Potential Functions

E. L. Lady

In Calculus III so far, we have considered functions where the argument is a on-dimensional variable, which we usually denotes by t and often think of as time, and the values are two or three-dimensional:

$$\mathbb{R} \xrightarrow{\mathbf{r}} \mathbb{R}^3$$
$$t \mapsto \mathbf{r}(t) = (x(t), y(t), z(t)).$$

We usually see this sort of function in the role of a curve.

We have also considered functions where the argument is two or three-dimensional and the value is one-dimensional:

$$\mathbb{R}^2 \longrightarrow \mathbb{R}$$
$$(x,\, y) \mapsto f(x, y).$$

There are also applications, however, where both the argument and the value of a function are two or three-dimensional. Mostly commonly, we think of the argument of such a function as representing a point and the value as being a vector.

$$\mathbb{R}^2 \xrightarrow{\mathbf{F}} \mathbb{R}^2$$
$$(x, y) \mapsto P(x, y)\,\mathbf{i} + Q(x, y)\,\mathbf{j}.$$

A function of this sort is called a **vector field**.

Vector fields are well known in physics. The best known examples are electrostatic fields, magnetic fields, and gravitational fields. Any such field can be thought of as a function by which a vector corresponds to every point of two or three-dimensional space.

We have already seen some vector fields in this course, without stopping to note them as such. Namely, if $f(x, y)$ is any function of two variables, then at every point in 2-space (or at least at those points where f is differentiable), there exists a vector given by the gradient of f: $\nabla f(x, y)$. (Needless to say, the same thing is true for functions defined in 3-space.)

Not every vector field is the gradient of some function. However electrostatic and gravitational fields are, and this fact is extremely familiar and useful in physics. In fact, if $P(x,y)\,\mathbf{i}+Q(x,y)\,\mathbf{j}$ is an electrostatic field in 2-space, then there is a function $V(x,y)$ called *electrical potential* (usually measured in volts) such that

$$P\,\mathbf{i}+Q\,\mathbf{j}=-\nabla V.$$

Likewise for a gravitational field there is a function $p(x,y)$ such that $P\,\mathbf{i}+Q\,\mathbf{j}=-\nabla p$. (An object with mass m placed at a position (x,y) in the gravitational field will be acted on by a force given by $mP\,\mathbf{i}+mQ\,\mathbf{j}$. The scalar $mp(x,y)$ is usually called the *potential energy* of the object at the given location and is measured in units of work, such as centimeter-dynes.)

In general, if a vector field $P(x,y)\,\mathbf{i}+Q(x,y)\,\mathbf{j}$ is the gradient of a function $f(x,y)$, then $-f(x,y)$ is called a *potential function* for the field. When the vector field represents force, then an objected placed in the field initially at rest will move in such a way as to maximize the function $f(x,y)$ most rapidly, i. e. to minimize the potential most rapidly. Vector fields that have potential functions are called *conservative fields.* This is because, if the vector field itself represents force, then conservative fields are precisely the ones for which there exists a law of conservation of energy: i. e. if one moves an object around in the field in such a way that it returns to its original location, the net work done on the object by the field will be zero.

From a purely mathematical point of view, the problem here is as follows: Give a vector field $P(x,y)\,\mathbf{i}+Q(x,y)\,\mathbf{j}$, determine whether or not there exists a function $f(x,y)$ such that $\nabla f=P\,\mathbf{i}+Q\,\mathbf{j}$, and, if so, find such a function. (We will consider only the problem in two dimensions, to make things simpler. In three dimensions, things are only a little bit harder.)

In order to have $\nabla f=P\,\mathbf{i}+Q\,\mathbf{j}$, we need $\dfrac{\partial f}{\partial x}=P$ and $\dfrac{\partial f}{\partial y}=Q$. If we look at only the first of these conditions, it seems fairly obvious that the only answer is to choose

$$f(x,y)=\int P(x,y)\,dx.$$

Here, the integral has to be interpreted as what ought to be (but is usually not) called a *partial integral*, which is to say that in doing the integration, y is treated as a constant.

Since this approach ignores Q altogether, it seems unlikely that it could give the right answer. And in fact, in most cases it will not. For instance, if we have

$$P(x,y)=y, \qquad Q(x,y)=-x,$$

then we get

$$f(x,y) = \int y\,dx = xy + C$$

(since y is treated as a constant when integrating). Testing this, we get $\dfrac{\partial f}{\partial x} = P$, as required. (In fact, this part could not have failed, since we obtained f by integrating P with respect to x.) On the other hand, $\dfrac{\partial f}{\partial y} = x$ but $Q = -x$, so $f(x,y)$ does not work in this respect.

