MA 1C RECITATION 04/16/15

1. Recall: General Multivariable Derivatives

The derivative of a general function $f : \mathbf{R}^n \to \mathbf{R}^m$ has the same definition as for scalar-valued functions, e.g. we have the directional derivative

$$\mathbf{f'}(\mathbf{a};\mathbf{u}) = \lim_{h\to 0} \frac{\mathbf{f}(\mathbf{a}+h\mathbf{u}) - \mathbf{f}(\mathbf{a})}{h}$$

and the other definitions are analogous. It is convenient to write

$$\mathbf{f}(x) = (f_1(\mathbf{x}, \dots, f_m(\mathbf{x})).$$

But for these functions, instead of the gradient, the total derivative (with respect to the standard basis) is given by the Jacobian:

$$D\mathbf{f}(\mathbf{a}) = \left[\frac{\partial f_i}{\partial x_j}\right] = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}.$$

Think about it this way: $\mathbf{f}$ eats things in $\mathbf{R}^n$ (a direction) and spits out things in $\mathbf{R}^m$ (how the image of $\mathbf{a}$ under $\mathbf{f}$ changes in that direction).

2. The General Chain Rule

The general chain rule is elegantly expressed in terms of linear algebra.

Theorem 2.1. *(General Chain Rule) Suppose that $\boldsymbol{h} = \boldsymbol{f} \circ \boldsymbol{g}$ and $\boldsymbol{g}$ is differentiable at $\boldsymbol{a}$ and $\boldsymbol{f}$ is differentiable at $\boldsymbol{g}(\boldsymbol{a})$, then*

$$D\boldsymbol{h}(\boldsymbol{a}) = D\boldsymbol{f}(\boldsymbol{g}(\boldsymbol{a})) \cdot D\boldsymbol{g}(\boldsymbol{a}),$$

where the $\cdot$ is just matrix multiplication.

Note that the usual one-variable chain rule is just a special case.

In an appendix, I'll write a proof of a chain rule, stated in more rigorous language.

Remark 2.2. If you're like me and have trouble keeping track of indices, I recommend computing the entire Jacobian of the functions you need to differentiate for the chain rule and multiply it out, even if they only ask for a single partial derivative.

This seems like a lot of extra work, but it's not hard and it can save time in case you make an arithmetic error in the calculation.

Date: April 16, 2015.

2.1. Another way to think of the chain rule.

The formula above is great, but I find it helpful to think about the chain rule as follows. Suppose that f depends on x and y, and that x and y in turn depend on t. Thus, to find $\frac{\partial f}{\partial t}$, we need to sum over all the ways in which f is affected by t, using the chain rule in each case. In symbols, this is written as

$$\frac{\partial f}{\partial t} = \frac{\partial f}{\partial x}\bigg|_{x(t)} \frac{\partial x}{\partial t} + \frac{\partial f}{\partial y}\bigg|_{y(t)} \frac{\partial y}{\partial t}.$$

This is just one coordinate of the chain rule matrix given by the formula above.

Example 1. Let's study a one-dimensional example. Things are easier here, since we can just use the gradient. Suppose that $\mathbf{r} : \mathbf{R} \to \mathbf{R}^n$ and $f : \mathbf{R}^n \to \mathbf{R}$ and $g \circ f = \mathbf{r}$, then

$$g'(t) = \nabla f(\mathbf{a}) \cdot \mathbf{r'}(t),$$

where $\mathbf{a} = \mathbf{r}(t)$.

Let $\mathbf{r}(t) = (a\cos(t), a\sin(t))$ and let $f(x, y) = x^2 - y^2$, and $g = f \circ \mathbf{r}$. Then

$$\nabla f(\mathbf{r}(t)) = (2x, -2y)\bigg|_{\mathbf{r}(t)} = (2a\cos(t), -2a\sin(t)).$$

We also have $\mathbf{r'}(t) = (-a\sin(t), a\cos(t))$. Therefore,

$$\begin{aligned} g'(t) &= \nabla f(\mathbf{r}(t)) \cdot \mathbf{r'}(t) \\ &= (2a\cos(t), -2a\sin(t)) \cdot (-a\sin(t), a\cos(t))^T \\ &= -2a^2\cos(t)\sin(t) - 2a^2\cos(t)\sin(t) \\ &= -2a^2\sin(2t). \end{aligned}$$

If you're geometrically minded, you can picture the saddle and the path that $g(t)$ takes and that the answer agrees with your intuition.

