Math 94 Professor: Padraic Bartlett

Lecture 3: Derivatives in $\mathbb{R}^n$

Week 3 *UCSB 2015*

This is the third week of the Mathematics Subject Test GRE prep course; here, we review the concepts of **derivatives** in higher dimensions!

1 Definitions and Concepts

We start by reviewing the definitions we have for the derivative of functions on $\mathbb{R}^n$:

Definition. The **partial derivative** $\frac{\partial f}{\partial x_i}$ of a function $f : \mathbb{R}^n \to \mathbb{R}$ along its i-th coördinate at some point $\mathbf{a}$, formally speaking, is the limit

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

(Here, e_i is the i-th basis vector, which has its i-th coördinate equal to 1 and the rest equal to 0.)

However, this is not necessarily the best way to think about the partial derivative, and certainly not the easiest way to calculate it! Typically, we think of the i-th partial derivative of f as the derivative of f when we "hold all of f's other variables constant" – i.e. if we think of f as a single-variable function with variable x_i, and treat all of the other x_j's as constants. This method is markedly easier to work with, and is how we actually *calculate* a partial derivative.

We can extend this to higher-order derivatives as follows. Given a function $f : \mathbb{R}^n \to \mathbb{R}$, we can define its **second-order partial derivatives** as the following:

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_j}\right).$$

In other words, the second-order partial derivatives are simply all of the functions you can get by taking two consecutive partial derivatives of your function f.

Definition. Often, we want a way to talk about **all** of the first-order derivatives of a function at once. The way we do this is with the **differential**, or **total derivative**. We define this as follows: the total derivative of a function $f : \mathbb{R}^n \to \mathbb{R}^m$ is the following matrix of partial derivatives:

$$D(f)\big|_{\mathbf{a}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{a}) & \frac{\partial f_1}{\partial x_2}(\mathbf{a}) & \dots & \frac{\partial f_1}{\partial x_n}(\mathbf{a}) \\ \frac{\partial f_2}{\partial x_1}(\mathbf{a}) & \frac{\partial f_2}{\partial x_2}(\mathbf{a}) & \dots & \frac{\partial f_2}{\partial x_n}(\mathbf{a}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1}(\mathbf{a}) & \frac{\partial f_n}{\partial x_2}(\mathbf{a}) & \dots & \frac{\partial f_n}{\partial x_n}(\mathbf{a}) \end{bmatrix}$$

For a function $f : \mathbb{R}^n \to \mathbb{R}$, this has the special name **gradient**, and is denoted

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots \frac{\partial f}{\partial x_n}\right)$$

For a function $f : \mathbb{R}^n \to \mathbb{R}$, a point $\mathbf{a}$ is called a **critical point** if it is a stationary point, or f is not differentiable in any neighborhood of $\mathbf{a}$. Similarly, a point $\mathbf{a} \in \mathbb{R}^n$ is called a **local maxima** of f if there is some small value r such that for any point $\mathbf{x}$ within distance r of $\vec{a}$, we have $f(\mathbf{x}) \leq f(\mathbf{a})$. (A similar definition holds for **local minima**.)

Definition. For functions $f : \mathbb{R}^n \to \mathbb{R}$, we can also define an object that generalizes the "second-derivative" from one-dimensional calculus to multidimensional calculus. We do this with the Hessian, which we define here. The Hessian of a function $f : \mathbb{R}^n \to \mathbb{R}$ at some point $\mathbf{a}$ is the following matrix:

$$H(f)\big|_{\mathbf{a}} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(\mathbf{a}) & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(\mathbf{a}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(\mathbf{a}) & \dots & \frac{\partial^2 f}{\partial x_n \partial x_n}(\mathbf{a}) \end{bmatrix}.$$

Finally: like with the normal second derivative, we can use $H(f)\big|_{\mathbf{a}}$ to create a "second-order" approximation to f at $\mathbf{a}$, in a similar fashion to how we used the derivative to create a linear (i.e. first-order) approximation to f. We define this here: if $f : \mathbb{R}^n \to \mathbb{R}$ is a function with continuous second-order partials, we define the **second-order Taylor approximation** to f at $\mathbf{a}$ as the function

$$T_2(f)\big|_{\mathbf{a}}(\mathbf{a} + \mathbf{h}) = f(\mathbf{a}) + (\nabla f)(\mathbf{a}) \cdot \mathbf{h} + \frac{1}{2} \cdot (h_1, \dots h_n) \cdot H(f)\big|_{\mathbf{a}} \cdot (h_1, \dots h_n)^T.$$

You can think of $f(\mathbf{a})$ as the constant, or zero-th order part, $(\nabla f)(\mathbf{a}) \cdot \mathbf{h}$ as the linear part, and $H(f)\big|_{\mathbf{a}}(\mathbf{h})$ as the second-order part of this approximation.

