The Newton-Raphson Method

1 Introduction

The Newton-Raphson method, or Newton Method, is a powerful technique for solving equations numerically. Like so much of the differential calculus, it is based on the simple idea of linear approximation. The Newton Method, properly used, usually homes in on a root with devastating efficiency.

The essential part of these notes is Section 2.1, where the basic formula is derived, Section 2.2, where the procedure is interpreted geometrically, and—of course—Section 6, where the problems are. Peripheral but perhaps interesting is Section 3, where the birth of the Newton Method is described.

2 Using Linear Approximations to Solve Equations

Let $f(x)$ be a well-behaved function, and let r be a root of the equation $f(x) = 0$. We start with an estimate x_0 of r. From x_0, we produce an improved—we hope—estimate x_1. From x_1, we produce a new estimate x_2. From x_2, we produce a new estimate x_3. We go on until we are 'close enough' to r—or until it becomes clear that we are getting nowhere.

The above general style of proceeding is called *iterative*. Of the many iterative root-finding procedures, the Newton-Raphson method, with its combination of simplicity and power, is the most widely used. Section 2.4 describes another iterative root-finding procedure, the *Secant Method*.

Comment. The initial estimate is sometimes called x_1, but most mathematicians prefer to start counting at 0.

Sometimes the initial estimate is called a "guess." The Newton Method is usually very very good if x_0 is close to r, and can be horrid if it is not. The "guess" x_0 should be chosen with care.

2.1 The Newton-Raphson Iteration

Let x_0 be a good estimate of r and let $r = x_0 + h$. Since the true root is r, and $h = r - x_0$, the number h measures how far the estimate x_0 is from the truth.

Since h is 'small,' we can use the linear (tangent line) approximation to conclude that

$$0 = f(r) = f(x_0 + h) \approx f(x_0) + hf'(x_0),$$

and therefore, unless $f'(x_0)$ is close to 0,

$$h \approx -\frac{f(x_0)}{f'(x_0)}.$$

It follows that

$$r = x_0 + h \approx x_0 - \frac{f(x_0)}{f'(x_0)}.$$

Our new improved (?) estimate x_1 of r is therefore given by

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

The next estimate x_2 is obtained from x_1 in exactly the same way as x_1 was obtained from x_0:

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}.$$

Continue in this way. If x_n is the current estimate, then the next estimate x_{n+1} is given by

$$\boxed{x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}} \tag{1}$$

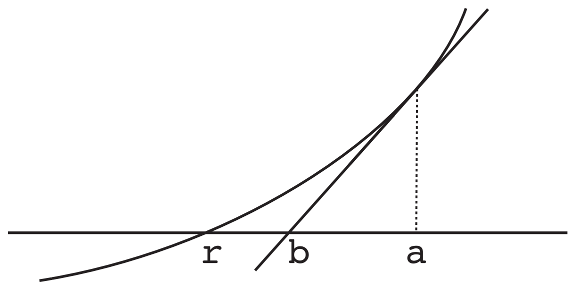
2.2 A Geometric Interpretation of the Newton-Raphson Iteration

In the picture below, the curve $y = f(x)$ meets the x-axis at r. Let a be the current estimate of r. The tangent line to $y = f(x)$ at the point $(a, f(a))$ has equation

$$y = f(a) + (x - a)f'(a).$$

Let b be the x-intercept of the tangent line. Then

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



Compare with Equation 1: b is just the 'next' Newton-Raphson estimate of r. The new estimate b is obtained by drawing the tangent line at $x = a$, and then sliding to the x-axis along this tangent line. Now draw the tangent line at $(b, f(b))$ and ride the new tangent line to the x-axis to get a new estimate c. Repeat.

We can use the geometric interpretation to design functions and starting points for which the Newton Method runs into trouble. For example, by putting a little bump on the curve at $x = a$ we can make b fly far away from r. When a Newton Method calculation is going badly, a picture can help us diagnose the problem and fix it.

It would be wrong to think of the Newton Method simply in terms of tangent lines. The Newton Method is used to find complex roots of polynomials, and roots of systems of equations in several variables, where the geometry is far less clear, but linear approximation still makes sense.

2.3 The Convergence of the Newton Method

The argument that led to Equation 1 used the informal and imprecise symbol $\approx$. We probe this argument for weaknesses.

