Math 1c TA: Padraic Bartlett

Recitation 3: Derivatives: Tools and Applications

Week 3 *Caltech 2011*

1 Random Question

Hopefully, in class or Apostol, you've seen a function $f : \mathbb{R}^2 \to \mathbb{R}$ such that its mixed partials $\frac{\partial^2 f}{\partial x \partial y}$ and $\frac{\partial^2 f}{\partial y \partial x}$ exist on all of $\mathbb{R}^2$, but at some point – say (0,0) – these two partials disagree, i.e.

$$\left.\frac{\partial^2 f}{\partial x \partial y}\right|_{(0,0)} \neq \left.\frac{\partial^2 f}{\partial y \partial x}\right|_{(0,0)}.$$

(One such function is $f(x,y) = \frac{xy^3 - x^3y}{x^2+y^2}$ for $(x,y) \neq (0,0)$ and 0 for $(x,y) = (0,0)$, in case you haven't seen this.)

Can you find a version of this function where all of the partials of higher orders exist, but are unequal? In other words,can you find a function $f : \mathbb{R}^2 \to \mathbb{R}$ such that it has its mixed partials $\frac{\partial^{2n} f}{\partial^n x \partial^n y}$ and $\frac{\partial^{2n} f}{\partial^n y \partial^n x}$ on all of $\mathbb{R}^2$, but at the origin these two mixed partials disagree?

Relevant theorems and definitions:

- A function is called $C^k(U)$, or C^k on some set U, iff all of its k-th order partial derivatives[1] are continuous on U. If the set U is unstated, assume that it's the entire domain of definition for your function (so, usually $\mathbb{R}^n$.)
- There is a theorem that says that if a function is differentiable, it must be continuous (this is not hard to show, if you want!)
- As well, there is a theorem that says if a function is C^k, it doesn't matter in which order you take its partials – i.e. $\frac{\partial^k f}{\partial x_{i_1} \ldots \partial x_{i_k}} = \frac{\partial^k f}{\partial x_{i'_1} \ldots \partial x_{i'_k}}$, for any two orderings $i_1, \ldots i_k$ and $i'_1, \ldots i'_k$ of the same set of numbers.

So, rephrased in this way, you want a function that is C^{2n-1}, not in C^{2n}, such that $\frac{\partial^{2n} f}{\partial^n x \partial^n y} \neq \frac{\partial^{2n} f}{\partial^n y \partial^n x}$ at the origin.

2 The Derivative, Take 2

2.1 Derivatives of functions from $\mathbb{R}^n \to \mathbb{R}^m$

Due to the sheer bulk of the material we have to cover in this lecture, we're going to kind of skip around a bit in our talk; our theme today is the derivative, but we're covering three

[1] A **k-th order partial derivative** of a function f is simply the function resulting from taking k partial derivatives of f. E.g: $\frac{\partial^2 f}{\partial x \partial y}$ is a second-order partial derivative.

slightly disjoint subjects here – the derivative in higher dimensions, the chain rule, and how to use derivatives to find extrema.

So: we start by first definining the derivative of a function from $\mathbb{R}^n$ to $\mathbb{R}^m$. From last lecture, recall that for the simpler case of a function $f : \mathbb{R}^n \to \mathbb{R}^1$, we said that

1. The partial derivative of f at some point $\mathbf{a}$ with respect to its i-th coördinate is the derivative of f at $\mathbf{a}$ if you hold all other coördinates constant and treat it as a single-variable function: equivalently, it is the limit

$$\lim_{h \to 0} \frac{f(\mathbf{a} + h \cdot \mathbf{e}_i) - f(\mathbf{a})}{h}.$$

2. The total derivative of f at some point $\mathbf{a}$ is a linear map $T_{\mathbf{a}} : \mathbb{R}^n \to \mathbb{R}^1$ – in other words, a $1 \times n$ matrix – such that it was a "linear approximation" of f close to $\mathbf{a}$. I.e.

$$\lim_{||\mathbf{h}|| \to 0} \frac{f(\mathbf{a} + \mathbf{h}) - f(\mathbf{a}) - T_{\mathbf{a}} \cdot \mathbf{h}}{||\mathbf{h}||} = 0.$$

Our definitions for a higher-dimensional function are almost identical. Specifically: in coming up with our definition of a partial derivative of f, we simply studied the ratio of (small changes in $f(x)$ in the i-th coördinate)/(small changes in the i-th coördinate). For a higher dimensional function, we can do precisely the same thing, and write that for a function $f : \mathbb{R}^n \to \mathbb{R}^m$, the i-th partial derivative of f is just

$$\lim_{h \to 0} \frac{f(\mathbf{a} + h \cdot \mathbf{e}_i) - f(\mathbf{a})}{h},$$

where the result of this limit will be **vector-valued**. Equivalently, we can also write that

$$\frac{\partial f}{\partial x_i} = \left(\frac{\partial f_1}{\partial x_i}, \dots \frac{\partial f_m}{\partial x_i} \right),$$

where each f_i is the function $\mathbb{R}^n \to \mathbb{R}$ given by f's i-th coördinate.