Definition. Finally, we have two useful physical phenomena, the **divergence** and **curl**, that have natural interpretations. Given a C^1 vector field $F : \mathbb{R}^3 \to \mathbb{R}^3$, we can defind the **divergence** and **curl** of F as follows:

- **Divergence**. The **divergence** of F, often denoted either as $\text{div}(F)$ or $\nabla \cdot F$, is the following function $\mathbb{R}^3 \to \mathbb{R}$:

$$\text{div}(F) = \nabla \cdot F = \frac{\partial F_1}{\partial x} + \frac{\partial F_2}{\partial y} + \frac{\partial F_3}{\partial z}.$$

- **Curl.** The **curl** of F, denoted $\text{curl}(F)$ or $\nabla \times F$, is the following map $\mathbb{R}^3 \to \mathbb{R}^3$:

$$\text{curl}(F) = \nabla \times F = \left(\left(\frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z}\right), \left(\frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x}\right), \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\right)\right).$$

Often, the curl is written as the "determinant" of the following matrix:

$$\det \begin{bmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ F_1 & F_2 & F_3 \end{bmatrix}$$

Given a function $F : \mathbb{R}^2 \to \mathbb{R}^2$, we can also find its curl by "extending" it to a function $F^\star : \mathbb{R}^3 \to \mathbb{R}^3$, where $F_1^\star(x,y,z) = F(x,y), F_2^\star(x,y,z) = F(x,y)$, and $F_3^\star(x,y,z) = 0$. If someone asks you to find the curl of a function that's going from $\mathbb{R}^2 \to \mathbb{R}^2$, this is what they mean.

Also, divergence naturally generalizes to working on any function $\mathbb{R}^n \to \mathbb{R}^n$; just take the sum of $\frac{\partial F_i}{\partial x_i}$ over all of the variables the function depends on.

We also have several theorems that we know about the derivative! We list a few here. Here's how we extend the product and chain rules:

Theorem. Suppose that f, g are a pair of functions $\mathbb{R}^n \to \mathbb{R}^m$, and we're looking at the inner product[1] $f \cdot g$ of these two functions. Then, we have that

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

Theorem. Take any function $g : \mathbb{R}^m \to \mathbb{R}^l$, and any function $f : \mathbb{R}^n \to \mathbb{R}^m$. Then, we have

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

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

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

[1] Recall that the **inner product** of two vectors $\mathbf{u}, \mathbf{v}$ is just the real number $\sum_{i=1}^m u_i v_i$.

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 $D(g)$!

Here's how the derivative interacts with finding maxima and minima:

Theorem. A function $f : \mathbb{R}^n \to \mathbb{R}$ has a local maxima at a critical 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[2] at $\mathbf{a}$. Similarly, it has a local minima if the Hessian is positive-definite at $\mathbf{a}$. If the Hessian takes on both positive and negative values there, it's a **saddle point**: there are directions you can travel where your function increase, and others where it will decrease. Finally, if the Hessian is identically 0, you have no information as to what your function may be up to: you could be in any of the three above cases.

In the section above, we talked about how to use derivatives to find and classify the **critical points** of functions $\mathbb{R}^n \to \mathbb{R}$. This allows us to find the global minima and maxima of functions over all of $\mathbb{R}^n$, if we want! Often, however, we won't just be looking to find the maximum of some function on all of $\mathbb{R}^n$: sometimes, we'll want to maximize a function *given a set of constraints.* For example, we might want to maximize the function $f(x, y, z) = x+y$ subject to the constraint that we're looking at points where $x^2 + y^2 = 1$. How can we do this?

Initially, you might be tempted to just try to use our earlier methods: i.e. look for places where Df is 0, and try to classify these extrema. The problem with this method, when we have a set of constraints, is that it usually **won't** find the maxima or minima on this constraint: because it's only looking for local maxima or minima over all of $\mathbb{R}^n$, it will ignore points that could be maxima or minima on our constrained surface! I.e. for the f, g we mentioned above, we know that $\nabla(f) = (1, 1)$, which is never 0; however, we can easily see by graphing that $f(x, y) = x + y$ should have a maximum value on the set $x^2 + y^2 = 1$, specifically at $x = y = \frac{1}{\sqrt{2}}$.

Theorem. So: how can we find these maxima and minima in general? The answer is the method of **Lagrange multipliers**, which we outline here.

[2] The Hessian $H(f)\big|_{\mathbf{a}}$ is positive-definite if and only if the matrix

$$\begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(\mathbf{a}) & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(\mathbf{a}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(\mathbf{a}) & \dots & \frac{\partial^2 f}{\partial x_n \partial x_n}(\mathbf{a}) \end{bmatrix}$$

is positive-definite. (The same relation holds for being negative-definite.)

Recall from Math 1a that a matrix is positive-definite if and only if all of its eigenvalues are real and positive. Similarly, a matrix is negative-definite if and only if all of its eigenvalues are real and negative. If some of a matrix's eigenvalues are 0, some are negative and others are positive, or if there are less real eigenvalues than the rank of the matrix (i.e. some eigenvalues are complex,) then the matrix is neither positive-definite or negative-definite.

Note also that because the Hessian is symmetric whenever the mixed partials of our function are equal, and symmetric matrices have only real eigenvalues, you really should never get complex-valued eigenvalues.