No numerical procedure works for *all* equations. For example, let $f(x) = x^2 + 17$ if $x \neq 1$, and let $f(1) = 0$. The behaviour of $f(x)$ *near* 1 gives no clue to the fact that $f(1) = 0$. Thus no method of successive approximation can arrive at the solution of $f(x) = 0$. To make progress in the analysis, we need to assume that $f(x)$ is in some sense smooth. We will suppose that $f''(x)$ (exists and) is continuous near r.

The tangent line approximation is—*an approximation.* Let's try to get a handle on the error. Imagine a particle travelling in a straight line, and let $f(x)$ be its position at time x. Then $f'(x)$ is the velocity at time x. If the acceleration of the particle were always 0, then the *change* in position from time x_0 to time $x_0 + h$ would be $hf'(x_0)$. So the position at time $x_0 + h$

would be $f(x_0) + hf'(x_0)$—note that this is the tangent line approximation, which we can also think of as the zero-acceleration approximation.

If the velocity *varies* in the time from x_0 to $x_0 + h$, that is, if the acceleration is not 0, then in general the tangent line approximation will not correctly predict the displacement at time $x_0 + h$. And the bigger the acceleration, the bigger the error. It can be shown that if f is twice differentiable then the error in the tangent line approximation is $(1/2)h^2 f''(c)$ for some c between x_0 and $x_0 + h$. In particular, if $|f''(x)|$ is large between x_0 and $x_0 + h$, then the error in the tangent line approximation is large. Thus we can expect *large second derivatives* to be bad for the Newton Method. This is what goes wrong in Problem 7(b).

In the argument for Equation 1, from $0 \approx f(x_0) + hf'(x_0)$ we concluded that $h \approx -f(x_0)/f'(x_0)$. This can be quite wrong if $f'(x_0)$ is close to 0: note that 3.01 is close to 3, but $3.01/10^{-8}$ is not at all close to $3/10^{-8}$. Thus we can expect *first derivatives close to* 0 to be bad for the Newton Method. This is what goes wrong in Problems 7(a) and 8.

These informal considerations can be turned into positive *theorems* about the behaviour of the error in the Newton Method. For example, if $|f''(x)/f'(x)|$ is not too large near r, and we start with an x_0 close enough to r, the Newton Method converges very fast to r. (Naturally, the theorem gives "not too large," "close enough," and "very fast" precise meanings.)

The study of the behaviour of the Newton Method is part of a large and important area of mathematics called *Numerical Analysis.*

2.4 The Secant Method

The Secant Method is the most popular of the many variants of the Newton Method. We start with *two* estimates of the root, x_0 and x_1. The iterative formula, for $n \geq 1$ is

$$x_{n+1} = x_n - \frac{f(x_n)}{Q(x_{n-1}, x_n)}, \quad \text{where} \quad Q(x_{n-1}, x_n) = \frac{f(x_{n-1}) - f(x_n)}{x_{n-1} - x_n}.$$

Note that if x_n is close to x_{n-1}, then $Q(x_{n-1}, x_n)$ is close to $f'(x_n)$, and the two methods do not differ by much. We can also compare the methods geometrically. Instead of sliding along the tangent line, the Secant Method slides along a nearby secant line.

The Secant Method has some advantages over the Newton Method. It is more stable, less subject to the wild gyrations that can afflict the Newton Method. (The differences are not great, since the geometry is nearly the same.) To use the Secant Method, we do not need the derivative, which can be expensive to calculate. The Secant Method, when it is working well, which is most of the time, is fast. Usually we need about 45 percent more iterations than with the Newton Method to get the same accuracy, but each iteration is cheaper. Your mileage may vary.

3 Newton's Newton Method

> Nature and Nature's laws lay hid in night:
> God said, Let Newton be! And all was light.
> *Alexander Pope, 1727*

It didn't quite happen that way with the Newton Method. Newton had no great interest in the numerical solution of equations—his only numerical example is a cubic. And there was a long history of efficient numerical solution of cubics, going back at least to Leonardo of Pisa ("Fibonacci," early thirteenth century).

At first sight, the method Newton uses doesn't look like the Newton Method we know. The derivative is not even mentioned, even though the same manuscript develops the Newtonian version of the derivative!