Similarly, for the total derivative, we want T_a to again be a "linear approximation" to f. However, this time, this means that we want T_a to be a function $\mathbb{R}^n \to \mathbb{R}^m$ – i.e. a $m \times n$ matrix – because we want it to have the same inputs and outputs as f. Therefore, to say that any such T_a is a close approximation to f around $\mathbf{a}$, we need to require that

$$\lim_{||\mathbf{h}|| \to 0} \frac{||f(\mathbf{a} + \mathbf{h}) - f(\mathbf{a}) - T_{\mathbf{a}} \cdot \mathbf{h}||}{||\mathbf{h}||} = 0,$$

where we have to take the magnitude of the numerator in this limit because it is now **vector-valued**, just as before.

Again, as before, we have the following two theorems relating the partial and total derivatives:

Theorem 1 *If a function $f : \mathbb{R}^n \to \mathbb{R}^m$ has a total derivative $T_{\mathbf{a}}$ at some point, then this total derivative is given by the matrix of partial derivatives $D(f)\Big|_{\mathbf{a}}$. In other words, if $T_{\mathbf{a}}$ exists, we have*

$$T_{\mathbf{a}} = D(f)\Big|_{\mathbf{a}} := \begin{bmatrix} \frac{\partial f_1}{\partial x_1}\Big|_{\mathbf{a}} & \cdots & \frac{\partial f_1}{\partial x_n}\Big|_{\mathbf{a}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}\Big|_{\mathbf{a}} & \cdots & \frac{\partial f_m}{\partial x_n}\Big|_{\mathbf{a}} \end{bmatrix}$$

Theorem 2 *If all of its first-order partials of a function $f : \mathbb{R}^n \to \mathbb{R}^m$ exist and are continuous in a small ball around some point $\mathbf{a}$ – i.e. f is C^1 in a ball around $\mathbf{a}$ – then $T_{\mathbf{a}}$ exists.*

To illustrate how these derivatives are calculated, we do one quick example:

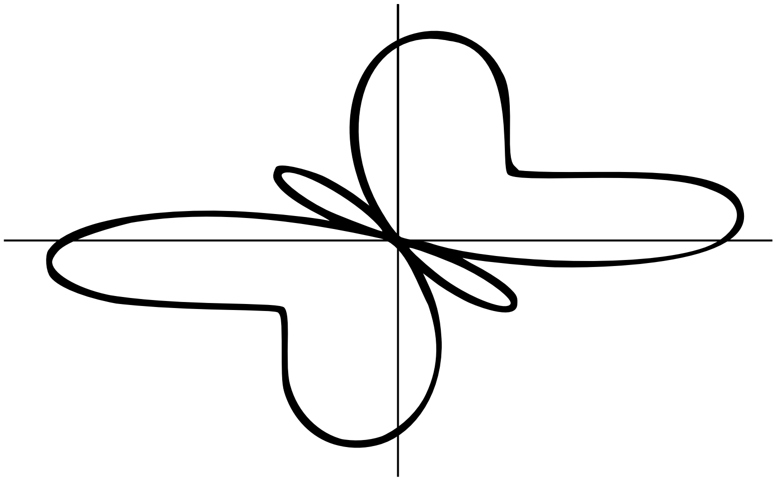
Example. Find the total derivative of the function $f : \mathbb{R} \to \mathbb{R}^2$ given by

$$f(t) = (\cos(t) \cdot (1 + \cos(4t) + \sin(2t)), \sin(t) \cdot (1 + \cos(4t) + \sin(2t)))\,.$$

How is this function changing at $t = 0$?

Solution. Before we start, notice two things about this function:

- Because it's a function $\mathbb{R} \to \mathbb{R}^n$, it's a **parametric curve** – i.e. its graph is going to look like some sort of a curve drawn in n-space (where n is 2 in our case.)
- In specific, it's a parametric curve of the form $(\cos(t) \cdot r(t), \sin(t) \cdot r(t)$, where we define $r(t)$ as the function $(1 + \cos(4t) + \sin(2t))$. So, recalling how polar coördinates work, we can see that our function is really the graph of the function $r(t)$ regarded in polar coördinates! In other words, it's the function



(which you may remember from Ma1a!)

