What is a derivative

R. Scott McIntire

Oct, 20, 2023

1 Derivative of a Function, $f : R \to R$

The derivative of a function, $f : R \to R$ is a linear approximation to a function. By placing a straight-edge up against the graph of a function at a point we get not just any approximation but the "best" one. Meaning, we could do no better with any other linear approximation. So, the error from this approximation should be higher order than linear. More specifically, we mean:

$$f(x+h) \quad = \quad f(x) + Df(x)(h) + o(h) \tag{1}$$

Here, $Df(x)$ is the linear function that is this best approximation of f at x and $o(h)$ is a higher order error (h^2, for example) in the sense that $\lim_{h\to 0} \frac{o(h)}{h} = 0$.

Although this geometric concept has been tied to an algebraic equation, is not clear how to compute the derivative, $Df(x)$ for every x. This seems like a great deal of work and when done we have to produce a function for every x – a relatively complicated thing. It turns out that there is a *representation* result that says that all linear functions, L, $L : R \to R$ can be mapped to a scalar value. Specifically, for each linear function, L, there is a scalar, a_L, so that[1]

$$L(h) \quad = \quad a_L\, h \tag{2}$$

This says nothing more than a linear function, L, can be completely determined by its slope, a_L.

For the linear function $Df(x)$ the associated scalar value we label $f'(x)$ and call it –as an abuse of language – the derivative of f at x. To compute this scalar value we use the definition, (1), at a point x:

$$f'(x)h \quad = \quad Df(x)(h) = f(x+h) - f(x) + o(h) \tag{3}$$

So that[2]

$$\begin{aligned} f'(x) &= \lim_{h\to 0} \frac{f(x+h)-f(x)}{h} + \frac{o(h)}{h} \\ &= \lim_{h\to 0} \frac{f(x+h)-f(x)}{h} + 0 \\ &= \lim_{h\to 0} \frac{f(x+h)-f(x)}{h} \end{aligned} \tag{4}$$

[1] We label the scalar with the index L to indicate the dependency on L.

[2] Geometrically, we are approximating the tangent line with lines passing through approximating cords to the function, f.

Although this seems like a great deal of work – for a given function we need to compute this limit for each value of x in its domain – it turns out that for many functions there is a *formula* that works for the given function *for all* points in its domain. For example, the derivative of the function $f(x) = x^2$ at any point can be computed from the formula: $f'(x) = 2\,x$. In order to avoid the limit process one can build a *calculus* of formulas in the following way:

1. Find formulas for a base collection of functions: polynomials, trig, and log functions, etc.
2. Create formulas for computing the derivative of functions formed in certain composite ways: sums, products, ratios, powers, and composition of functions.

1.1 Application to Optimization

At a local maximum or minimum of a function, x, the tangent line to the graph of the function at x would be flat. That is the linear function, $Df(x)$, should be the zero function; and therefore, $f'(x)$, the corresponding slope of the graph of this linear function should be 0. So, potentially, we have a practical way of finding maxima or minima. This makes sense, to find the maxima or minima of a function over an interval, $[a, b]$, examine all of the local maxima or minima *plus* the values of the function at a and b. This should be a relatively small list. We find the local maxima and minima by solving for all points x such that $f'(x) = 0$.

Example: Find the point x^* which maximizes the function $f(x) = x(12 - 2x)^2$ over the interval $[0, 6]$. We find all points x such that $f'(x) = 0$. It turns out that the $f'(x) = 144 - 96x + 12x^2$. Solving for x, $144 - 96x + 12x^2 = 0$ (same as $12 - 8x + x^2 = 0$) we have two solutions: $\{2, 6\}$. The "x" values that are candidates which produce the maximum value of the function f are then: $\{0, 2, 6\}$ – the union of the local maxima and the boundary points. Now it is a simple matter to walk this list and find an x which produces the maximum value of f. In this case, the "x" which produces the maximum value is $x^* = 2$.

We did not show how to compute the derivative, this comes from following the two step plan above. Is the example of any practical value? Well, suppose you wanted to take a square of sheet metal 12 inches on a side, and cut out a square in each corner so that after folding the sides you made a box with the most volume. Then you are really trying to find the length x of the square to cut out so that the volume of the resulting box: $f(x) = x(12 - 2x)^2$ is maximized – the problem above.