Newton's version of the Method is mainly a pedagogical device to explain something quite different. Newton *really* wanted to show how to solve the following 'algebraic' problem: given an equation $F(x, y) = 0$, express y as a series in powers of x.

But before discussing his novel *symbolic* calculations, Newton tried to motivate the idea by doing an analogous calculation with *numbers,* using the equation

$$y^3 - 2y - 5 = 0.$$

We describe, quoting (in translation) from Newton's *De Methodis Serierum et Fluxionum,* how he deals with the equation. Like any calculation, Newton's should be followed with pencil in hand.

> "Let the equation $y^3-2y-5 = 0$ be proposed for solution and let the number 2 be found, one way or another, which differs from the required root by less than its tenth part. I then set $2+p = y$ and in place of y in the equation I substitute $2 + p$. From this there arises the new equation
>
> $$p^3 + 6p^2 + 10p - 1 = 0.$$
>
> whose root p is to be sought for addition to the quotient. Specifically, (when p^3+6p^2 is neglected because of its smallness) we have

> $10p - 1 = 0$, or $p = 0.1$ narrowly approximates the truth. Accordingly, I write 0.1 in the quotient and, supposing $0.1 + q = p$, I substitute this fictitious value for it as before. There results
>
> $$q^3 + 6.3q^2 + 11.23q + 0.061 = 0.$$
>
> And since $11.23q + 0.061 = 0$ closely approaches the truth, in other words very nearly $q = -0.0054 \ldots$."

Newton puts $-0.0054 + r$ for q in $q^3 + 6.3q^2 + 11.23q + 0.061 = 0$, Neglecting the terms in r^3 and r^2, he concludes that $r \approx -0.00004852$. His final estimate for the root is $2 + p + q + r$, that is, 2.09455148.

As we go through Newton's calculation, it is only with hindsight that we see in it the germs of the method we now call Newton's. When Newton discards terms in powers of p, q, and r higher than the first, he is in effect doing linear approximation. Note that $2 + p$, $2 + p + q$, and $2 + p + q + r$ are, more or less, the numbers y_1, y_2, and y_3 of Problem 3.

Newton substitutes $0.1 + q$ for p in $p^3 + 6p^2 + 10p - 1 = 0$. Surely he knows that it is more sensible to substitute $2.1 + q$ for y in the original equation $y^2 - 2y - 5 = 0$. But his numerically awkward procedure, with an ever changing equation, is the right one for the series expansion problems he is really interested in. And Newton goes on to use his method to do something really new: he finds infinite series for, among others, the sine and cosine functions.

Comment. When Newton asks that we make sure that the initial estimate "differs from the required root by less than its tenth part," he is trying (with no justification, and he is wrong) to quantify the idea that we should start close to the root. His use of the word "quotient" may be confusing. He doesn't really mean quotient, he is just making an analogy with the usual 'long division' process.

Newton says that $q = -0.0054$. But $-0.61/11.23$ is about -0.00543188. Here Newton truncates deliberately. He is aiming for 8 place accuracy, but knows that he can work to less accuracy at this stage. Newton used a number of tricks to simplify the arithmetic—an important concern in the Before Calculators Era.

Historical Note. Newton's work was done in 1669 but published much later. Numerical methods related to the Newton Method were used by al-Kāshī, Viète, Briggs, and Oughtred, all many years before Newton.

Raphson, some 20 years after Newton, got close to Equation 1, but only for *polynomials* $P(y)$ of degree 3, 4, 5, …, 10. Given an estimate g for a root, Raphson computes an 'improved' estimate $g+x$. He sets $P(g+x)=0$, expands, discards terms in x^k with $k \geq 2$, and solves for x. For polynomials, Raphson's procedure is equivalent to linear approximation.

Raphson, like Newton, seems unaware of the connection between his method and the derivative. The connection was made about 50 years later (Simpson, Euler), and the Newton Method finally moved beyond polynomial equations. The familiar geometric interpretation of the Newton Method may have been first used by Mourraille (1768). Analysis of the convergence of the Newton Method had to wait until Fourier and Cauchy in the 1820s.

4 Using the Newton-Raphson Method

4.1 Give Newton a Chance

- Give Newton the right equation. In 'applied' problems, that's where most of the effort goes. See Problems 10, 11, 12, and 13.