So: to find this function's total derivative, we merely need to find $D(f)$, which we do by taking derivatives of f's first and second coördinates one at a time:

$$\begin{aligned} D(f) &= \begin{bmatrix} \frac{\partial f_1}{\partial t} \\ \frac{\partial f_2}{\partial t} \end{bmatrix} \\ &= \begin{bmatrix} -\sin(t)\cdot(1+\cos(4t)+\sin(2t)) + \cos(t)\cdot(-4\sin(4t)+2\cos(2t)) \\ \cos(t)\cdot(1+\cos(4t)+\sin(2t)) + \sin(t)\cdot(-4\sin(4t)+2\cos(2t)) \end{bmatrix}. \end{aligned}$$

As both of these partial derivatives are continuous on all of $\mathbb{R}$, we know that our function is $C^1(\mathbb{R})$ and thus that it has a total derivative on all of $\mathbb{R}$, given by the above matrix. In particular, at $t = 0$, we have that

$$T_0(f) = \begin{bmatrix} -\sin(0)\cdot(1+\cos(0)+\sin(0)) + \cos(0)\cdot(-4\sin(0)+2\cos(0)) \\ \cos(0)\cdot(1+\cos(0)+\sin(0)) + \sin(0)\cdot(-4\sin(0)+2\cos(0)) \end{bmatrix}. \quad = \begin{bmatrix} 2 \\ \cos(0)\cdot 2 \end{bmatrix},$$

and thus that at time $t = 0$, our function is increasing equally in both the x and y directions at a rate equal to twice that of t.

2.2 Tools for taking derivatives

Switching gears somewhat, we now turn from the theory of derivatives to a more practical approach – how do we calculate these things? For functions $\mathbb{R}^1 \to \mathbb{R}^1$, in particular, we had things like the product and chain rule; are there analogues for functions $\mathbb{R}^n \to \mathbb{R}^m$?

Well: yes! Kind-of. The product rule is kind of tricky, as there isn't a single well-defined notion of "product" to use. However, if we take two functions $f, g : \mathbb{R}^n \to \mathbb{R}^m$, and assume that by product we mean the **dot product**, then we have a version of the product rule, that you used on your last HW:

$$\nabla(f \cdot g)\Big|_{\mathbf{a}} = f(\mathbf{a}) \cdot (\nabla g)\Big|_{\mathbf{a}} + g(\mathbf{a}) \cdot (\nabla f)\Big|_{\mathbf{a}}.$$

The chain rule, however, is pretty clear-cut, as there's only one way to define composition! Specifically, take any function $g : \mathbb{R}^m \to \mathbb{R}^l$, and any function $f : \mathbb{R}^n \to \mathbb{R}^m$. Then, the chain rule says the following thing about the matrix of derivatives $D(g \circ f)$ of $g \circ f$ at a point $\mathbf{a} \in \mathbb{R}^n$:

$$D(g \circ f)\Big|_{\mathbf{a}} = D(g)\Big|_{f(\mathbf{a})} \cdot D(f)\Big|_{\mathbf{a}}.$$

One interesting/cautionary tale to notice from the above calculations is that the partial derivative of $g \circ f$ with respect to one variable x_i can depend on **many** of the variables and coördinates in the functions f and g!

I.e. something many first-year calculus students are tempted to do on their sets is to write

$$\frac{\partial (g \circ f)_i}{\partial x_j}\Big|_{\mathbf{a}} = \frac{\partial g_i}{\partial x_j}\Big|_{f(\mathbf{a})} \cdot \frac{\partial f_i}{\partial x_j}\Big|_{\mathbf{a}}.$$

DO NOT DO THIS. Do not do this. Do not do this. Ever. Because it is wrong. Indeed, if you expand how we've stated the chain rule above, you can see that $\frac{\partial (g \circ f)_i}{\partial x_j}\Big|_{\mathbf{a}}$ – the (i, j)-th entry in the matrix $D(g \circ f)$ – is actually equal to the i-th row of $D(g)\Big|_{f(\mathbf{a})}$ multipled by the j-th column of $D(f)\Big|_{\mathbf{a}}$ – i.e. that

$$\frac{\partial (g \circ f)_i}{\partial x_j}\Big|_{\mathbf{a}} = \left[\begin{array}{ccc} \frac{\partial g_i}{\partial x_1}\Big|_{f(\mathbf{a})} & \cdots & \frac{\partial g_i}{\partial x_m}\Big|_{f(\mathbf{a})} \end{array}\right] \cdot \left[\begin{array}{c} \frac{\partial f_1}{\partial x_j}\Big|_{\mathbf{a}} \\ \vdots \\ \frac{\partial f_m}{\partial x_j}\Big|_{\mathbf{a}} \end{array}\right].$$

Notice how this is much more complex! In particular, it means that the partials of $g \circ f$ depend on all sorts of things going on with g and f, and aren't restricted to worrying about just the one coördinate you're finding partials with respect to.

