is negative, or vice versa, since we do not care which value is positive and which is negative. We only care whether they have different signs, and if they do, then their product must be negative.

In comparison to other methods, including some that we will discuss, bisection tends to converge rather slowly, but it is also guaranteed to converge. These qualities can be seen in the following result concerning the accuracy of bisection.

Theorem *Let f be continuous on $[a, b]$, and assume that $f(a)f(b) < 0$. For each positive integer n, let p_n be the nth iterate that is produced by the bisection algorithm. Then the sequence $\{p_n\}_{n=1}^{\infty}$ converges to a number p in (a, b) such that $f(p) = 0$, and each iterate p_n satisfies*

$$|p_n - p| \leq \frac{b-a}{2^n}.$$

It should be noted that because the nth iterate can lie anywhere within the interval (a, b) that is used during the nth iteration, it is possible that the error bound given by this theorem may be quite conservative.

Example We seek a solution of the equation $f(x) = 0$, where

$$f(x) = x^2 - x - 1.$$

Because $f(1) = -1$ and $f(2) = 1$, and f is continuous, we can use the Intermediate Value Theorem to conclude that $f(x) = 0$ has a solution in the interval $(1, 2)$, since $f(x)$ must assume every value between -1 and 1 in this interval.

We use the method of *bisection* to find a solution. First, we compute the midpoint of the interval, which is $(1 + 2)/2 = 1.5$. Since $f(1.5) = -0.25$, we see that $f(x)$ changes sign between $x = 1.5$ and $x = 2$, so we can apply the Intermediate Value Theorem again to conclude that $f(x) = 0$ has a solution in the interval $(1.5, 2)$.

Continuing this process, we compute the midpoint of the interval $(1.5, 2)$, which is $(1.5+2)/2 = 1.75$. Since $f(1.75) = 0.3125$, we see that $f(x)$ changes sign between $x = 1.5$ and $x = 1.75$, so we conclude that there is a solution in the interval $(1.5, 1.75)$. The following table shows the outcome of several more iterations of this procedure. Each row shows the current interval (a, b) in which we know that a solution exists, as well as the midpoint of the interval, given by $(a + b)/2$, and the value of f at the midpoint. Note that from iteration to iteration, only one of a or b changes, and the endpoint that changes is always set equal to the midpoint.

a	b	$m = (a+b)/2$	$f(m)$
1	2	1.5	-0.25
1.5	2	1.75	0.3125
1.5	1.75	1.625	0.015625
1.5	1.625	1.5625	-0.12109
1.5625	1.625	1.59375	-0.053711
1.59375	1.625	1.609375	-0.019287
1.609375	1.625	1.6171875	-0.0018921
1.6171875	1.625	1.62109325	0.0068512
1.6171875	1.62109325	1.619140625	0.0024757
1.6171875	1.619140625	1.6181640625	0.00029087

The correct solution, to ten decimal places, is 1.6180339887, which is the number known as the *golden ratio.* □