

Nonlinear Equations and Optimization I

April 4, 2006

Contents

1	Theory of Nonlinear Equations	2
1.1	Nonlinear Equations	2
1.2	Attainable Accuracy	3
1.3	Order of Convergence	4
2	Fixed Point Iteration	5
2.1	Theory	6
2.2	Application	8

1 Theory of Nonlinear Equations

1.1 Nonlinear Equations

The problem we will consider in this lecture is solving a nonlinear equation $f(x) = 0$ for x .

For our example we will use Kepler's equation

$$M = E - e \sin(E)$$

which occurs in celestial mechanics. Here M is the mean anomaly, E is the eccentric anomaly, and e , $0 < e < 1$, is the eccentricity of an elliptic orbit. We will take $M = 1$ and $e = 0.5$ and solve the equation for E . Thus the equation we want to solve is

$$f(E) = E - 1 - 0.5 \sin(E) = 0.$$

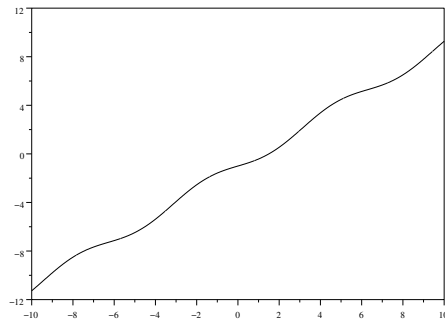
The first step in solving any equation numerically is to graph the equation to obtain an idea of the number and approximate location of the roots.

```
-->function y = kepler (e)
--> y = e - 1 - 0.5*sin(e)
-->endfunction

-->x = -10:0.01:10;

-->y = kepler(x);

-->plot2d(x, y)
```



The graph shows that there is one root near $x = 1.5$.

1.2 Attainable Accuracy

Irrespective of the algorithm being used, there is a limit to the attainable accuracy in solving a nonlinear equation due to the errors in function evaluation caused by rounding.

Example

The following is an example of the small but often unpredictable rounding errors which occur in the numerical evaluation of a function. Let

$$f(x) = (x - 1)^6 = x^6 - 6x^5 + 15x^4 - 20x^3 + 15x^2 - 6x + 1.$$

We will evaluate $f(x)$ near the root $x = 1$:

```
-->x = 0.995 : 0.0001: 1.005;
```

```
-->y1 = (x - 1).^6;
```

```
-->y2 = x.^6 - 6*x.^5 + 15*x.^4 - 20*x.^3 + 15*x.^2 - 6*x + 1;
```

```
-->plot2d(x, y1)
```

```
-->plot2d(x, y2)
```

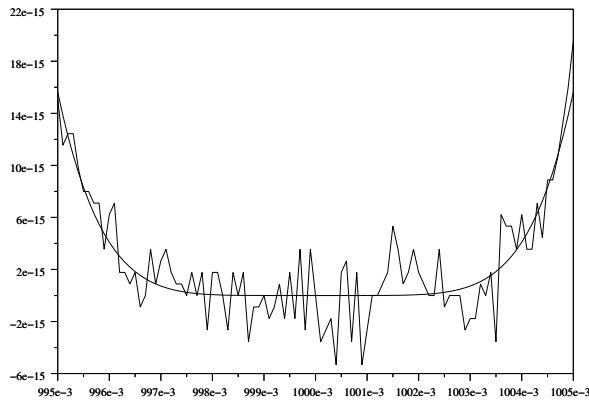


Figure 1: Rounding Error in Function Evaluation

When evaluated in the expanded form, it shows small (of the order of $\varepsilon_{\text{mach}}$) seemingly random fluctuations which will clearly make it difficult to determine the root of $f(x)$.

To analyse this effect quantitatively the most important thing to note is that the equation we are trying to solve will be, when viewed computationally, affected by rounding error. Let x be the exact solution of $f(x) = 0$, let $\hat{f}(x)$ the

function actually computed, and let \hat{x} the computed solution. We will assume that the computed solution is an exact solution of the computed function $\hat{f}(x)$, $\hat{f}(\hat{x}) = 0$, that is we assuming that the only error is that arising from function evaluation.