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a function whose extremal values $\{\mathbf{x}\}$ we would like to find, given the constraints $g(\mathbf{x}) = c$, for some constraining function $g(\mathbf{x})$. Then, we have the following result: if $\mathbf{a}$ is an extremal value of f restricted to the set $S = \{\mathbf{x} : \forall i, g(\mathbf{x}) = c\}$, then either one of $\nabla(f)\big|_{\mathbf{a}}$ is 0, doesn't exist, or there is some constant λ such that

$$\nabla(f)\big|_{\mathbf{a}} = \lambda\nabla(g)\big|_{\mathbf{a}}.$$

Theorem. We have a pair of rather useful theorems about the divergence and curl of functions, which we state here:

- For any C^2 function F, div(curl(F)) is always 0.
- For any C^2 function F, curl(grad(F)) is always 0.

2 Worked Examples

Example. (Lagrange multipliers; level curves.) Consider the function

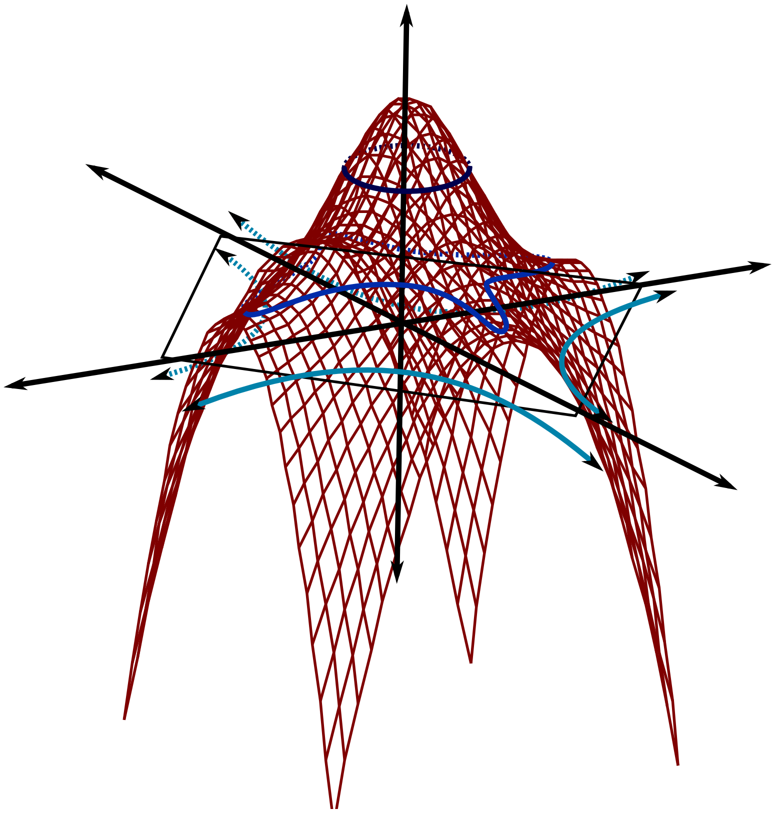
$$g(x, y) = e^{-x^2-y^2} - x^2y^2.$$

(a) Draw several level curves of this function.

(b) Let $f(x, y) = x + y$, and let S be the constraint set given by the level curve $\{(x, y) : g(x, y) = c\}$. For what values of c does $f\big|_S$ have a global maximum? For what values does it fail to have a global maximum: i.e. for what values of c is f unbounded on S?

(c) For $c = \frac{1}{4}$, find the global maximum of f on the above constraint set $S = \{(x, y) : g(x, y) = c\}$.

Solution. We graph $g(x, y) = z$ in red, along with three level curves in different shades of blue, in the following picture.



Roughly speaking, there are three kinds of level curves for our function:

1. Level curves $g(x, y) = c$, where c is close to 1. There, because we need g to be close to 1, we need to have x and y very small (so that the $e^{-x^2-y^2}$ part is as close to 1 as we can get it, and the $-x^2y^2$ part is not too large.) In particular, this forces us to have a roughly circular shape, as for very small values of (x, y) the x^2y^2 part is insignificant

and our function looks roughly like $e^{-x^2-y^2}$, which is roughly $1-x^2-y^2$ (via Taylor series) for small values of (x, y).

2. Level curves $g(x, y) = c$, where c is greater than 0, but not by much. For these values of c, we wind up having kind of a "four-armed" shape, with arms stretching out along the x- and y- axes. This is because when one of our coordinates is nearly zero, the other can become much larger (because our function is roughly $e^{-x^2-y^2}$ then), whereas when the coordinates are roughly the same, the dominant term is now the $-x^2y^2$ term, and we need to have both x and y be much smaller.

3. Level curves $g(x, y) = c$, where c is ≤ 0. In these cases, our level curves look like hyperbola-style curves, one in each quadrant. This is because on each axis, our function $g(x, y)$ can never be 0, as the $e^{-x^2-y^2}$-part is always positive and the $-x^2y^2$ part is zero on the axes.

This graphing and subsequent analysis suggests an answer to part (b), as well:

Claim. Our function $f(x, y)$ has a global maximum on the curve $g(x, y) = c$ if and only if $1 \geq c > 0$.

Proof. If $c > 1$, then there are no points (x, y) such that $g(x, y) = c$, because $e^{-x^2-y^2}$ is bounded above by $e^0 = 1$, while $-x^2y^2$ is bounded above by 0.

