Approximations

Paul E. Johnson

January 26, 2006

1 Linear Approximation

1.1 Linear equation

We start here because it is particularly easy.

Suppose $y = f(x) = a + bx$. That's a straight line. Draw a graph:

The derivative of that line, $\frac{dy}{dx} = f'(x) = b$, the same quantity you ordinarily know as the "slope".

Consider you want to build an approximation to the value $y = f(x)$ but you have to base your approximation from a "starting estimate" at a point x_0. We want to "estimate" the value of y at some point, x.

Quite obviously, our "best estimate" is

$$\hat{y} = f(x) = f(x_0) + b \cdot (x - x_0) \tag{1}$$

The "estimate" using this formula is always exactly right, of course, because $f(x)$ is a straight line.

1.2 Nonlinear equation

Now imagine that $f(x)$ is not a straight line. Begin at x_0 and you can calculate $f(x_0)$. Then the linear approximation to $f(x_0)$ is

$$\hat{y} = f(x_0) + f'(x)(x - x_0) \tag{2}$$

Note that, depending on how sharply the curve changes, the linear approximation may be either a good or bad estimate. Make a picture:

2 Taylor Series Approximation

Here is an intuition. Suppose your approximation, $\hat{y}$ is too low. That happened because the curve $f(x)$ is concave upwards–the slope is rapidly increasing. That means the second derivative, $f''(x) > 0$.

So, to improve the approximation, one might add another term that includes $f''(x)$. Suppose, for example, you tried:

$$\hat{y} = f(x_0) + f'(x_0)(x - x_0) + f''(x_0)(x - x_0) \tag{3}$$

That might get you closer to the mark. It could be that the guess becomes too large after that correction.

The problem with 4 is that it is completely unsystematic. There's no "guarantee" that the error, $\hat{y} - f(x)$ will be within any given limits.

That's where Taylor's theorem enters the picture. The theorem holds that you can keep adding terms to the approximation and it will get "closer and closer" to the correct value. In fact, if you set your tolerable error level, you can add enough terms to make the actual error smaller than the tolerable level.

What is the magical formula? A Taylor series is:

$$\hat{y} = f(x_0) + f'(x_0)(x - x_0) + (\frac{1}{2})f''(x_0)(x - x_0)^2 + (\frac{1}{6})f'''(x_0)(x - x_0)^3 + \tag{4}$$

$$+ \ldots + (\frac{1}{N!})f^N(x_0)(x - x_0)^N$$

For many well-behaved functions (insert boring math details here), as $N \to \infty$, the series converges–approaches a finite value. That means, for x near to the starting estimate x_0, then the value of $\hat{y}$ is equal to $f(x)$. In fact, it gets "arbitrarily close".

In most applied problems, it is not necessary to raise N to a very high number. In fact, most of the time, applications only increase N to 2 or 3.

3 Newton's method of finding roots

Given a function $g(x)$, we want to find a maximum or minimum value. Assuming the function is differentiable, we want a point where $\frac{\partial y}{\partial x} = g'(x) = 0$.

3.1 Relabel the derivative at x as $h(x)$.

Since I get tired of typing $g'(x)$ or the symbol $\frac{\partial y}{\partial x}$, I will now just assign a new letter to the derivitive. In the past, I confused people by re-using f here, so I'll try to avoid that mistake by introducing the letter h . From this point forward,

$$h(x) = g'(x)$$

3.2 Roots.

In order to find a critical point–max or min–we need to find a value of x such that $h(x) = 0$. Such values of x are called "roots."

3.3 Crude guess-timation

Suppose $g(x)$ is "U" shaped and suppose the derivative $g'(x)$ (same as $h(x)$) can be calculated. Select some point $x_0$and find out whether the slope $g'(x)$ (same as $h(x)$)is positive or negative. Suppose $g'(x_0) < 0$ (meaning $h(x_0) < 0$), then we should move to the right to a point x_1 and calculate $g'(x_1)$. Keep going, over and over. As x_N gets closer and closer to the bottom of the "U" shape, then $g'(x_N)$ will get smaller and smaller, until, when the exact bottom of the curve is found, $g'(x) = 0$. That is, we find where $h(x) = 0$.

That might get boring, unless you have a good algorithm with which to choose each successive guess. One of the first proposals was offered by Newton, one of the inventors of the calculus of infinite differences.

3.4 Newton's method

Newton's method of finding the roots of a function, such as $h(x)$, is rather simple.

We think of x_0 as the initial guess and then x_1 is the "new improved guess." So, repeatedly the following calculation is made. First, calculate a "newer improved guess" with this formula:

$$new\,guess = x_0 - \frac{h(x_0)}{h'(x_0)}$$

$$x_0 = x_1$$

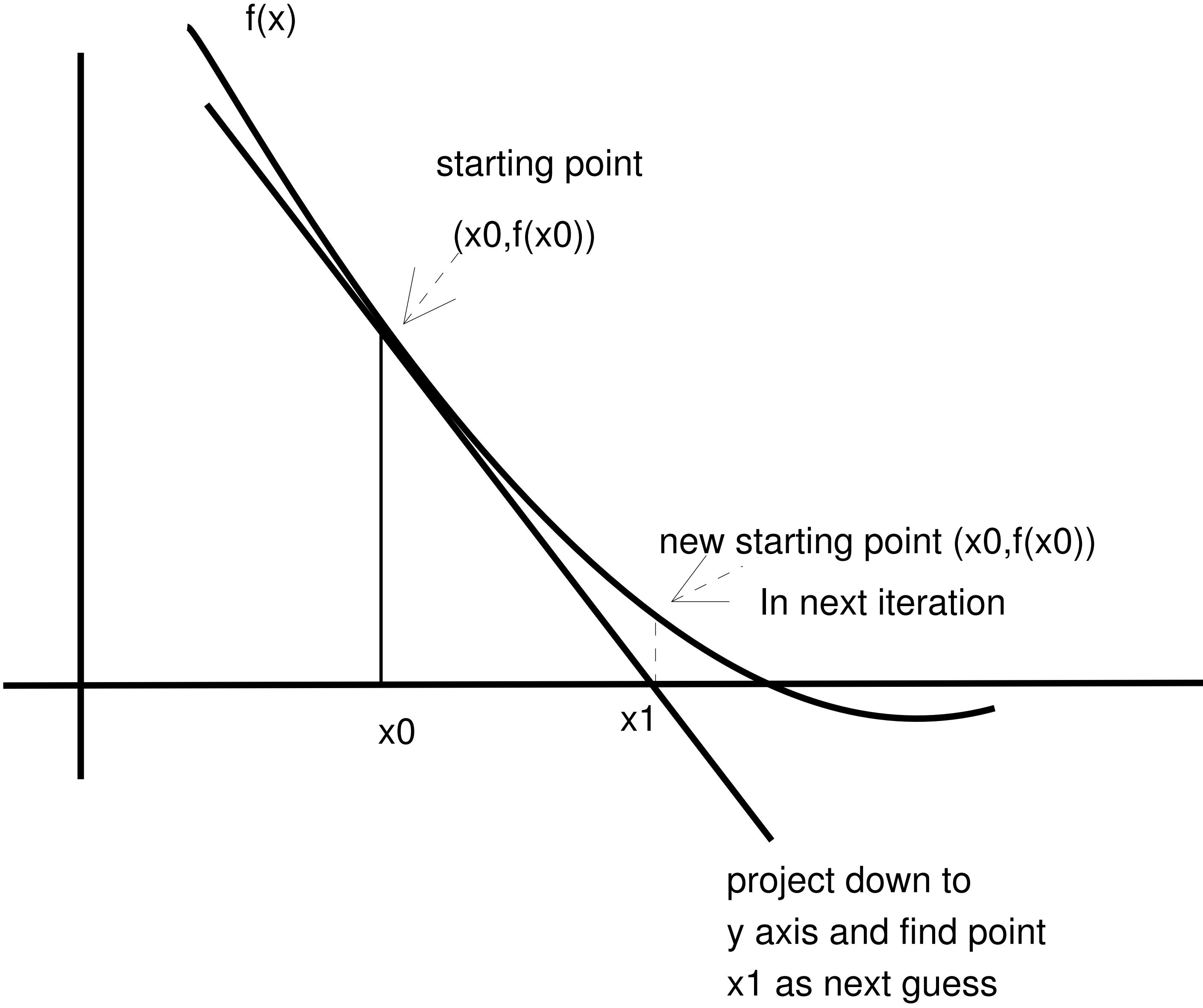
$$x_1 = new\,guess$$

Make that calculation repeatedly, until $|x_1 - x_0|$ is smaller than the target value. Often, one will see the stoping criterion stated as a tolerance parameter, such as

$$\frac{|x_1 - x_0|}{x_0} < 10^{-6}$$

Note that, if the difference between x_1 and x_0 is very very small, then that must mean that $h(x_0)$ has become very small.

How to explain this algorithm? Perhaps a picture is worth a thousand words.



It is not too difficult to justify this approach. If one calculates a linear approximation to forecast $h(x_1)$ from a point x_0,

$$\hat{y} = h(\hat{x_1}) = h(x_0) + h'(x_0)(x_1 - x_0)$$

Recall we are looking for the point where the function, $h(x)$, equals 0.

So think backwards for a minute.

Don't estimate y, but rather set $\hat{y} = 0$.

Solve for x_1, the value for which $\hat{y} = 0$.

$$0 = h(x_0) + h'(x_0)(x_1 - x_0)$$

which implies:

$$-h(x_0) = h'(x_0)(x_1 - x_0)$$

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

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

4 Think backwards again.

We were talking about adjusting x to find the critical points. If you replace x in the above with a parameter, say θ or β, then everything works OK. You are finding optimal estimates.

In statistics, we have some function–a maximum likelihood equation or a sum of squared residuals. The input data, the y's and x's, is just seen as "numerical constants" in the formula. The variables that we are adjusting to maximize or minimize things, are the values of the parameters.

So, if you translate that idea into the above story, we would not be maximizing $g(x)$, but rather a function that depends on β, as in maximum likelihood model, where the MLE is the value $\hat{\beta}$ that maximizes this:

$$lnL(\beta|X,y) = \sum_{i=1}^{N} prob(y|X,\beta)$$

Or, in least squares analysis, the value $\hat{b}$ that minimizes the sum of squared errors between the observed y_i and the predicted value $f(X,b)$:

$$S(b|y,x) = \sum_{i=1}^{N} (y_i - f(X_i,b))'(y_i - f(X_i,b))$$

In the OLS case, where we assume the linear model

$$y = f(X,b) = Xb + e$$

it is not difficult to find the minimum of the sum of squares.

It is very difficult to find the minimum of the sum of squares if you put some other function in place of $f(X,b)$. And, in the maximum likelihood case, if the probability model is something not Normal, or the role of the parameters is complicated, then maximization is difficult.

When $lnL()$ or $S()$ can't be solved with algebra, then the maxima and minima have to be found by numerical approximation. And that's where approximations and algorithms like Newton's come into play.