Let $\hat{x} = x + \Delta x$ and $\hat{f}(x) = f(x) + \Delta f(x)$. Then, using the first order Taylor series approximation, we have

$$\begin{aligned} 0 = \hat{f}(\hat{x}) &= f(\hat{x}) + \Delta f(\hat{x}) \\ &= f(x + \Delta x) + \Delta f(\hat{x}) \\ &\approx f(x) + f'(x)\Delta x + \Delta f(\hat{x}) \\ &= f'(x)\Delta x + \Delta f(\hat{x}) \end{aligned}$$

From which we get

$$\Delta x \approx -\frac{\Delta f(\hat{x})}{f'(x)} \quad (1)$$

$\Delta f(\hat{x})$, which is the difference between $f(x)$ and the function actually computed, can be interpreted as the rounding error in evaluating $f(x)$. This difference will typically be of order $\varepsilon_{\text{mach}}$.

What formula (1) shows is that the accuracy which we can determine the solution of $f(x) = 0$ depends on the slope of the function at the root. When the graph of $f(x)$ is flat, i.e the derivative $f'(x)$ is small near the root, it will be difficult to estimate the root accurately no matter what method is used.

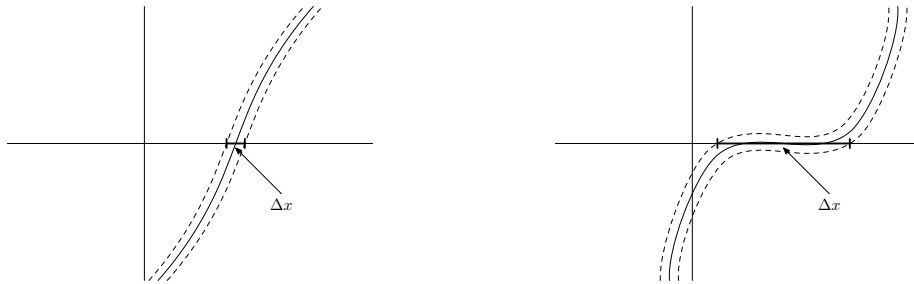


Figure 2: Errors in the Solution of Nonlinear Equations

1.3 Order of Convergence

Algorithms for solving nonlinear equations, like many other numerical algorithms, compute a sequence of approximations x_0, x_1, \dots to the solution, the sequence terminating when we have judged that we have a sufficiently accurate approximation. Let x_1, x_2, \dots be a sequence of approximations to the some number x . In the present context we can think of a series of approximations to the solution of a nonlinear equation.

The error in the k -the approximation is

$$e_k = x_k - x.$$

If the approximations x_k converge to x then $\lim_{k \rightarrow \infty} e_k = 0$. What we want is a measure of how fast e_k converges to 0.

The sequence x_k is said to converge to x with **order of convergence** r if

$$\lim_{k \rightarrow \infty} \frac{\|e_{k+1}\|}{\|e_k\|^r} = C$$

for some constant $C > 0$. This says that (at least for large k)

$$\|e_{k+1}\| \approx C\|e_k\|^r,$$

that is, the error in the $(k + 1)$ -th approximation is proportional to the r -th power of the error in the k -th approximation.

Some particular cases are:

1. **Linear Convergence**, $r = 1$. Here we have

$$\|e_{k+1}\| \approx C\|e_k\|$$

and the error decreases by a factor of C at each iteration. Note that C must be less than one for convergence to occur. The number C is sometimes called the **rate of convergence** in this case.

2. **Superlinear Convergence**, $r > 1$.

3. **Quadratic Convergence**, $r = 2$. Here we have

$$\|e_{k+1}\| \approx C\|e_k\|^2$$

and the error at each iteration is proportional to the square of the error at the previous iteration.

The higher the order of convergence r the faster the convergence. For example, the sequence

$$e_k = 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, \dots$$

shows linear convergence (with $C = 0.1$), while the sequence

$$e_k = 10^{-1}, 10^{-2}, 10^{-4}, 10^{-8}, \dots$$

shows quadratic convergence (with $C = 1$).