3. Tangent Spaces

If f is differentiable at a point, we can make sense of the notion of tangent space. However, in the multivariable context, we define the tangent space *with respect to a level set*, rather than to a graph, like tangent lines in single-variable calculus. The notation I use here isn't standard, but I don't know if there's really any standard notation in Apostol.

Suppose that f is differentiable at $\mathbf{a} \in L_c(f) = f^{-1}(c)$. If $\nabla f(\mathbf{a}) \neq 0$, then the **tangent space** $\theta_{\mathbf{a}}(L_c(f))$ to $L_c(f)$ at $\mathbf{a}$ is

$$\theta_{\mathbf{a}}(L_c(f)) = \{\mathbf{x} \in \mathbf{R}^n \mid \nabla f(\mathbf{a}) \cdot \mathbf{x} = 0\}.$$

Again, note that the tangent space lives in $\mathbf{R}^n$, not on the graph of the function.

Why is this the correct formula? Recall that the gradient points in the direction of greatest increase. Thus, if you are orthogonal to the gradient, you are in a direction of 0 increase (think about the angle formula for thee absolute value of the dot product). Thus, by moving in any direction in the plane $\nabla f(\mathbf{a} \cdot \mathbf{x} = 0$ keeps you inside the level set, because on the level set the function does not change.

Example 2. Let $f(x, y) = x^2 + y^2$. Then the level set $f^{-1}(1)$ is the circle of radius 1 in $\mathbf{R}^2$. At $(1, 0)$, we have $\nabla f(1, 0) = (2, 0)$ and so

$$\theta_{(1,0)}(L_1(f)) = \{\mathbf{x} \in \mathbf{R}^2 \mid (2, 0) \cdot \mathbf{x} = 0\},$$

which is just a vertical line.

Example 3. (Reconciling our intuition for tangent spaces) Thinking of tangent spaces on graphs is intuitive, so how can we, say, define a tangent plane to the *graph* of $f(x,y) = x^2 + y^2$? In order to do this, we need to do the following trick. Set

$$g(x,y,z) = f(x,y) - z = x^2 + y^2 - z$$

and consider the level set $L_0(g)$. This level set consists precisely of the points $(x,y,z) \in \mathbf{R}^3$ such that $x^2 + y^2 = z$, in other words, it's just the graph of of f.

Then we see that the gradient of g at, say, $(1,1,2)$ is $(2,2,-1)$. Thus, we have

$$\theta_{(1,1,2)}(L_0(g)) = \{\mathbf{x} \in \mathbf{R}^3 \mid (2,2,-1) \cdot \mathbf{x} = 0\}.$$

It becomes intuitive what this is if you draw a picture. The vector $(2,2,-1)$ points straight out from the graph of f, so the points orthogonal form the tangent plane at that point.

It's useful to use this trick on your homework problem, e.g. for finding a vector that is normal to a surafce.

4. Critical Points

We now return to the scalar field ($f : \mathbf{R}^n \to \mathbf{R}$) case. Last time we saw that if a point is a local extremum of a scalar function f, then the gradient of f is zero at that point.

However, the converse statement is not true. Consider $f(x,y) = x^2 - y^2$ at the point $(0,0)$. Since we are always interested in finding extrema, even if we have to search for them among a given set of points, we give these points a name. A point where the gradient of f is zero is called a **critical point** (or **stationary point**) of f.