- Give Newton an equation of the form $f(x)=0$. For example, $xe^x=1$ is not of the right form: write it as $xe^x-1=0$. There are many ways to make an equation ready for the Newton Method. We can rewrite $x\sin x=\cos x$ as $x\sin x-\cos x=0$, or $x-\cot x=0$, or $1/x-\tan x=0$, or .... *How* we rewrite can have a dramatic effect on the behaviour of the Newton Method. But mostly it is not worth worrying about.

- A Newton Method calculation can go bad in various ways. We can usually tell when it does: the first few x_n refuse to settle down. There is almost always a simple fix: spend time to find a good starting x_0.

- A graphing program can help with x_0. Graph $y=f(x)$ and eyeball where the graph crosses the x-axis, zooming in if necessary. For simple problems, a graphing program can even produce a final answer. But to solve certain scientific problems we must find, without human intervention, the roots of tens of thousands of equations. Graphing programs are no good for that.

- Even a rough sketch can help. It is not immediately obvious what $y=x^2-\cos x$ looks like. But the roots of $x^2-\cos x=0$ are the x-coordinates of the points where the familiar curves $y=x^2$ and $y=\cos x$ meet. It is easy to see that there are two such points, symmetric across the y-axis. Already at $x=1$ the curve $y=x^2$ is above $y=\cos x$. A bit of fooling around with a calculator gives the good starting point $x_0=0.8$.

4.2 The Newton Method can go bad

- Once the Newton Method catches scent of the root, it usually hunts it down with amazing speed. But since the method is based on *local* information, namely $f(x_n)$ and $f'(x_n)$, the Newton Method's sense of smell is deficient.

- If the initial estimate is not close enough to the root, the Newton Method may not converge, or may converge to the wrong root. See Problem 9.

- The successive estimates of the Newton Method may converge to the root too slowly, or may not converge at all. See Problems 7 and 8.

4.3 The End Game

- When the Newton Method works well, which (with proper care) is most of the time, the number of correct decimal places roughly doubles with each iteration.

- If we want to compute a root correct to say 5 decimal places, it seems sensible to compute until two successive estimates agree to 5 places. While this is theoretically unsound, it is a widely used rule of thumb. And the second estimate will likely be correct to about 10 places.

- We can usually *verify* that our final answer is close enough. Suppose, for example, that b is our estimate for a root of $f(x) = 0$, where f is continuous. If $f(b - 10^{-8})$ and $f(b + 10^{-8})$ have *different signs*, then there must be a root between $b - 10^{-8}$ and $b + 10^{-8}$, so we know that the error in b has absolute value less than 10^{-8}.

5 A Sample Calculation

We use the Newton Method to find a non-zero solution of $x = 2\sin x$. Let $f(x) = x - 2\sin x$. Then $f'(x) = 1 - 2\cos x$, and the Newton-Raphson iteration is

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{x_n - 2\sin x_n}{1 - 2\cos x_n} = \frac{2(\sin x_n - x_n\cos x_n)}{1 - 2\cos x_n}. \tag{2}$$

Let $x_0 = 1.1$. The next six estimates, to 3 decimal places, are:

$$\begin{array}{llllll} x_1 = & 8.453 & x_3 = & 203.384 & x_5 = & -\ \ 87.471 \\ x_2 = & 5.256 & x_4 = & 118.019 & x_6 = & -203.637. \end{array}$$

Things don't look good, and they get worse. It turns out that $x_{35} < -64000000$. We could be stubborn and soldier on. Miracles happen—but not often. (One happens here, around $n = 212$.)

To get an idea of what's going wrong, use a graphing program to graph $y = x - 2\sin x$, and recall that x_{n+1} is where the tangent line at x_n meets the x-axis. The bumps on $y = x - 2\sin x$ confuse the Newton Method terribly.

Note that choosing $x_0 = \pi/3 \approx 1.0472$ leads to immediate disaster, since then $1 - 2\cos x_0 = 0$ and therefore x_1 does not exist. Thus with $x_0 = 1.1$ we are starting on a (nearly) flat part of the curve. Riding the tangent line takes us to an x_1 quite far from x_0. And x_1 is also on a flat part of the curve, so x_2 is far from x_1. And x_2 is on a flat part of the curve: the chaotic ride continues.

