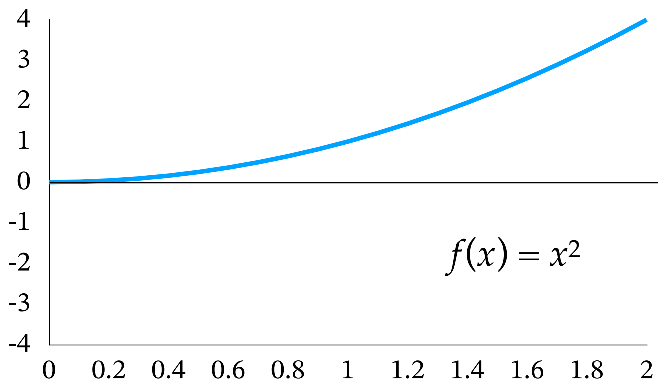
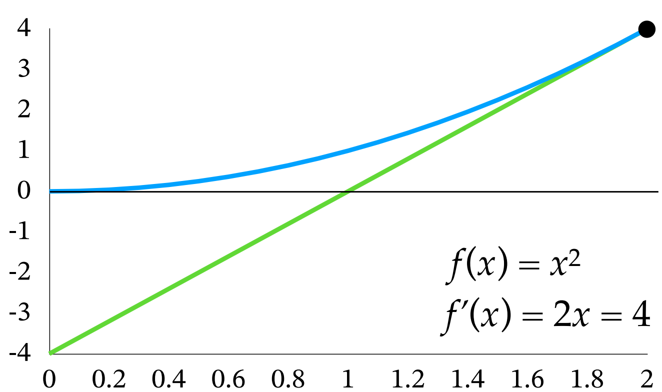
Newton's Algorithm: Zero (Root) Finding

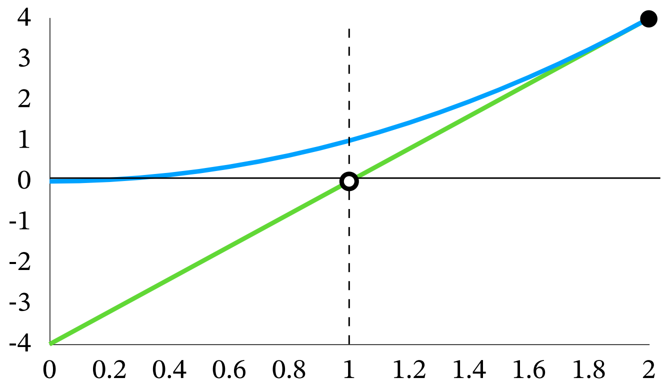
Consider the function $f(x) = x^2$, the graph of which is shown below. The goal is to find the value of x that returns $f(x) = 0$ (the root).



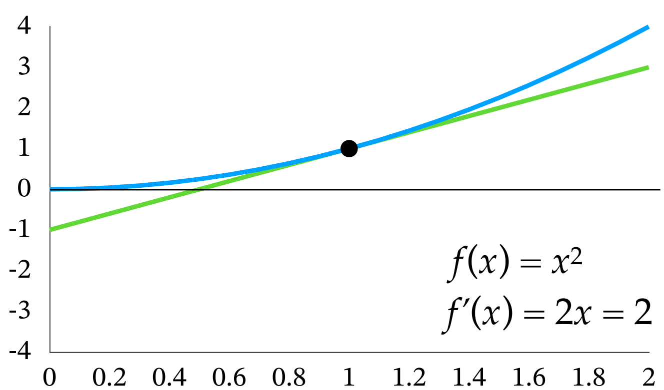
The tangent of the function at a particular value of x (i.e., the first derivative, or gradient), is a good approximation to the quadratic at that particular value of x. Differentiating the function gives the derivative equation, which is $f'(x) = 2x$. The graph below shows the tangent line that results from evaluating the function at $x = 2$.



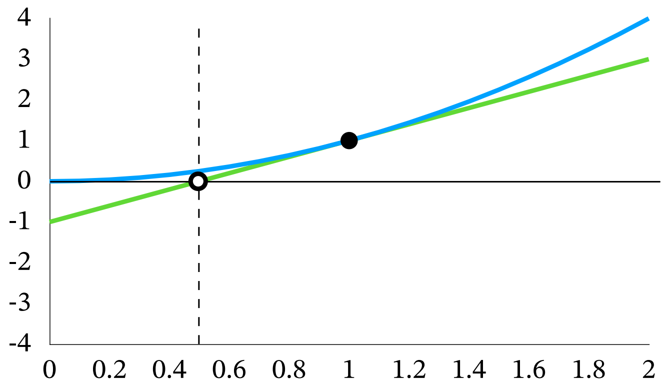
The point at which the tangent line crosses 0 is an approximation to the true root of the function. The graphic below shows this projection.