So: suppose that $1 \geq c > 0$. Then, if (x, y) are such that $g(x, y) = c$, we know that in particular

$$\begin{aligned} & e^{-x^2-y^2} \geq c \\ \Rightarrow \quad & -x^2 - y^2 \geq \ln(c) \\ \Rightarrow \quad & x^2 + y^2 \leq -\ln(c) \\ \Rightarrow \quad & \sqrt{x^2+y^2} \leq \sqrt{-\ln(c)} \\ \Rightarrow \quad & ||(x, y)|| \leq \sqrt{-\ln(c)}, \end{aligned}$$

i.e. the point (x, y) can be no further than $\sqrt{-\ln(c)}$ from the origin. (Because $1 \geq c > 0$, we know that $-\infty < \ln(c) \leq 0$, and therefore that this is a well-defined finite and real-valued bound on distances.)

Therefore, the set of points such that $g(x, y) = c$ is bounded. We also know that it is closed, because it is the level curve of a continuous function. Therefore, we know that any continuous function (in particular, f) will attain its global maxima and minima on this set, and do so at the critical points identified by the method of Lagrange multipliers.

Finally, suppose that $c \leq 0$. In this case, our claim is that f does not attain its global maximum on $g(x, y) = c$. To prove this, pick any value of n: we want to find a point (x, y) on our curve such that $f(x, y) > n$.

To do this, we simply use the intermediate value theorem. Pick any n, and choose x such that $-x^2 < c - 1$, and also $x > n$. Then, we know that

$$g(x, 0) = e^{-x^2-0} - x^2 \cdot 0 = e^{-x^2} > 0 \geq c,$$

while

$$g(x,1) = e^{-x^2-1} - x^2 \cdot 1 = e^{-x^2} - x^2 < e^{-x^2} - c - 1 < c,$$

because $e^{-x^2} < 1$.

Therefore,because $g(x,0) > c$ and $g(x,1) < c$, by the intermediate value theorem, there is some value of y between 0 and 1 such that $g(x,y) = c$. At this point (x,y), we know that

$$f(x,y) = x + y \geq n + 0 \geq n,$$

which is what we wanted to prove: i.e. we've shown that we can find points on our curve along which $f(x,y)$ is arbitrarily large, and therefore that there is no global maximum. □

Finally, with this theoretical discussion out of the way, we can turn to the calculational part of (c), which asks us to find the global maximum of our function f on the constraint set $g(x,y) = \frac{1}{4}$. First, note that by our above discussion, we know that a global maximum does exist, because when $1 \geq c > 0$ we've shown that our constraint set is closed and bounded. Furthermore, to find this maximum, it suffices to use the method of Lagrange multipliers to find all of the critical points of our function restricted to this curve, and simply select the largest value amongst these critical points. (Again, this is because $g(x,y) = c$ is closed and bounded, which means that our global maximum must occur a critical point.)

So: we calculate. We are looking for any points (x,y) such that either

- $\nabla(f)$ or $\nabla(g)$ are 0,
- $\nabla(f)$ or $\nabla(g)$ are undefined, or
- there is some nonzero constant λ such that $\nabla(f) = \lambda\nabla(g)$.

Because

$$\nabla(f)(x,y) = (1,1)\,,$$

we can immediately see that $\nabla(f)$ is never undefined or zero.

Similarly, because

$$\nabla(g) = \left(-2xe^{-x^2-y^2} - 2xy^2, -2ye^{-x^2-y^2} - 2yx^2\right),$$

we can see that the first component of $\nabla(g)$ is zero if and only if

$$\begin{aligned} &0 = -2xe^{-x^2-y^2} - 2xy^2 \\ \Leftrightarrow &0 = -2x\left(e^{-x^2-y^2} + y^2\right) \\ \Leftrightarrow &0 = x, \text{ because } e^{-x^2-y^2} + y^2 \text{ is strictly positive.} \end{aligned}$$

Similarly, we can see that the second component of $\nabla(g)$ is zero if and only if

$$\begin{aligned} &0 = -2ye^{-x^2-y^2} - 2yx^2 \\ \Leftrightarrow &0 = -2y\left(e^{-x^2-y^2} + x^2\right) \\ \Leftrightarrow &0 = y, \text{ because } e^{-x^2-y^2} + x^2 \text{ is strictly positive.} \end{aligned}$$

So $\nabla(g)$ is always defined and is only zero at $(0,0)$, which is not a point on our curve $g(x,y)=\frac{1}{4}$. Therefore, the only points we're concerned with are ones at which $\nabla(f)=\lambda\nabla(g)$; i.e. points such that

$$\begin{aligned}\nabla(f)=(1,1)=\lambda\nabla(g)=\lambda\left(-2xe^{-x^2-y^2}-2xy^2,-2ye^{-x^2-y^2}-2yx^2\right)\\ \Leftrightarrow -2xe^{-x^2-y^2}-2xy^2=-2ye^{-x^2-y^2}-2yx^2,\end{aligned}$$

because the above equation is equivalent to forcing both the left and right coordinates of $\nabla(g)$ to equal the same quantity (namely, $\frac{1}{\lambda}$.)