2 Fixed Point Iteration

Given an equation

$$f(x) = 0$$

rewrite it in an equivalent form

$$x = g(x).$$

Start with an approximation x_0 to the solution and compute a sequence of approximations via the **iteration**

$$x_{k+1} = g(x_k).$$

That is

$$\begin{aligned}x_1 &= g(x_0) \\x_2 &= g(x_1) \\x_3 &= g(x_2) \\&\vdots\end{aligned}$$

We will see that under certain conditions the sequence of approximations x_k converges to the solution x of $f(x) = 0$.

For example, Kepler's equation

$$M = E - e \sin(E)$$

can be written

$$E = M + e \sin(E).$$

For $M = 1$, $e = 0.5$ this gives the iteration

$$E_{k+1} = 1 + 0.5 \sin(E_k).$$

Starting with approximate solution $E_0 = 1.5$, we compute

$$\begin{aligned}E_1 &= 1 + 0.5 \sin(1.5) = 1.4987475 \\E_2 &= 1 + 0.5 \sin(1.4987475) = 1.4987028 \\E_3 &= 1 + 0.5 \sin(1.4987028) = 1.4987012 \\E_4 &= 1 + 0.5 \sin(1.4987012) = 1.4987011\end{aligned}$$

which is clearly converging.

2.1 Theory

We will look at the general theory of the iteration

$$x_{k+1} = g(x_k). \tag{2}$$

First, suppose the the sequence x_k converges to a number x^* , so that

$$\lim_{k \rightarrow \infty} x_k = x^*.$$

Taking limits of both sides of Equation (2) we have

$$\lim_{k \rightarrow \infty} x_{k+1} = \lim_{k \rightarrow \infty} g(x_k)$$

giving

$$x^* = g(x^*)$$

(This requires that $g(x)$ be continuous at a^* .)

Thus we see that if an iteration

$$x_{k+1} = g(x_k)$$

converges, then it converges to a solution of

$$x = g(x).$$

Such a solution is called a **fixed point** of the function $g(x)$. This name comes from the fact that if we start an iteration $x_{k+1} = g(x_k)$ at a fixed point x_0 , then it remains at that point, since, if x_0 satisfies $x_0 = g(x_0)$, then $x_1 = x_0$ and so on.

Now we turn to the question of when the iteration (2) converges. Let x^* be a fixed point of $g(x)$ so $x^* = g(x^*)$, and let $x_k = x^* + e_k$, where e_k can be thought of as the error in the approximation x_k to x^* . We will assume that x_k is close to x^* so that e_k is small and use the approximation

$$g(x + \epsilon) \approx g(x) + g'(x)\epsilon$$

valid for small ϵ . The iteration can be written

$$\begin{aligned} x^* + e_{k+1} &= g(x^* + e_k) \\ &\approx g(x^*) + g'(x^*)e_k \end{aligned}$$

Now, since $x^* = g(x^*)$,

$$e_{k+1} \approx g'(x^*)e_k.$$

The sequence x_k will converge to x^* provided e_k converges to zero. Let $g'(x^*) = K$ then

$$e_{k+1} \approx K e_k$$

and e_k will be multiplied by a factor of K at each iteration. Thus¹ e_k will converge to zero provided $|K| = |g'(x^*)| < 1$. Conversely, if $|K| = |g'(x^*)| > 1$, then e_k will increase and the iteration $x_{k+1} = g(x_k)$ will diverge away from the fixed point x^* (but may converge to a different fixed point). Finally, if $|K| = |g'(x^*)| = 1$, then the iteration may or may not converge.

Further, from

$$e_{k+1} \approx K e_k$$

we see that when $0 < |K| < 1$, the iteration is *linearly convergent*. When $K = g'(x^*) = 0$ convergence turns out to be at least quadratic. We will see an example of this later with Newton's method.

In our example involving Kepler's equation

$$E = 1 + 0.5 \sin(E)$$

we have

$$g(E) = 1 + 0.5 \sin(E)$$

and

$$g'(E) = 0.5 \cos(E).$$

Since $\cos(E)$ is never greater than one in magnitude, we have

$$|g'(E)| \leq 0.5$$

and our iteration will converge.