The trouble was caused by the choice of x_0. Let's see whether we can do bettter. Draw the curves $y = x$ and $y = 2\sin x$. A quick sketch shows that they meet a bit past $\pi/2$. But we will be sloppy and take $x_0 = 1.5$. Here are the next six estimates, to 19 places—the computations were done to 50.

$$
\begin{aligned}
x_1 &= 2.0765582006304348291 & \qquad x_4 &= 1.8954942764727706570\\
x_2 &= 1.9105066156590806258 & \qquad x_5 &= 1.8954942670339809987\\
x_3 &= 1.8956220029878460925 & \qquad x_6 &= 1.8954942670339809471
\end{aligned}
$$

The next iterate x_7 agrees with x_6 in the first 19 places, indeed in the first 32, and the true root is equal to x_6 to 32 places.

Comment. The equation $x = 2\sin x$ can be rewritten as $2/x - 1/\sin x = 0$. If x_0 is *any* number in $(0, \pi)$, Newton quickly takes us to the root. The reformulation has changed the geometry: there is no longer a flat spot inconveniently near the root. Rewriting the equation as $(\sin x)/x = 1/2$ also works nicely.

6 Problems

1. Use the Newton-Raphson method, with 3 as starting point, to find a fraction that is within 10^{-8} of $\sqrt{10}$. Show (without using the square root button) that your answer is indeed within 10^{-8} of the truth.

2. Let $f(x) = x^2 - a$. Show that the Newton Method leads to the recurrence
$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right).$$

Heron of Alexandria (60 CE?) used a pre-algebra version of the above recurrence. It is still at the heart of computer algorithms for finding square roots.

3. Newton's equation $y^3 - 2y - 5 = 0$ has a root near $y = 2$. Starting with $y_0 = 2$, compute y_1, y_2, and y_3, the next three Newton-Raphson estimates for the root.

4. Find all solutions of $e^{2x} = x + 6$, correct to 4 decimal places; use the Newton Method.

5. Find all solutions of $5x + \ln x = 10000$, correct to 4 decimal places; use the Newton Method.

6. A calculator is defective: it can only add, subtract, and multiply. Use the equation $1/x = 1.37$, the Newton Method, and the defective calculator to find $1/1.37$ correct to 8 decimal places.

7. (a) A devotee of Newton-Raphson used the method to solve the equation $x^{100} = 0$, using the initial estimate $x_0 = 0.1$. Calculate the next five Newton Method estimates.

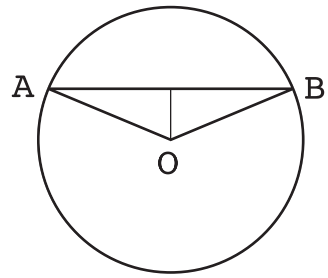
   (b) The devotee then tried to use the method to solve $3x^{1/3} = 0$, using $x_0 = 0.1$. Calculate the next ten estimates.

8. Suppose that
$$f(x) = \begin{cases} e^{-1/x^2} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$
   The function f is continuous everywhere, in fact differentiable arbitrarily often everywhere, and 0 is the only solution of $f(x) = 0$. Show that if $x_0 = 0.0001$, it takes more than one hundred million iterations of the Newton Method to get below 0.00005.

9. Use the Newton Method to find the smallest and the second smallest positive roots of the equation $\tan x = 4x$, correct to 4 decimal places.

10. The circle below has radius 1, and the *longer* circular arc joining A and B is twice as long as the chord AB. Find the length of the chord AB, correct to 18 decimal places.

11. Find, correct to 5 decimal places, the x-coordinate of the point on the curve $y = \ln x$ which is closest to the origin. Use the Newton Method.



12. It costs a firm $C(q)$ dollars to produce q grams per day of a certain chemical, where

$$C(q) = 1000 + 2q + 3q^{2/3}$$

The firm can sell any amount of the chemical at \$4 a gram. Find the break-even point of the firm, that is, how much it should produce per day in order to have neither a profit nor a loss. Use the Newton Method and give the answer to the nearest gram.

13. A loan of A dollars is repaid by making n equal monthly payments of M dollars, starting a month after the loan is made. It can be shown that if the monthly interest rate is r, then

$$Ar = M\left(1 - \frac{1}{(1+r)^n}\right).$$

A car loan of \$10000 was repaid in 60 monthly payments of \$250. Use the Newton Method to find the monthly interest rate correct to 4 significant figures.