Solving, we can see that this is equivalent to

$$\begin{aligned}0=2xe^{-x^2-y^2}+2xy^2-2ye^{-x^2-y^2}-2yx^2\\ \Leftrightarrow 2(x-y)e^{-x^2-y^2}-2xy(x-y)=0.\end{aligned}$$

If $x-y=0$, i.e. $x=y$, this equation holds. Otherwise, we can divide through by $2(x-y)$, and get

$$e^{-x^2-y^2}=xy.$$

Plugging this into our constraint equation $g(x,y)=\frac{1}{4}$ gives us

$$e^{-x^2-y^2}-(xy)^2=\frac{1}{4}\Rightarrow(xy)-(xy)^2=\frac{1}{4}\Rightarrow xy=\frac{1}{2},$$

by thinking of "xy" as one term and using the quadratic formula. But, if we think about what this means for the equation $e^{-x^2-y^2}=xy$, and specifically use $y=\frac{1}{2x}$, we have

$$\frac{1}{2}=xy=e^{-x^2-y^2}=e^{-x^2-\frac{1}{4x^2}}.$$

This is impossible! In specific, by taking a single-variable derivative, you can easily see that the largest value of $-x^2-\frac{1}{4x^2}$ happens at $x=\frac{1}{\sqrt{2}}$, at which this is -1. This means that the largest that $e^{-x^2}-\frac{1}{4x^2}$ gets is $e^{-1}=\frac{1}{e}$, which is smaller than $\frac{1}{2}$.

Therefore, the only points at which $\nabla(f)=\lambda\nabla(g)$ are those at which $x=y$. Plugging this into our constraint $g(x,y)=\frac{1}{4}$ yields

$$\begin{aligned}e^{-2x^2}-x^4=\frac{1}{4}\\ \Rightarrow x\equiv\pm.65.\end{aligned}$$

The function $f(x,y)=x+y$ is equal to 1.3 at the point $(.65,.65)$ and is equal to -1.3 at $(-.65,-.65)$. Therefore, by our discussion earlier about how f must attain its global minima and maxima at the critical points discovered by the Lagrange multiplier process, we can safely conclude that $(.65,.65)$ is roughly the point at which $f(x,y)$ attains its global maxima, which is roughly 1.3.

Example. (Tangent planes.) Let S be the surface in $\mathbb{R}^3$ formed by the collection of all points (x,y,z) such that $e^{xyz}=e$. Find the tangent plane to S at $(1,1,1)$.

Solution. One way to attack this problem is to apply natural logs to both sides, which lets us write S as the collection of all points (x, y, z) such that $xyz = 1$; i.e. all points $x, y \neq 0$ such that $z = \frac{1}{xy}$. In other words, we can write S as the graph of the function $f(x, y) = \frac{1}{xy}$. We know that the gradient of $f(x, y)$ is just

$$\left(-\frac{y}{(xy)^2}, -\frac{x}{(xy)^2}\right),$$

which at 1 is just $(-1, -1)$. Therefore, by using the formula for describing the first-order Taylor approximation – i.e. tangent plane – of functions of the form $f(x, y) = z$, we have that the tangent plane to our surface at $(1, 1, 1)$ is just

$$\begin{aligned}(z-1) &= \nabla(f)\Big|_{(1,1,1)} \cdot (x-1, y-1) = (-1, -1) \cdot (x-1, y-1) \\ \Rightarrow & z - 1 + x - 1 + y - 1 = 0.\end{aligned}$$

Alternately, we also discussed a second formula in class for finding tangent planes to surfaces of the form $g(x, y, z) = C$, at some point (a, b, c). Specifically, we observed that the gradient of g at the point (a, b, c) was orthogonal to the tangent plane to our surface at this point: in other words, that we could define our tangent plane as just the set of all vectors orthogonal to the gradient of g through this point. As a formula, this was

$$\begin{aligned}0 &= \nabla(g)\Big|_{(1,1,1)} \cdot (x-1, y-1, z-1) \\ \Leftrightarrow 0 &= \left(yze^{xyz}, xze^{xyz}, xye^{xyz}\right)\Big|_{(1,1,1)} \cdot (x-1, y-1, z-1) \\ \Leftrightarrow 0 &= (1, 1, 1) \cdot (x-1, y-1, z-1) \\ \Leftrightarrow 0 &= z - 1 + x - 1 + y - 1.\end{aligned}$$

Reassuringly, we get the same answer no matter which method we pick.

Example. (Chain rule.) Let $g : \mathbb{R}^4 \to \mathbb{R}$ be defined by the equation $(w, x, y, z) = (wz - yx)$, and $h_\lambda : \mathbb{R}^2 \to \mathbb{R}^4$ be defined by the equation $h_\lambda(a, b) = (a, \lambda a, b, \lambda b)$.

(a) Calculate the derivative of $g \circ h_\lambda$ using the chain rule.

(b) Geometrically, explain why your answer in (a) is "obvious," in some sense.

Solution. So, we know that both g and h_λ are continuous functions on all of their domains; therefore, we know that their composition is continuous everywhere. Therefore, we know that the total derivative of $g \circ h_\lambda$ is just given by the partial derivatives of $g \circ h_\lambda$: i.e.