¹This argument can be made rigorous using the mean value theorem.

In summary: Let x^* be a fixed point of $g(x)$,

$$x^* = g(x^*)$$

and consider the iteration

$$x_{k+1} = g(x_k).$$

Then

1. If $|g'(x^*)| < 1$ the iteration will converge to x^* provided the initial approximation x_0 is sufficiently close to x^* .
2. If $|g'(x^*)| > 1$ then the iteration will diverge from x^* .

2.2 Application

Many algorithms for solving linear equations can be viewed as fixed point methods. Here we will look at direct application of the method.

To apply the fixed point method to an equation $f(x) = 0$ we need to rewrite it in an equivalent fixed point form $x = g(x)$. This can usually be done in many ways.

Consider, for example, the equation

$$f(x) = x^2 - x - 2 = 0.$$

This has two solutions $x = -1$ and $x = 2$. The equation $f(x) = 0$ can be written in the following fixed point forms:

1. $x = x^2 - 2, \quad g(x) = x^2 - 2.$
2. $x = \sqrt{x+2}, \quad g(x) = \sqrt{x+2}.$
3. $x = 1 + 2/x, \quad g(x) = 1 + 2/x.$

To determine whether the fixed point iteration converges, we need to examine the derivative of $g(x)$ at the fixed points. In general, we will have only approximate values for the fixed points, obtained, for example, by graphing $f(x)$. This will usually allow us to determine whether $|g'(x)| < 1$ and hence whether the fixed point iteration converges. For the examples above:

1. $g'(x) = 2x, g'(-1) = -2, g'(2) = 4$, and the fixed point iteration diverges in both cases.
2. $g'(x) = 1/(2\sqrt{x+2}), g'(-1) = 1/2, g'(2) = 1/4$, and the fixed point iteration converges in both cases.
3. $g'(x) = -2/x^2, g'(-1) = -2, g'(2) = -1/2$, and the fixed point iteration converges for the second fixed point but not the first.

Example

We will see how the example above works in practice, using the third example above. This uses the iteration

$$x_{k+1} = 1 + \frac{2}{x_k}$$

We will start at $x_0 = 1$ and perform 20 iterations:

```
-->function y = g(x)
-> y = 1+2/x
-->endfunction

-->x = zeros(1,21);

-->x(1) = 1;
-->for k = 1:20
--> x(k+1) = g(x(k));
-->end

-->x
x =

    column 1 to 7
!  1.    3.    1.6666667    2.2    1.9090909    2.047619    1.9767442 !

    column 8 to 12
!  2.0117647    1.994152    2.0029326    1.9985359    2.0007326 !

    column 13 to 17
!  1.9996338    2.0001831    1.9999085    2.0000458    1.9999771 !

    column 18 to 21
!  2.0000114    1.9999943    2.0000029    1.9999986 !

We can see the iteration converging to  $x = 2$  although fairly slowly. Computing the errors:

-->e = abs(x-2)
e =

    column 1 to 6
!  1.    1.    0.3333333    0.2    0.0909091    0.0476190 !
```

```

        column 7 to 11
!  0.0232558   0.0117647   0.0058480   0.0029326   0.0014641 !
        column 12 to 16
!  0.0007326   0.0003662   0.0001831   0.0000915   0.0000458 !
        column 17 to 21
!  0.0000229   0.0000114   0.0000057   0.0000029   0.0000014 !

```

We see that after a few steps the error approximately halves at each iteration in accord with theory. Computing the ratio of the errors at each step confirms this:

```

-->ee = e(2:21)./e(1:20)
ee =

```

```

        column 1 to 6
!  1.   0.3333333   0.6   0.4545455   0.5238095   0.4883721 !
        column 7 to 11
!  0.5058824   0.4970760   0.5014663   0.4992679   0.5003663 !
        column 12 to 16
!  0.4998169   0.5000916   0.4999542   0.5000229   0.4999886 !
        column 17 to 20
!  0.5000057   0.4999971   0.5000014   0.4999993 !

```

For this example, $K = |g'(2)| = 1/2$ and the observed errors are consistent with the relation

$$e_{k+1} \approx K e_k$$

derived earlier.