The moral here is basically if you're applying the chain rule without doing a *lot* of derivative calculations, you've almost surely messed something up. So, when in doubt, just find the matrices $D(f)$ and $D(g)$!

We work one example, to illustrate how to do these kinds of calculations:

Example. If $f(x) = (x, x^2, x^3)$ and $g(x) = \sin(xyz)$, use the chain rule to find $D(g \circ f)\Big|_{\mathbf{a}}$.

Solution. If we straightforwardly apply the chain rule, we have that

$$\begin{aligned} D(g \circ f)\Big|_{\mathbf{a}} &= D(g)\Big|_{f(\mathbf{a})} \cdot D(f)\Big|_{\mathbf{a}} \\ &= \left[\begin{array}{ccc} \frac{\partial g}{\partial x}\Big|_{f(\mathbf{a})} & \frac{\partial g}{\partial y}\Big|_{f(\mathbf{a})} & \frac{\partial g}{\partial z}\Big|_{f(\mathbf{a})} \end{array}\right] \cdot \left[\begin{array}{c} \frac{\partial f_1}{\partial x}\Big|_{\mathbf{a}} \\ \frac{\partial f_2}{\partial x}\Big|_{\mathbf{a}} \\ \frac{\partial f_3}{\partial x}\Big|_{\mathbf{a}} \end{array}\right] \\ &= \left[\begin{array}{ccc} yz \cdot \cos(xyz)\Big|_{(a,a^2,a^3)} & xz \cdot \cos(xyz)\Big|_{(a,a^2,a^3)} & xy \cdot \cos(xyz)\Big|_{(a,a^2,a^3)} \end{array}\right] \cdot \left[\begin{array}{c} 1\Big|_{\mathbf{a}} \\ 2x\Big|_{\mathbf{a}} \\ 3x^2\Big|_{\mathbf{a}} \end{array}\right] \\ &= \left[\begin{array}{ccc} a^5 \cdot \cos(a^6) & a^4 \cdot \cos(a^6) & a^3 \cdot \cos(a^6) \end{array}\right] \cdot \left[\begin{array}{c} 1 \\ 2a \\ 3a^2 \end{array}\right] \\ &= a^5 \cdot \cos(a^6) + 2a^5 \cdot \cos(a^6) + 3a^5 \cdot \cos(a^6) \\ &= 6a^5 \cdot \cos(a^6). \end{aligned}$$

As a quick sanity check, we can verify that this makes sense by just looking at the function $g \circ f$ directly: $g \circ f(x) = \sin(x \cdot x^2 \cdot x^3) = \sin(x^6)$, and therefore $(g \circ f)'(a) = 6a^5 \cdot \cos(a^6)$ by applying the one-dimensional version of the chain rule.

2.3 Applications: Extrema

Changing gears once again, we now turn to one of the classical applications of the derivative: finding extremal points!

Specifically, we have the following definitions, for a function $f : \mathbb{R}^n \to \mathbb{R}$:

Definition. A point $\mathbf{a} \in \mathbb{R}^n$ is called a **local maxima** of a function $f : \mathbb{R}^n \to \mathbb{R}$ iff there is some small value r such that for any point $\mathbf{x}$ in $B_{\mathbf{a}}(r)$ not equal to $\mathbf{a}$, we have $f(\mathbf{x}) \leq f(\mathbf{a})$.

A similar definition holds for local minima.

So: how can we use the derivative to find such local maxima? Well, it's clear that (if our function is differentiable in a neighborhood around this point) that no matter how we move to leave this point, our function must not increase – in other words, for any direction $\mathbf{v} \in \mathbb{R}^n$, the directional derivative $f'(\mathbf{a}, \mathbf{v})$ must be ≤ 0. But this means that in fact all of the directional derivatives must be **equal** to 0!, because if $f'(\mathbf{a}, \mathbf{v})$ was < 0, then $f'(\mathbf{a}, -\mathbf{v})$ would be > 0.