$T(g \circ h_\lambda) = D(g \circ h_\lambda)$. Therefore, we can use the chain rule:

$$\begin{aligned}
D(g \circ h_\lambda)\Big|_{(a,b)} &= D(g)\Big|_{h_\lambda(a,b)} \cdot D(h_\lambda)\Big|_{(a,b)} \\
&= \begin{bmatrix} z & -y & -x & w \end{bmatrix}\Big|_{h_\lambda(a,b)} \cdot \begin{bmatrix} 1 & 0 \\ \lambda & 0 \\ 0 & 1 \\ 0 & \lambda \end{bmatrix} \\
&= \begin{bmatrix} \lambda b & -b & -\lambda a & a \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ \lambda & 0 \\ 0 & 1 \\ 0 & \lambda \end{bmatrix} \\
&= [\lambda b - \lambda b, \lambda a - \lambda a] \\
&= [0, 0].
\end{aligned}$$

Notice that this is geometrically somewhat obvious because g is just the determinant of the matrix $\begin{pmatrix} w & x \\ y & z \end{pmatrix}$, while the function h_λ just outputs the rank-1 matrix $\begin{pmatrix} a & b \\ \lambda a & \lambda b \end{pmatrix}$. Because the determinant of a rank 1 matrix is 0, we have that $g \circ h_\lambda$ is identically 0, and therefore also has derivative 0.

Example. (Taylor series; directional derivatives.) Let $g(x, y) = \sin(xy)$.

(a) Calculate the directional derivative of $g(x, y)$ at $(1, 2)$ in the direction $(3, 4)$.

(b) Calculate the second-order Taylor approximation of $g(x, y)$ at $(0, 0)$.

Solution. Because the gradient of g is just

$$\nabla(g) = (y \cos(xy), x \cos(xy)),$$

we know that the directional derivative at $(1, 2)$ in the direction $(3, 4)$ is just given to us by the dot product of $\nabla(g)(1, 2)$ with the **unit-length vector in the direction** $(3, 4)$, given by $\frac{1}{||(3,4)||} \cdot (3, 4) = \frac{1}{\sqrt{9+16}}(3, 4) = \left(\frac{3}{5}, \frac{4}{5}\right)$:

$$\nabla(g)(1, 2) \cdot \left(\frac{3}{5}, \frac{4}{5}\right) = (2\cos(1), \cos(2)) \cdot \left(\frac{3}{5}, \frac{4}{5}\right) = \frac{6\cos(1) + 4\cos(2)}{5}.$$

To calculate the Taylor approximation of g at $(0, 0)$, we just need to construct the following function:

$$T_2(g)\big|_{(0,0)}(h_1, h_2) = g(0, 0) + \nabla(g)\big|_{(0,0)} \cdot (x, y) + H(g)\big|_{(0,0)}(x, y).$$

To do this, simply note that the Hessian $H(g)$ of g is just

$$\begin{aligned}
H(g)\big|_{(0,0)}(h_1,h_2) &= \frac{1}{2}[h_1,h_2]\left[\begin{array}{cc} -y^2\sin(xy) & \cos(xy)-xy\sin(xy) \\ \cos(xy)-xy\sin(xy) & -x^2\sin(xy) \end{array}\right]\Bigg|_{(0,0)}\left[\begin{array}{c} h_1 \\ h_2 \end{array}\right] \\
&= \frac{1}{2}[h_1,h_2]\left[\begin{array}{cc} 0 & 1 \\ 1 & 0 \end{array}\right]\cdot\left[\begin{array}{c} h_1 \\ h_2 \end{array}\right] \\
&= \frac{1}{2}[h_1,h_2]\left[\begin{array}{c} h_2 \\ h_1 \end{array}\right] \\
&= \frac{1}{2}(h_1h_2+h_1h_2) \\
&= h_1h_2,
\end{aligned}$$

and therefore that

$$\begin{aligned}
T_2(g)\big|_{(0,0)}(h_1,h_2) &= g(0,0)+\nabla(g)\big|_{(0,0)}\cdot(x,y)+H(g)\big|_{(0,0)}(x,y) \\
&= \sin(0)+(0\cos(0),0\sin(0))\cdot(x,y)+xy \\
&= xy.
\end{aligned}$$

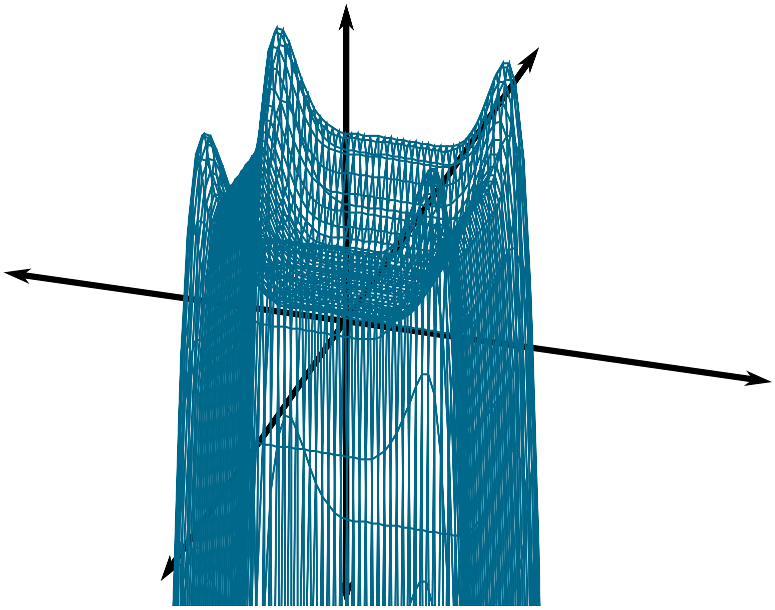
Therefore, the second-order approximation to $\sin(xy)$ at the origin is just $T_2(x,y)=xy$.