Why are these sometimes called "stationary points"? I think the definition makes the most sense if you think about the problem physically. Say f measures the temperature in a certain $\mathbf{R}^n$ space. Recalling our reasoning on the last problem from last week, we saw that the gradient points in the *direction of greatest change*, in other words, it will always point to the locally "hottest point." Suppose you take the gradient of a scalar field at point p, then move a tiny bit in the direction the gradient is pointing at p, then evaluate the gradient and repeat the process. If the temperature function is bounded, then by following this process, you will eventually will arrive at the "hottest point," and since you are at the hottest point, no direction will point to a hotter point, and so the gradient will be zero, and you will remain stationary.

This thinking is slightly misleading in general, since it applies to "coldest points" as well, but it gives you an idea behind the terminology. Actually, for functions $f(x,y)$ in two variables, the stationary points correspond to peaks, pits, and saddle points. In Math 2a, you will study how points converges to these stationary points, which is a fascinating subject in itself.

To study critical points, we need to look at higher derivatives, much like how we studied maxima in the one-variable case by looking at the second-derivatives at that point. To do this, we introduce the following important object.

Definition 4.1. The **Hessian** matrix of a scalar field $f : \mathbf{R}^n \to \mathbf{R}$ is defined to be

$$H(\mathbf{x}) = [D_{ij} f(\mathbf{x})]_{i,j=1}^n$$

Remark 4.2. Note that

$$H(f)(x) = J(\nabla f)(x).$$

In particular, the Hessian matrix is symmetric, so there exists a basis of $\mathbf{R}^n$ consisting of the eigenvectors of H.

Observe that if $\mathbf{x}$ is an eigenvector of H, then the sign of the derivative in the direction of $\mathbf{x}$ is $\mathbf{x}H(\mathbf{a})\mathbf{x}^T$. Therefore, the sign of the derivative of f in the various directions at $\mathbf{a}$ corresponds with the signs of the eigenvalues of H. In particular, we have three cases.

- (H is *negative definite*) If all the eigenvalues of H are negative, then f has a maximum.
- (H is *positive definite*) If all the eigenvalues of H are positive, then f has a minimum.
- If there are eigenvalues of both signs, then f has a saddle point.

Remark 4.3. Note that this is just a generalization of the second-derivative criteria for extrema in the one-variable case, i.e. if the function is concave down, then we have a local maximum; if concave up, then we have a local minimum. We just have an additional ambiguity here with the saddle point since we are in the multiple-variable setting.

Example 4. Find and classify the critical points of $f(x, y) = x^2 + y^2$.

Solution. We have $\nabla f = (2x, 2y)$. Therefore, the only critical point of f is $(0, 0)$. Computing the Hessian, we find that

$$H_f = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}.$$

Therefore, the eigenvalues of H_f are 2 and 2, so H_f is positive definite, implying that $(0, 0)$ is a local minimum. □

Critical points are subtle objects. They may not always behave like you'd imagine.

For instance, $f(x, y, z) = \sin(x^2 + y^2 + z^2)$ is an example that shows that critical points don't need to be isolated, since every point in the 3-sphere is a critical point of f. (Work it out!)

A function may also have no critical points whatsoever. Consider the (slightly modified) Gaussian function $f(x, y) = \int_x^y e^{-t^2}\, dt$. Then $\nabla f(x, y) = (-e^{-x^2}, e^{-y^2})$, which is never zero.

Warning 4.4. To find extrema for a function restricted to a region, you must check the critical points *and the boundary points*, just like in the single-variable case. This is because if a function is defined on a region, you may have global extrema that are not critical points.

4.1. How to find extrema of multivariable functions.

(a) Find and classify all critical points in the defined region and classify them as local/relative extrema or saddle points by looking at the Hessian at that point.

(b) Find the extrema on the boundary by evaluating the function at the boundary points, and see if any of these are global extrema (larger or smaller than your local extrema found above). Since the gradient may not be zero here, you don't have to worry about finding relative extrema or saddle points.

(c) Neatly and clearly describe all the local extrema, the saddles, and the global extrema on the interior of the region defined, as well as the global extrema that may lie on the boundary.