2 Extension to $f : R^n \to R$

$$f(\mathbf{x} + \mathbf{h}) \quad = \quad f(\mathbf{x}) + Df(\mathbf{x})(\mathbf{h}) + o(\mathbf{h}) \tag{5}$$

Here, $Df(\mathbf{x})$ is a linear function from R^n to R and $o(\mathbf{h})$ is the higher order error term, higher order in the sense that $\lim_{\mathbf{h}\to\mathbf{0}} \frac{o(\mathbf{h})}{\|\mathbf{h}\|} \to 0$.

Again, the function $Df(\mathbf{x})$ is a rather abstract and difficult to work with. There is as in the scalar case a *representation* result: Any linear function, L, from R^n to R can be represented by a vector, $\mathbf{a}_L$:

$$L(\mathbf{h}) \quad = \quad \mathbf{a}_L \cdot \mathbf{h} \tag{6}$$

We denote the vector that represents the linear function $Df(\mathbf{x})$ by $\nabla f(\mathbf{x})$. Equation (5) becomes:

$$f(\mathbf{x}+\mathbf{h}) \quad = \quad f(\mathbf{x}) + \nabla f(\mathbf{x}) \cdot \mathbf{h} + o(\mathbf{h}) \tag{7}$$

As in the scalar case, we use equation (5) to compute the components of the vector $\nabla f(\mathbf{x})$ once we chose a coordinate system. To this end let $\mathbf{e}_i$ be a basis for R^n. The i^{th} component in this coordinate system of this vector is

$$\nabla f(\mathbf{x})_i \quad = \quad \nabla f(\mathbf{x}) \cdot \mathbf{e_i} = \lim_{h \to 0} \left(\frac{f(\mathbf{x} + h\mathbf{e_i}) - f(\mathbf{x})}{h} + \frac{o(h\mathbf{e_i})}{\|h\mathbf{e}_i\|} \right) \tag{8}$$

If $\tilde{f}$ is the function that takes an n tuple of the components of a vector, $\mathbf{x}$, giving the value of $f(\mathbf{x})$ in such a way that $\tilde{f}(x_1, x_2, \ldots, x_n) = f(\mathbf{x})$ *and* $f(\mathbf{x} + h\mathbf{e_i}) = \tilde{f}(x_1, \ldots, x_i + h, \ldots, x_n)$, then equation, (8) yields:

$$\begin{aligned} \nabla f(\mathbf{x})_i \quad &= \quad \lim_{h \to 0} \left(\frac{\tilde{f}(x_1, \ldots, x_i + h, \ldots, x_n) - \tilde{f}(x_1, \ldots, x_i, \ldots, x_n)}{h} \right) + \lim_{h \to 0} \frac{o(h\mathbf{e_i})}{\|h\mathbf{e}_i\|} \\ &= \quad \frac{\partial \tilde{f}}{\partial x_i}(x_1, x_2, \ldots, x_n) + 0 \\ &= \quad \frac{\partial \tilde{f}}{\partial x_i}(x_1, x_2, \ldots, x_n) \end{aligned} \tag{9}$$

Here, we are using the notional $\frac{\partial \tilde{f}}{\partial x_i}$ to "take a derivative" in one direction. That is, we leave all of the other "slots" of $\tilde{f}$ constant and just compute the change in one slot, x_i.[3]

Note, different choices of coordinate systems will yield different values for the components of f. It turns out that in two dimensions if one uses polar coordinates and the function $\tilde{f}$ is written using these coordinates that the components of the gradient of f are: $[\frac{1}{r}\frac{\partial \tilde{f}}{\partial \theta}, \frac{\partial \tilde{f}}{\partial r}]$.

Note: The "local" direction of maximum increase at a point of the $f(\mathbf{x})$ is given by the gradient of f. To see this notice that we can discuss direction of a vector in a unique sense by normalizing it to be of unit length. That is, we can speak of a vector's direction as its unit vector.