Example. (Using derivatives to study local extrema.) Let

$$f(x,y) = -(x^8+y^8)+4(x^6+y^6)-4(x^4+y^4).$$

Find all of the critical points of f, and classify them as local maxima, minima, or saddle points.

Solution. We start by graphing our function:



Roughly speaking, it looks like we have four global maxima, at least four saddle points between these maxima, and probably a bunch of weird things going on in the interior part of our function which are hard to determine from our picture. Probably a local minima in there.

Picture aside, our task here is pretty immediate:

1. First, we want to calculate $\nabla(f)$, and find all of the points where it is either undefined or 0. These are our **critical points**.

2. We then want to calculate $H(f)$, the Hessian of f, for each critical point. If the Hessian is **positive-definite**[3], then we know that this point is a local minimum; if it is **negative-definite**, then it's a local maximum; if it **has both a positive eigenvalue and a negative eigenvalue**, it's a saddle point; and if it's **anything else**, we have no idea what's going on, and will need to explore its behavior using other methods.

So: by calculating, we can see that

$$D(f) = (-8x^7 + 24x^5 - 16x^3, -8y^7 + 24y^4 - 16y^3),$$

and therefore that this is equal to 0 whenever

$$\begin{aligned}
&0 = -8x^7 + 24x^5 - 16x^3 \\
\Leftrightarrow &x = 0, \text{ or} \\
&0 = -8x^4 + 24x^2 - 16 \\
\Leftrightarrow &0 = (x^2 - 2)(x^2 - 1) \\
\Leftrightarrow &x = \pm\sqrt{2}, \pm 1,
\end{aligned}$$

and

$$\begin{aligned}
&0 = -8y^7 + 24y^5 - 16y^3 \\
\Leftrightarrow &y = 0, \pm\sqrt{2}, \pm 1.
\end{aligned}$$

So we have twenty-five critical points, consisting of five choices of x and five choices of y. To classify these points, we look at the matrix of second-order-partials formed in the Hessian:

$$\begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(\mathbf{a}) & \ldots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(\mathbf{a}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(\mathbf{a}) & \ldots & \frac{\partial^2 f}{\partial x_n \partial x_n}(\mathbf{a}) \end{bmatrix} = \begin{bmatrix} -56x^6 + 120x^4 - 48x^2 & 0 \\ 0 & -56y^6 + 120y^4 - 48y^2 \end{bmatrix}.$$

[3] We say that the Hessian is positive-definite if the associated matrix $\begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(\mathbf{a}) & \ldots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(\mathbf{a}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(\mathbf{a}) & \ldots & \frac{\partial^2 f}{\partial x_n \partial x_n}(\mathbf{a}) \end{bmatrix}$ of second partial derivatives is positive-definite: i.e. it has n eigenvalues and they're all strictly positive. Negative-definite is similar, except we ask that all of the eigenvalues exist and are strictly negative.

When $x = \pm 1$, the polynomial $-56x^6+120x^4-48x^2$ is 16, which is positive; when $x = \pm\sqrt{2}$, this polynomial is -64, which is negative; finally, when $x = 0$ this polynomial is 0. Therefore, at the points

$$(\pm 1, \pm 1)$$

the Hessian is positive-definite, and therefore our function has a local minimum, while at the points

$$(\pm\sqrt{2}, \pm\sqrt{2})$$

the Hessian is negative-definite, and therefore our function has a local maximum, while at

$$(\pm\sqrt{2}, \pm 1), (\pm 1, \pm\sqrt{2}),$$

the Hessian has both a negative and a positive eigenvalue (try $(1, 0), (0, 1)$ for two eigenvectors!), and therefore our function has a saddle point.

This leaves just the points with a zero-coordinate, at which the Hessian is useless to us. There, we need to analyze how small changes in our function

$$f(x, y) = -(x^8 + y^8) + 4(x^6 + y^6) - 2(x^4 + y^4)$$

change its values at such points!

So: for very small values of x, y, we know that $x^4 \gg x^6, x^8$ and $y^4 \gg y^6, y^8$; therefore, very very close to the origin, our function is roughly just $-2(x^4+y^4)$, which is a upside-down parabola with a maximum at the origin. Therefore, we can see that this point is actually a local maximum, because (using our approximation) at all values very close to the origin that are not the origin, our function is roughly $-2(x^4 + y^4)$ and therefore quite decidedly < 0, its value at the origin. So $(0, 0)$ is a local maxima!