In fact, for this particular example, as we shall see, it is impossible to find any function $f(x,y)$ that works for both x and y. And, when one thinks about it, this is not surprising. Because in looking for $f(x,y)$ such that

$$\frac{\partial f}{\partial x} = P, \qquad \frac{\partial f}{\partial y} = Q$$

we are attempting to solve **two** equations for only one unknown f. Now these equations involve functions rather than numbers, so we can't automatically assume that the usual rules of algebra apply. Nonetheless, when we have more equations than unknowns it's not surprising that we are unable to find a solution.

An obvious thing to do at this point is to simply give up on this kind of problem. A more intelligent approach, though, is to realize that some of these problems are solvable and some are not, and we need a way of deciding which is the case.

A good way to get insight in a case like this is to deviseconstruct an example by starting with the answer first. So suppose we start with $f(x,y) = \sin x^2 y$. Then set

$$P(x,y) = \frac{\partial f}{\partial x} = 2xy\cos x^2 y, \qquad Q(x,y) = \frac{\partial f}{\partial y} = x^2\cos x^2 y.$$

If we momentarily forget the known answer and use the method above, we get

$$f(x,y) = \int 2xy\cos x^2 y = \sin x^2 y + C.$$

(The integral is an easy substitution, setting $u = x^2 y$, $du = 2xy\,dx$. It only looks confusing because one has to remember to treat y as a constant.)

One really ought to try several more examples of this sort before coming to any conclusion, but it seems as though the method given produces the correct answer whenever any answer exists. (In fact, we shall see that this is *almost* true. We need to add one more refinement.)

So there is one clear approach to the problem of finding an integral for a vector field $P(x,y)\,\mathbf{i} + Q(x,y)\,\mathbf{j}$. Namely, solve for $f(x,y)$ by integrating P, and if this doesn't work (after trying a minor adjustment to be described below), then there is no answer.

This is actually a fairly workable approach. It's just that it's a bit inelegant.

To get an even better answer, go back to the example $P = 2xy\cos x^2y$, $Q = x^2\cos x^2y$, and compare what one might (but seldom does) refer to as the "cross derivatives," namely $\dfrac{\partial P}{\partial y}$ and $\dfrac{\partial Q}{\partial x}$. We get (by applying the product rule)

$$\frac{\partial P}{\partial y} = \frac{\partial}{\partial y}(2xy\cos x^2y) = 2x\cos x^2y - 2x^3y\sin x^2y$$

$$\frac{\partial Q}{\partial x} = \frac{\partial}{\partial x}(x^2\cos x^2y) = 2x\cos x^2y - 2x^3y\sin x^2y.$$

Trying a few more examples will convince you that this is apparently not a coincidence. This following principle seems to be true.

If there exists a function $f(x,y)$ *such that* $\nabla f = P\,\mathbf{i} + Q\,\mathbf{j}$, *then* $\dfrac{\partial P}{\partial y} = \dfrac{\partial Q}{\partial x}$.

It's easy to see why this must be true. In fact, if $\nabla f = P\,\mathbf{i} + Q\,\mathbf{j}$ then $P = \dfrac{\partial f}{\partial x}$ and $Q = \dfrac{\partial f}{\partial y}$ and so

$$\frac{\partial P}{\partial y} = \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right) = \frac{\partial^2 f}{\partial y\,\partial x}$$

$$\frac{\partial Q}{\partial y} = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right) = \frac{\partial^2 f}{\partial x\,\partial y}.$$

It is known that if these mixed partials here are continuous (as is almost always the case at points where they exist), then they will be equal.

The principle stated above is a **necessary condition** for the problem $\nabla f = P\,\mathbf{i} + Q\,\mathbf{j}$ to have a solution. The reasoning given so far does not show that it is also a sufficient condition.

In other words, it might be possible for $\dfrac{\partial P}{\partial y} = \dfrac{\partial Q}{\partial x}$ to be true and yet for there not to exist any function $f(x, y)$ with $\nabla f = P\,\mathbf{i} + Q\,\mathbf{j}$. But in fact, the condition we have found is (more or less) sufficient as well. We just have to give a different explanation of this fact (as well as explaining what the "more or less" means.)

This involves considering the question:

If $f(x, y) = \displaystyle\int P\,dx$ and $\dfrac{\partial P}{\partial y} = \dfrac{\partial Q}{\partial x}$, why should $\dfrac{\partial f}{\partial y} = Q$? (There should be no doubt about why $\dfrac{\partial f}{\partial x} = P$.)

This is not so hard to see. If $f = \displaystyle\int P\,dx$ then

$$\begin{aligned}\frac{\partial f}{\partial y} &= \frac{\partial}{\partial y}\int P\,dx \\ &= \int \frac{\partial P}{\partial y}\,dx \\ &= \int \frac{\partial Q}{\partial x}\,dx.\end{aligned}$$

At this point, it is tempting to finish the problem by writing $\displaystyle\int \frac{\partial Q}{\partial x}\,dx = Q$, showing that $\dfrac{\partial f}{\partial y} = Q$ and thus finishing the proof. This step is slightly shaky, however, for the reason that integration involves an arbitrary constant. What we really seem to have shown is that $\dfrac{\partial f}{\partial y}$ and Q differ by a constant.