Question: What is the direction to change a given value x so as to increase f as much as possible? So, find a direction, $\mathbf{w}$ (unit vector), so that for a vector of small enough length h, $f(\mathbf{x} + h\mathbf{w})$ is larger than all other vectors of the same length. We claim that the direction of maximum increase (locally) is $\nabla f(\mathbf{x})$ (normalized).

$$f(\mathbf{x} + h\mathbf{w}) - f(\mathbf{x}) \quad = \quad \nabla f(\mathbf{x}) \cdot h\mathbf{w} + o(h\mathbf{w}) \tag{10}$$

But the Cauchy-Schwartz inequality says: $|\nabla f(\mathbf{x}) \cdot h\mathbf{w}| \leq \|\nabla f(\mathbf{x})\| |h|$. The maximum is only achieved if $\nabla f(\mathbf{x})$ and $\mathbf{w}$ are in the same direction. That is $\mathbf{w} = \frac{\nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|}$.

Therefore, to increase f in the fast possible way (locally) move in the direction of $\nabla f(\mathbf{x})$; conversely, to decrease f in the fastest possible way (locally) move in the direction of $-\nabla f(\mathbf{x})$

[3]What does x have to do with the i^{th} slot of $\tilde{f}$. Actually nothing, as we can put anything in that slot, x, y, z, 5, etc. Keep this in mind, as two sections from now we will see the notation: $\frac{\partial F}{\partial y'}$. In this case, it simply means take a derivative in the third slot of the function F.

2.1 Application to Optimization and Data Science

By the same reasoning of the last section in order to find a maximum or minimum one should find all of the places where a function is flat; that is, where the linear approximation is zero. But this is the place where the gradient is zero – the zero linear function is represented by the $\mathbf{0}$ vector. Therefore, find all points $\mathbf{x}$ such that $\nabla f(\mathbf{x}) = \mathbf{0}$. Then compare these points with any boundary points to see which is the largest or smallest.

This might be hard in general – true also in the scalar case from the previous section. How might one go about finding the minimum value of a function, $f(\mathbf{x})$ when solving for $\nabla f(\mathbf{x}) = \mathbf{0}$ is too hard? We know that at any point x, the "local" direction to move to decrease f as much as possible is to move along the direction of the negative gradient. This suggests a scheme for finding the minimum.

$\mathbf{x}_0 \leftarrow \mathbf{x0}$
$n \leftarrow 0$
while $\|\mathbf{x_n} - \mathbf{x_{n-1}}\| > \epsilon$ *and* $n \leq N$ **do**
  $n \leftarrow n + 1$
  $\mathbf{x}_n \leftarrow \mathbf{x}_{n-1} - \eta \nabla f(\mathbf{x}_{n-1})$
end while

if $n > N$ **then**
  ERROR : Did not converge within N steps.
else
  $\mathbf{x}^* \leftarrow \mathbf{x}_n$
end if

This algorithm depends on the parameters:

x0 : An Initial guess;

N**:** The maximum number of iterations allowed;

ϵ**:** The error tolerance for convergence;

η**:** The "learning" parameter.

As it is analytically difficult to compute the gradient for real world problems, schemes like the one above are used to find solutions to sophisticated constrained optimization problems. They are also used to train Artificial Deep Neural Networks.

3 Extension to $f : R^n \rightarrow R^m$

$$\mathbf{f}(\mathbf{x} + \mathbf{h}) \quad = \quad \mathbf{f}(\mathbf{x}) + \mathbf{Df}(\mathbf{x})(\mathbf{h}) + \mathbf{o}(\mathbf{h}) \tag{11}$$

Here $\mathbf{Df}(\mathbf{x})$ is a linear function from R^n to R^m and $\mathbf{o}(\mathbf{h})$ is a higher order error term, higher order in the sense that $\lim_{\mathbf{h}\to\mathbf{0}} \frac{\|\mathbf{o}(\mathbf{h})\|}{\|\mathbf{h}\|} \to 0$.

There is a *representation* result that says that any linear map from R^n to R^m can be represented by a matrix once coordinates have been chosen.