5. Appendix: Proof of the Chain Rule

Theorem 5.1. *(The Chain Rule) Let $f : D \subseteq \boldsymbol{R}^n \to \boldsymbol{R}^m$ and $g : E \subseteq \boldsymbol{R}^k \to \boldsymbol{R}^n$. Let $a \in Int(E)$ and $g(a) = b \in Int(D)$ and assume that $g'(a)$ and $f'(b) = f'(g(a))$ exist. Then $h = f \circ g$ is defined in a neighborhood of a (by continuity of g at a) and $h'(a)$ exists and is given by*

$$h'(a) = f'(g(a)) \circ g'(a).$$

(Under the standard basis, this is just the Jacobian formula given above.)

Proof. Take a small nonzero $y \in \mathbf{R}^k$. Then

$$h(a + y) - h(a) = f(g(a + y)) - f(g(a)) = f(b + v) - f(b)$$

with $b = g(a + y) - g(a) \in \mathbf{R}^n$. Since g is continuous at a (by differentiability at a), the vector v is small as well (i.e. $y \to 0$ implies that $v \to 0$). Now

$$v = g'(a)(y) + ||y|| \cdot E_g(a, y)$$

where $E_g(a, y) \to 0$ as $y \to 0$ and

$$f(b + v) - f(b) = f'(b)(v) + ||v|| \cdot E_f(b, v)$$

where $E_f(b, v) \to 0$ as $v \to 0$. Then

$$\begin{aligned} h(a + y) - h(a) &= f(b + v) - f(b) \\ &= f'(b)(g'(a)y + ||y||E_g(a, y)) + ||v||E_f(b, v) \\ &= f'(b) \cdot g'(a)(y) + ||y||E(a, y) \end{aligned}$$

by by linearity of $f'(b)$ and where

$$(*) \qquad E(a, y) = f'(b)(E_g(a, y)) + \frac{||v||}{||y||}E_f(b, v).$$

We need to show that $E(a, y) \to 0$ as $y \to 0$. Since $E_g(a, y) \to 0$ as $y \to 0$ and linear transformations are continuous, the first term in $(*)$ goes to zero.

We now prove a little lemma.

> **Lemma.** If $T : \mathbf{R}^p \to \mathbf{R}^q$ is a linear transformation, then for some $c > 0$, we have
>
> $$||T(v)|| \le c||v||$$
>
> for all $v \in \mathbf{R}^p$. *Proof of Lemma.* Let $d \ge ||T(e_i)||$ for any $i = 1, \dots, p$ for $\{e_i\}$ the standard basis of $\mathbf{R}^p$. We can write $v = \sum_{i=1}^{p} \alpha_i e_i$, and

we have

$$\begin{aligned}||T(v)|| &= ||\sum_{i=1}^{p} \alpha_i T(e_i)|| \\ &\leq \sum_{i=1}^{p} |\alpha_i| \, ||T(e_i)|| \\ &=\leq \sum_{i=1}^{p} |\alpha_i| \cdot d \\ &= d \sum_{i=1}^{p} |\alpha_i| \cdot 1 \\ &\leq d \sqrt{\sum_{i=1}^{p} \alpha_i^2} \cdot \sqrt{\sum_{i=1}^{p} 1^2} \\ &\leq d\sqrt{n}||v||.\end{aligned}$$

Taking $c = d\sqrt{n}$, we are done. $\diamond$

Now, we have

$$\begin{aligned}||v|| &\leq ||g'(a)(y)|| + ||y|| \cdot ||E_g(a, y)|| \\ &\leq c \cdot ||y|| + ||y|| \cdot ||E_g(a, y)||\end{aligned}$$

or $\frac{||v||}{||y||} \leq C + ||E_g(a, y)||$, so $\frac{||v||}{||y||}$ is bounded and $E_f(b, v) \to 0$ as $y \to 0$ (since $v \to 0$ as well). Therefore, the second term in $(*)$ also goes to 0. $\square$