This motivates the following definitions, and basically proves the following theorem:

Definition. A point $\mathbf{a}$ is called a stationary point of some function $f : \mathbb{R}^n \to \mathbb{R}$ iff $\nabla(f)\Big|_{\mathbf{a}} = (0, \ldots, 0)$. A point $\mathbf{a}$ is called a **critical point** iff it is a stationary point or f is not differentiable in any neighborhood of $\mathbf{a}$.

Theorem 3 *A function $f : \mathbb{R}^n \to \mathbb{R}$ attains its local maxima and minima only at critical points.*

However, it bears noting that not every critical or stationary point is a local maxima or minima! A trivial example would be the function $f(x, y) = x^2 - y^2$: the origin is a stationary point, yet neither a local minima or maxima (as $f(0, \epsilon) < 0 < f(\epsilon, 0)$, and thus there are positive and negative values of f attained in any ball around the origin, where it is 0.)

How can we tell which stationary points do what? Well, in one-variable calculus, we used the idea of the "second derivative" to determine what was going on! In specific, we knew that if the second derivative of a function f at some point a was negative, then tiny increases in our variable at that point would cause the first derivative to decrease, and tiny decreases in our variable at that point would cause the negative of the first derivative to increase – i.e. cause the first derivative to decrease, and therefore make the function itself decrease! Therefore, the second derivative being negative at a stationary point implied that that point was a local maxima.

In higher dimensions, things are tricker – we no longer have this idea of a "single" second derivative, but instead have many different second derivatives, like $\frac{\partial^2 f}{\partial x \partial y}$ and $\frac{\partial^2 f}{\partial z^2}$. Yet, we can still use the same ideas as before to figure out what's going on!

In particular, in one dimension, we said that we wanted tiny positive changes of our variables to make the first functions decrease. In other words, given any of the partials $\frac{\partial f}{\partial x_i}$, we want any positive changes in the direction of this partial to make our function decrease – i.e. we want the directional derivative of $\frac{\partial f}{\partial x_i}$ to be negative in any direction $\mathbf{v}$, where all of the coördinates of $\mathbf{v}$ are positive. (Positivity here stems from the same reason that in one dimension, we have that the first derivative is increasing for all of the points to the left of a maxima and decreasing for all of the points to the right of a maxima.)

So: this condition, if we write it out, is just asking that for every i and nonzero $\mathbf{v}$, that

$$\left(\frac{\partial^2 f}{\partial x_1 \partial x_i}, \frac{\partial^2 f}{\partial x_2 \partial x_i}, \ldots \frac{\partial^2 f}{\partial x_n \partial x_i}\right) \cdot (v_1^2, v_2^2, \ldots v_n^2)$$

is negative. If you choose to write this out as a matrix, this actually becomes the claim that for any $\mathbf{v}$, we have

$$\mathbf{v}^T \cdot \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{bmatrix} \cdot \mathbf{v} < 0.$$

From linear algebra, you may hopefully remember that this condition is called being **negative-definite**, and is equivalent to having all n of your eigenvalues existing and being negative.

So! Just a bit more complicated than the 1-dimensional case :) But doable! We restate what we just (very loosely) discussed above here in the following theorem and definition:

Definition. The Hessian of a function $f : \mathbb{R}^n \to \mathbb{R}$, $H(f)$, is the following $n \times n$ matrix:

$$\begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{bmatrix}$$

Theorem 4 *A function $f : \mathbb{R}^n \to \mathbb{R}$ has a local maxima at a stationary point* $\mathbf{a}$ *if all of its second-order partials exist and are continuous in a neighborhood of* $\mathbf{a}$*, and the Hesssian of f is negative-definite at* $\mathbf{a}$*. Similarly, it has a local minima if the Hessian is positive-definite at* $\mathbf{a}$*; as well, if it has both positive and negative eigenvalues, it has a saddle point at* $\mathbf{a}$ *(i.e. there are directions one can go in to either decrease or increase your function, as you want.)*

A quick example, to illustrate how this gets used:

Example. For $f(x, y) = x^2 + y^2, g(x, y) = -x^2 - y^2$, and $h(x, y) = x2 - y^2$, find local minima and maxima.

Solution. First, by taking partials, it is clear that the only point at which the gradient of these functions is $(0, 0)$ is the origin. There, we have that

$$H(f)\Big|_{(0,0)} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, H(g)\Big|_{(0,0)} = \begin{bmatrix} -2 & 0 \\ 0 & -2 \end{bmatrix}, H(h)\Big|_{(0,0)} = \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix},$$

and thus that f is positive-definite at (0,0), g is negative-definite at (0,0), and h is neither at (0,0) by examining the eigenvalues. Thus f has a local minima at $(0, 0)$, g has a local maxima at (0,0), and h has a saddle point at (0,0).