Using (11) one can derive a representation of the derivative using functions commensurate with a choice of basis vectors, $\mathbf{e}_i \quad i \in [1, n]$. Given this choice of basis, let $\tilde{f}_j$ be the coordinate functions for $\mathbf{f}$. We claim without proof that the matrix representing the derivative is given by $[\mathbf{Df}(\mathbf{x})]$, whose $i^{\text{th}}, j^{\text{th}}$ entries are:

$$[\mathbf{Df}(x)]_{i,j} \quad = \quad \frac{\partial \tilde{\mathbf{f}}_j}{\partial x_i} \tag{12}$$

Why do we study matrices? One answer is that they represent the linear mappings from $R^n \to R^m$ and so if we want to use calculus[4] on the functions from R^n to R^m we need to understand the linear functions, on these spaces and their representations – matrices.

4 Extension to $f : R^\infty \to R$

Consider the function $J(y) = \int_a^b F(x, y(x), y'(x))\, dx$. J is a function from a *function space*, H, to R; that is, $J : H \to R$. This is similar to functions, $f : R^n \to R$, but where n is infinity. One can show that, for instance, the functions $\{\sin(n\, x)\}_{n=1}^{\infty}$ are linearly independent; therefore, the space H is infinite dimensional.[5]

We define the derivative of J as before, the best linear approximation:[6]

$$J(y + h) = J(y) + DJ(y)(h) + o(h) \tag{13}$$

Here $DJ(y) : H \to R$; that is, it is a linear function from function space to R and $o(h)$ is a higher order error term that takes a function, h to a number such that $\lim_{h\to 0} \frac{o(h)}{\|h\|} = 0$. However, its not clear what the norm of the function h is. One can define an inner product (generalization of a dot product) on the space of functions which is a generalization of the dot product for vectors in a finite dimensional space. In turn, this inner product defines a norm on the space of functions, H. We define the following inner product function, $<\cdot, \cdot>$ by:

$$< f, g > \quad \equiv \quad \int_a^b f(x) g(x)\, dx \tag{14}$$

From this, just as with dot products, we can define the norm, or length of a function, using this inner product:

$$\|f\| \quad = \quad \sqrt{< f, f >} \tag{15}$$

[4] We can use calculus to *locally* approximate a given non-linear function, which can be useful in itself. We also know from the last section that we can iterate on the derivative and solve/understand *global* issues of such a non-linear function.

[5] The notation $\{\sin(nx)\}_{n=1}^{\infty}$ is sloppy, we should instead write: $\{f_n\}_{n=1}^{\infty}$ where f_n is the function defined by $f_n(x) = \sin(nx)$.

[6] We write $o(h)$ as the left over higher order terms and do not create new function names, even though from equation to equation, the $o(h)$ functions may change.

As a brief aside we discuss the inner product.

This definition of an inner product seems strange and seems very much removed from the dot product in R^n. We now show that it follows by looking at finite dimensional approximations to H. One way to get a finite version of a continuous function, f, on the interval $[a, b]$ is to sample it at, say, n discrete points on $[a, b]$, $\{x_i\}_{i=1}^n$, which are evenly spaced points over $[a, b]$. We can think of the following vectors as approximations to f and g:

$$\begin{aligned} \widetilde{f} &= [f(x_1), f(x_2), \ldots, f(x_n)] \\ \widetilde{g} &= [g(x_1), g(x_2), \ldots, g(x_n)] \end{aligned}$$

The natural inner product of these vectors would be the usual dot product. As we increase the number of points the approximations get closer to the original continuous functions. However, at the same time, due to the increase in the number of points, the induced norm of the approximations gets larger and larger. We would like an inner product that converged to something finite as the number approximations convert to the original function. One way to do this is to normalize these dot products so that they would remain finite. The simplest thing to do is to divide by the number of points in the discretization. That is, we define an inner product $\ll \cdot, \cdot \gg$ for these vectors as:

$$\ll \widetilde{f}, \widetilde{g} \gg \equiv \frac{1}{n} \sum_{i=1}^{n} f(x_i) g(x_i) \tag{16}$$

Taking the limit of the above we can define an inner product for the original functions f and g. That is, for our function space, define an inner product, $\ll \cdot, \cdot \gg$ by:

$$\begin{aligned} \ll f, g \gg &\equiv \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} f(x_i) g(x_i) \\ &= \frac{1}{b-a} \lim_{n \to \infty} \sum_{i=1}^{n} f(x_i) g(x_i) \frac{b-a}{n} \\ &= \frac{1}{b-a} \int_a^b f(x) g(x)\, dx \end{aligned}$$

Now, equation (14) should make sense as an analog of a dot product for vectors. It is a different inner product than the one we just defined – but differs only by a constant multiple.