For the other values, we can do a similar (but more in-depth) analysis. For convenience's sake, let $g(z) = -z^8 + 4z^6 - 4z^4$; we can then write $f(x, y) = g(x) + g(y)$. By the same logic as above, for arbitrarily small values of z we can write $g(z)$ as approximately $-4z^4$, as $z^4 \gg z^6, z^8$ and thus the z^4 terms dominate the function $g(z)$.

In general, we can extend our observation above to an approximation of $(c+\epsilon)^n$ for any constant c, power n, and very small ϵ, by using the binomial theorem:

$$\begin{aligned}
(z + \epsilon)^n &= \sum_{i=0}^{n} \binom{n}{k} \cdot z^{n-k}\epsilon^k \\
&= z^n + n\epsilon z^{n-1} + \frac{n(n-1)}{2} z^{n-2}\epsilon^2 + (\text{terms scaled by } \epsilon^4) \\
&\approx z^n + n\epsilon z^{n-1} + \frac{n(n-1)}{2} z^{n-2}\epsilon^2 \\
\Rightarrow g(z + \epsilon) &= -(z + \epsilon_1)^8 + 4(z + \epsilon_1)^6 - 4(z + \epsilon_1)^4 \\
&\approx g(z) + (-8z^7\epsilon + 24z^5\epsilon - 16z^3\epsilon) + (-56z^6\epsilon_1^2 + 120z^4\epsilon_1^2 - 48z^2\epsilon_1^2) \\
\Rightarrow g(z + \epsilon) - g(z) &\approx (-8z^7\epsilon + 24z^5\epsilon - 16z^3\epsilon) + (-56z^6\epsilon_1^2 + 120z^4\epsilon_1^2 - 48z^2\epsilon_1^2).
\end{aligned}$$

This is kind of horrible-looking, but we can work with it. In particular, it tells us that at $z = \pm\sqrt{2}$, we have

$$\begin{aligned} g((\pm\sqrt{2})+\epsilon) - g((\pm\sqrt{2})) &\approx (-8(\pm\sqrt{2})^7\epsilon + 24(\pm\sqrt{2})^5\epsilon - 16(\pm\sqrt{2})^3\epsilon) \\ &\qquad + (-56(\pm\sqrt{2})^6\epsilon^2 + 120(\pm\sqrt{2})^4\epsilon^2 - 48(\pm\sqrt{2})^2\epsilon^2) \\ &= 0 + (-56\cdot 8\epsilon^2 + 120\cdot 4\epsilon^2 - 48\cdot 2\epsilon^2) \\ &= -64\epsilon^2, \end{aligned}$$

and at $z = \pm 1$ we have

$$\begin{aligned} g((\pm 1)+\epsilon) - g((\pm 1)) &\approx (-8(\pm 1)^7\epsilon + 24(\pm 1)^5\epsilon - 16(\pm 1)^3\epsilon) \\ &\qquad + (-56(\pm 1)^6\epsilon^2 + 120(\pm 1)^4\epsilon^2 - 48(\pm 1)^2\epsilon^2) \\ &= 0 + (-56\epsilon^2 + 120\epsilon^2 - 48\epsilon^2) \\ &= 16\epsilon^2. \end{aligned}$$

(Note that we used our earlier observation that $\pm 1, \pm\sqrt{2}$ are roots of $-8z^7 + 24z^5 - 16z^3$ to simplify the first parenthetical expression to 0.) In other words, small changes of $g(z)$ near 0 or $\pm\sqrt{2}$ yield decreases in our function, while small changes near ± 1 yield increases!

This lets us classify our remaining points: we can now see that the points $(0, \pm\sqrt{2})$, $(\pm\sqrt{2}, 0)$ are local maxima, and that the points $(0, \pm 1)$ and $(\pm 1, 0)$ are saddle points. Success!

Example. Take the vector field $V(x, y) = (x^2y^2, x^2 + y^2)$. Show that this vector field is neither the curl nor the gradient of any function.

Proof. This is relatively straightforward. To show that V is not the gradient of any vector field, we simply need to calculate the curl of V. If it is nonzero, then we know that it cannot be a gradient.

Because V is a vector field on $\mathbb{R}^2$, in order to calculate its curl we treat it like a vector field on $\mathbb{R}^3$ that has a 0 in its third component and does not depend on z. Then,

$$\begin{aligned} \operatorname{curl}(V) &= \left(\left(\frac{\partial V_3}{\partial y} - \frac{\partial V_2}{\partial z}\right), \left(\frac{\partial V_1}{\partial z} - \frac{\partial V_3}{\partial x}\right), \left(\frac{\partial V_2}{\partial x} - \frac{\partial V_1}{\partial y}\right)\right) \\ &= \left(0-0, 0-0, \left(2x - 2yx^2\right)\right), \end{aligned}$$

which is not identically equal to 0.

Similarly, we can show that V is not a curl by calculating its divergence: if this is nonzero, then V cannot be written as the curl of any vector field. We do this here:

$$\operatorname{div}(V) = \frac{\partial V_1}{\partial x} + \frac{\partial V_2}{\partial y} = 2xy^2 + 2y,$$

which is clearly nonzero. □