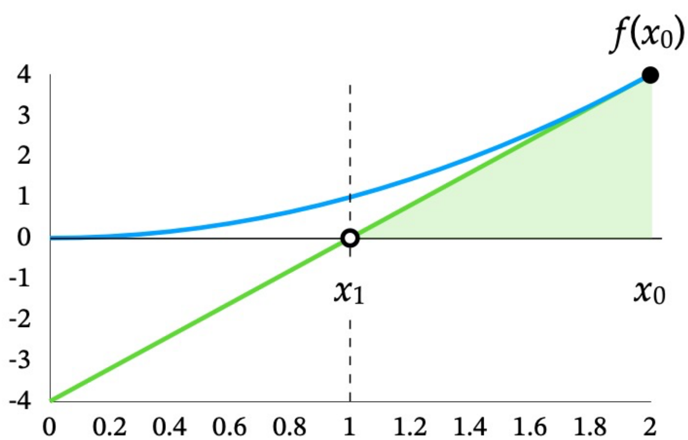
Newton's algorithm uses this approximation as the starting point for the next update, which involves the tangent line of the function evaluated at $x = 1$, as shown below.



The point at which the new tangent line crosses 0 is the updated estimate of the root, as shown in the graph below.



Successive updates to the algorithm get closer and closer to the function's root. To formalize the algorithm, consider a starting value of x and its successive update, x_0 and x_1, respectively. Revisiting the initial graph, suppose that $x_0 = 2$, the graphic for which is shown below.



The slope of the tangent line evaluated at x_0 equals rise over run, as follows.

$$f'(x_0) = \frac{f(x_0)}{x_0 - x_1} \tag{1}$$

The goal is to determine the value of x_1, which is the point at which the tangent line crosses 0. You first rearrange the equation to isolate $x_0 - x_1$

$$x_0 - x_1 = \frac{f(x_0)}{f'(x_0)} \tag{2}$$

then you solve for x_1.

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} \tag{3}$$

More generally, the updated root at iteration $t + 1$ is as follows.

$$x_{t+1} = x_t - \frac{f(x_t)}{f'(x_t)} \tag{4}$$

Newton's Algorithm: Optimization

The goal of root finding was to find the point at which the function $f(x)$ equals 0. For finding a minimum or a maximum, we need to find the point at which the *derivative* of a function equals 0. That is, instead of finding the point at which $f(x)$ equals 0, we find the point at which $f'(x)$ equals 0. The use of Newton's algorithm for finding a minimum or maximum is a perfect analogy with its application to root finding: $f(x_t)$ becomes $f'(x_t)$, and $f'(x_t)$ becomes $f''(x_t)$.

$$x_{t+1} = x_t - \frac{f'(x_t)}{f''(x_t)} \tag{5}$$

To illustrate Newton's algorithm, consider using maximum likelihood to estimate the mean and variance, an operation that involves finding the value of μ and σ^2 where the respective partial derivatives of the log-likelihood equal 0. With complete data, the mean and variance are independent (i.e., the cross-product derivative equals 0), so we can treat estimation as a pair of 1-dimensional optimization problems. The Newton updating step from Equation 5 can be written as follows, where ℓ is the log-likelihood function, and θ is the parameter of interest.

$$\theta_{t+1} = \theta_t - \frac{\ell'(\theta_t)}{\ell''(\theta_t)} \tag{6}$$

The univariate normal log-likelihood for a sample of N cases is as follows.

$$\ell\left(\mu, \sigma^2 | \text{data}\right) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\left(\sigma^2\right) - \frac{1}{2}\left(\sigma^2\right)^{-1} \sum_{i=1}^{N}(Y_i - \mu)^2 \tag{7}$$

Applying differential calculus rules gives the first and second derivatives of the log-likelihood function with respect to the mean

$$\begin{aligned} \ell'(\mu) &= \left(\sigma^2\right)^{-1} \sum_{i=1}^{N}(Y_i - \mu) \\ \ell''(\mu) &= -\frac{N}{\sigma^2} \end{aligned} \tag{8}$$

and the corresponding derivatives with respect to the variance are as follows.

$$\begin{aligned} \ell'\left(\sigma^2\right) &= -\frac{N}{2}\left(\sigma^2\right)^{-1} + \frac{1}{2}\left(\sigma^2\right)^{-2} \sum_{i=1}^{N}(Y_i - \mu)^2 \\ \ell''\left(\sigma^2\right) &= \frac{N}{2}\left(\sigma^2\right)^{-2} - \left(\sigma^2\right)^{-3} \sum_{i=1}^{N}(Y_i - \mu)^2 \end{aligned} \tag{9}$$

Newton's algorithm is a second-order approximation, meaning that it uses a quadratic expression to approximate the maximum of the log-likelihood function. Visually, estimation can be viewed as generating a parabola at point θ_t and jumping to the value

θ_{t+1} where the derivative of the quadratic approximation equals 0. A graphical depiction of the updating step for the variance is shown below.