Going back to the problem at hand we note that there is a *representation* result for linear functions on our function space, H – technically, for a Hilbert space. All linear functions[7], also called functionals, which map the function space to R can be represented by a function in the function space – a generalization of R^n. In R^n we found functionals could be represented by a *vector*. Analogously, for every linear functional, L, in our function space there is a *function*, a_L, such that:

$$L(f) = < a_L, f > \tag{17}$$

We denote by $\nabla J(y)$ the function that represents the derivative, $DJ(f)$. That is

$$DJ(y)(h) = < \nabla J(y), h > \tag{18}$$

[7]There are some details we won't bother with.

Remember, the representation, $\nabla J(y)$, gives us a concrete thing with which we can compute. The derivative is somewhat abstract, it is the representative that really gives us something to work with in terms of computations. This was the case with the derivative of a scalar function as well. There, we represented the derivative. a linear function, as a scalar – the slope of the linear function. The slope, a number, is much easier to use in computations.

We now try to compute $\nabla J(y)$. To recap, we know

$$DJ(y)(h) \quad = \quad J(y+h) - J(y) + o(h) \tag{19}$$

Since $DJ(y)(h) = <\nabla J(y), h>$, we have:

$$<\nabla J(y), h> \quad = \quad J(y+h) - J(y) + o(h) \tag{20}$$

To find the function $\nabla J(y)$ we need to re-write the RHS of the last equation so that it becomes:

$$<\nabla J(y), h> \quad = \quad <w, h> + o(h) \tag{21}$$

Then we will have identified our function $\nabla J(y)$, it will be w. Why? Because equation (21) implies that w and $\nabla J(y)$ must be the same function.

How do we get h to come out of the inside of the functional $J(y+h)$? Answer, use Taylor's expansion on F. Continuing, substituting the definition of J in equation (20) and using Taylor's expansion on F we have:[8]

$$\begin{aligned} <\nabla J(y), h> &= \int_a^b F(x, (y+h)(x), (y'+h')(x))\, dx - \int_a^b F(x, y(x), y'(x))\, dx + o(h) \\ &= \int_a^b \frac{\partial F}{\partial y}(x, y(x), y'(x))h(x)\, dx + \int_a^b \frac{\partial F}{\partial y'}(x, y(x), y'(x))h'(x)\, dx + o(h) \end{aligned} \tag{22}$$

Here, we used Taylor's expansion on F:

$$F(x, y+h, y+h') \quad = \quad F(x, y, y') + \frac{\partial F}{\partial y}(x, y, y')h + \frac{\partial F}{\partial y'}(x, y, y')h' + o(h) + o(h') \tag{23}$$

If equation (22) looked instead like

$$<\nabla J(y), h> \quad = \quad \int_a^b \frac{\partial F}{\partial y}(x, y(x), y'(x))h(x)\, dx + o(h) \tag{24}$$

then since (24) must be true for all h in the function space we must have that $\nabla J(f) = \frac{\partial F}{\partial y}(x, y(x), y'(x))$.

Going back to the equation (22), we need to replace the term involving h' with one involving h. Using integration by parts, we can rewrite the second term in (22):

$$\begin{aligned} \int_a^b \frac{\partial F}{\partial y'}(x, y(x), y'(x))h'(x)\, dx &= \frac{\partial F}{\partial y'}(x, y(x), y'(x))h'(x)\Big|_{x=a}^{x=b} \\ &- \int_a^b \frac{d}{dx}\left[\frac{\partial F}{\partial y'}(x, y(x), y'(x))\right] h(x) \end{aligned} \tag{25}$$

[8] This is just the definition of the derivative of F: $F(x+a, f+h, f+h') - F(x, f, f') = \nabla F(x, f, f')\cdot[a, h, h'] + o([a, h, h'])$. In our case, $a = 0$.

