Solving Nonlinear Equations

Leigh J. Little SUNY Brockport October 10, 2018

Contents

1 What Does Solving an Equation Mean?

Solving equations comprises a significant part of science and mathematics. One of the main goals of scientific computing is to enable computers to solve equations. There are many equations that can be solved analytically. For example, all of the equations below can be solved for the unknown variable (with the understanding that some solutions might be expressed in terms of elementary functions or contain complex numbers):

$$\begin{aligned}
3(x-2)+6x &= 2(x+4)+8(x+2) \\
z^2-2z+4 &= 0 \\
t^4-t^3+t^2-t+1 &= 0 \\
e^{-4s} &= 5 \\
\cos y &= \frac{\pi}{3}
\end{aligned}$$

Not only do the above equations have solutions, but there is a great deal of mathematical theory that guarantees these equations have solutions.

However, most equations cannot be solved analytically (in much the same way that most functions don't have an antiderivative). For example, none of the equations below can be solved for the unknown variable:

$$\begin{aligned}
\cos x &= x \\
ye^y &= 7 \\
t^t &= \sqrt{2} \\
x^5-4x^4+3x^3-2x^2+x-1 &= 0
\end{aligned}$$

In addition, whether or not the equations even have solutions is a much more complicated problem. The last one does have 5 solutions because the Fundamental Theorem of Algebra says that a polynomial of degree n has n roots, but the others are less obvious.

Equations can also take on more elaborate forms. For example, the equation below is known as an integral equation

$$\int_0^x t+1\,dt = 1.$$

This can be solved using the Fundamental Theorm of Calculus.

$$\begin{aligned}
\int_0^x t+1\,dt &= 1 \\
\left.\left(\frac{1}{2}t^2+t\right)\right|_0^x &= 1 \\
\frac{1}{2}x^2+x &= 1 \\
\frac{1}{2}x^2+x-1 &= 0
\end{aligned}$$

$$
\begin{aligned}
x^2 + 2x - 2 &= 0 \\
x &= -1 \pm \sqrt{3}
\end{aligned}
$$

The above discussion leads to the following questions: When does an equation have a solution? If it has as solution, but the solution can't be expressed in terms of elementary functions, what do we mean by 'solve the equation'? The first question is beyond the scope of the course, but for the second question, the answer is typically expressed as 'find a value x accurate to t digits that satisfies the equation ....'. In other words, find an approximate solution to the equation.

There is no general approach for solving all equations. What has developed over time is a set of methods for generating approximate solutions to various classes equations with systems of linear equations, nonlinear equations, differential and partial differential equations being the most common.

2 Example: Solving $t^t = 3$

Consider the equation

$$t^t = 3.$$

This equation does have a solution. This can be seen by graphing the function $f(t) = t^t$ and verifying that it crosses the line $y = 3$ (see the Figure 1). How can we approximate the solution? One way would be to create

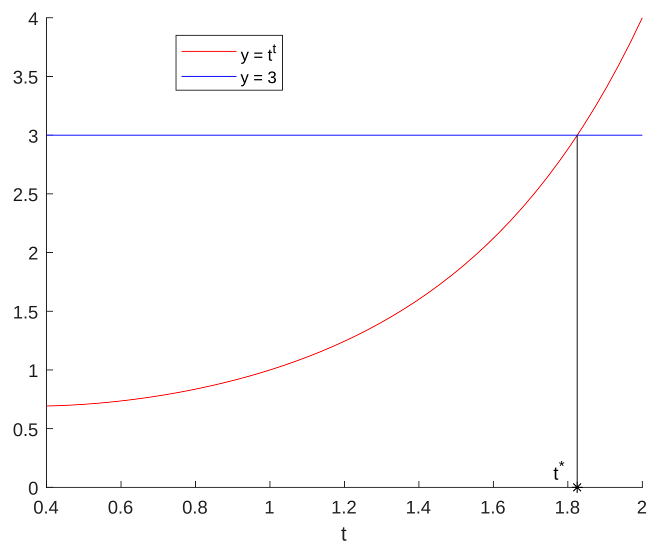

Figure 1: Graphical solution of $t^t = 3$.

a very precise graph and then zoom in on the intersection point. This works, but is not a good approach because it is time consuming and difficult for a computer to replicate.

Note that at $t = 0.4, t^t \approx 0.6931$ and that at $t = 2, t^t = 4$. Now we can make use of some important facts about t^t. The first is that t^t is a continuous function for $t > 0$. The second is that somewhere in the interval $[0.4, 2], t^t$ goes from being less than 3 to being greater than 3. This means that somewere in the interval $[0.4, 2]$ there has to be a point t^* where $t^t = 3$. The aim is to find an approximate value for t^*.

To obtain an approximate solution, take the interval $[0.4, 2]$ and find the midpoint

$$\text{midpoint} = \frac{0.4 + 2}{2} = 1.2.$$

Compute $1.2^{1.2} = 1.244$. We now have two intervals

$$[0.4, 1.2] \quad \text{and} \quad [1.2, 2]$$

Based on the value of t^t at 1.2, we know the solution has to be in the second interval. This is because the value of t^t at 1.2 is less than 3 and the value of t^t at 2 is greater than 3.

We can safely discard the first interval and repeat the above process on the interval $[1.2, 2]$ First, find the midpoint of this new interval

$$\text{midpoint} = \frac{1.2 + 2}{2} = 1.6.$$

Compute the value of t^t at this midpoint

$$1.6^{1.6} = 2.12.$$

This again gives 2 intervals

$$[1.2, 1.6] \quad \text{and} \quad [1.6, 2]$$

Based on the value of $1.6^{1.6}$, we know that the answer has to lie in the second interval. Thus we can discard the first interval and repeat the process on the interval $[1.6, 2]$.

This is known as the bisection method. It begins with some interval that the solution is known to lie in. The method then proceeds by dividing the interval into two equal parts and then discarding the part that is not needed. Eventually, this will produce an answer to the desired accuracy. This process is what we are looking for when it comes to solving equations. It consists of steps that ultimately reduce to a series of calculations and an `if-then` test to determine how to proceed. Repeating the process on smaller and smaller intervals is a looping operation.