If we restrict our function space, H, to the functions which have the restriction that $h(a) = h(b) = 0$, then the previous equation becomes:

$$\int_a^b \frac{\partial F}{\partial y'}(x, y(x), y'(x))h'(x)\,dx \quad = \quad -\int_a^b \frac{d}{dx}\left[\frac{\partial F}{\partial y'}(x, y(x), y'(x))\right] h(x) \tag{26}$$

we claim that that with this restriction of function, $DJ(y)$ is still unique. With this choice of h, equation (22) becomes:

$$\begin{aligned} <\nabla J(y), h> &= \int_a^b F(x, (y+h)(x), (y'+h')(x)\,dx - \int_a^b F(x, y(x), y'(x))\,dx + o(h) \\ &= \int_a^b \frac{\partial F}{\partial y}(x, y(x), y'(x))h(x)\,dx + \int_a^b \frac{\partial F}{\partial y'}(x, y(x), y'(x))h'(x)\,dx + o(h) \\ &= \int_a^b \left[\frac{\partial F}{\partial y}(x, y(x), y'(x))h(x) - \frac{d}{dx}\frac{\partial F}{\partial y'}(x, y(x), y'(x))\right] h(x)\,dx + o(h) \end{aligned} \tag{27}$$

Using (21) we find that

$$(\nabla J(y))(x) \quad = \quad \frac{\partial F}{\partial y}(x, y(x), y'(x)) - \frac{d}{dx}\left[\frac{\partial F}{\partial y'}(x, y(x), y'(x))\right] \tag{28}$$

We ignored a few technical details – when h is small in our distance measure is h' small? However, the point is that as far away as we are from the scalar case, the essence of the calculations are the same – identify the linear thing acting on a small change to the input, h, and show that what remains are "higher order" terms.

4.1 Application to Physics

Certain systems in physics have the property that the dynamics of the system are governed by minimizing (or rendering stationary – hitting a flat spot) a certain functional. In this case, the derivative should be "zero"; meaning, in this case, the zero function.[9] From this condition we can determine the dynamics of the system. That is for certain physical systems, the path trajectory of a particle, $y(x)$, is determined by find the function y where the gradient, $\nabla J(y)$, is the zero function. Using (28), $\nabla J(y) \equiv 0$ means:

$$\frac{\partial F}{\partial y}(x, y(x), y'(x)) - \frac{d}{dx}\left[\frac{\partial F}{\partial y'}(x, y(x), y'(x))\right] \quad = \quad 0 \tag{29}$$

Example:

The dynamics of a mass is determined by finding a trajectory where the Lagrangian is locally flat – a place where the derivative of the Lagrangian is zero. The Lagrangian of a mass hanging from a spring is[10]

$$J(y) \quad = \quad \int_0^T \overbrace{\frac{1}{2}ky(t)^2}^{\text{Potential Energy}} \quad - \overbrace{\frac{1}{2}my'(t)^2}^{\text{Kinetic Energy}} \quad dt \tag{30}$$

[9] The zero function is the function which sends all inputs to the value 0.

[10] We replace the "dummy" variable x in our equations with t.

In this case, $F(t, y, y') = \frac{1}{2}ky^2 - \frac{1}{2}my'^2$. Therefore, we need to find a function, y, such that $(\nabla J(y))(t) = 0 \quad \forall t \in [0, T]$. This equation becomes (using the dummy variable t):

$$\frac{\partial F}{\partial y}(t, y(t), y'(t)) - \frac{d}{dt}\left[\frac{\partial F}{\partial y'}(t, y(t), y'(t))\right] = 0 \tag{31}$$

The partial derivatives of F are computed as:

$$\frac{\partial F}{\partial y} = ky \tag{32}$$

$$\frac{\partial F}{\partial y'} = -my' \tag{33}$$

Substituting into (31) we have:

$$ky(t) - \frac{d}{dt}\left[-my'(t)\right] = 0 \tag{34}$$

Or,

$$my''(t) + ky(t) = 0 \tag{35}$$

This is a second order linear differential equation. Adding initial conditions, one can solve for the particle dynamics.

The general form of a solution to this differential equation is:

$$y(t) = a_1 \cos(\sqrt{k/m}\, t) + a_2 \sin(\sqrt{k/m}\, t) \tag{36}$$

Given the initial position and velocity we can determine a_1 and a_2.