At this point, it seems that only a technicality stands in the way of what we want, and many students will be willing to take the rest of the proof on faith. However, this technicality is actually a *little* bit bigger than it appears, and it brings up an important point that needs to be considered in solving problems in practice.

Consider what happens if we make a slight change to the problem considered above. Let

$$P(x, y) = 2xy\cos x^2 y, \qquad\qquad Q(x, y) = y^3 + x^2\cos x^2 y.$$

Since the problem has been changed, the previous solution
$f(x,y) = \int 2xy\cos x^2y = \sin x^2y + C$ clearly will no longer work.

Notice, however, that since the change made to Q has involved adding a function of y, the partial derivative $\dfrac{\partial Q}{\partial x}$ is not changed, and so the condition $\dfrac{\partial Q}{\partial x} = \dfrac{\partial P}{\partial y}$ still holds. And, astonishingly enough, the new problem does have a solution, namely

$$f(x,y) = \frac{1}{4}\,y^4 + \sin x^2y + C.$$

The reason that this new solution works is that the change made to the old solution only involved y and hence did not affect $\dfrac{\partial f}{\partial x}$.

To understand this better, we need to think about the integral

$$f(x,y) = \int P(x,y)\,dx$$

more carefully. As already mentioned, this is a *partial integral,* where during the integration y is treated as if it were a constant. But since this is true, when we do the integral we need to add not an arbitrary constant, as one would normally do, but an arbitrary function of y. For instance, in the problem in question, one needs to write

$$f(x,y) = \int 2xy\cos x^2y\,dx = \sin x^2y + \varphi(y).$$

One then needs to choose $\varphi(y)$ in such a way that the desired condition $\dfrac{\partial f}{\partial y} = Q$ will be true.

For practical purposes, this is usually not hard to do, provided that one starts out with functions $P(x,y)$ and $Q(x,y)$ which have reasonable formulas. However it brings up an important theoretical point, and it is in fact this fine point which accounts for the rather arcane proof one finds in books for the fact that a function $f(x,y)$ exists whenever $\dfrac{\partial P}{\partial y} = \dfrac{\partial Q}{\partial x}$.

The point is that writing $f = \int P\,dx$ doesn't really make sense, because $\int P\,dx$ isn't actually a well defined function, since integration involves an arbitrary constant. Usually this is not a problem, but in this case the "constant" of integration actually depends on y, and it becomes essential to determine this "constant" in such a way that $f(x, y)$ becomes a reasonable function of y (continuous, differentiable) as well as of x.

In most practical problems, where one has specific formulas, this is not a major problem. However in order to write down a valid proof, one needs to finesse this point. The proofs one finds in books accomplish this by using a definite integral instead of an indefinite integral.

The theorem one finds in books also says something about working in a "simply connected" region. Basically, a region in the plane is simply connected if it doesn't have any holes in it. In other words, the requirement is that whenever you draw a closed curve (not crossiong itself) in the region, everything inside that curve is still in the region. (The definition of "simply connected" is not quite this simple for a region in three-space. Most regions that one encounters in practice in three space are simply connected, even when they have holes in them.)

For instance, if we delete the origin from the xy-plane then the resulting region is no longer simply connected, since if we draw a circle around the origin, then the entire curve lies in the given region, but one of the points inside (i. e. the origin) does not.

Now the explanation given above doesn't suggest any reason why the geometry of the domain of a vector field should cause problems in solving for a function that has that field as its gradient. This is, in fact, one indication that the explanation given above is overly simplistic. And yet that explanation is **almost** correct.

To understand this issue, consider another example (which unfortunately, has to be a little more complicated than the previous ones). Let

$$P = \frac{-y}{x^2 + y^2}, \qquad Q = \frac{x}{x^2 + y^2}.$$

Notice that the domain of the vector field $P\,\mathbf{i} + Q\,\mathbf{j}$ consists of the whole plane except the origin (where the denominator becomes 0) and is thus not simply connected. Computing $\dfrac{\partial P}{\partial y}$ and $\dfrac{\partial Q}{\partial x}$ is unpleasant, but an application of the quotient rule shows that

$$\frac{\partial P}{\partial y} = \frac{y^2 - x^2}{x^2 + y^2} = \frac{\partial Q}{\partial x}.$$

Therefore it makes sense to try to find a function f such that $\nabla f = P\,\mathbf{i} + Q\,\mathbf{j}$. In order to make the answer slightly nicer I will compute f in this case by integrating Q rather than P, although the same difficulty arises in either case. Thus we write

$$f(x,y)=\int Q\,dy=\int \frac{x\,dy}{x^2+y^2}\,.$$

This integral is not actually very difficult if one remembers that x is to be treated as if it were a constant. However it's importat to notice that $x=0$ is a special case:

$$f(0,y)=\int Q(0,y)\,dy=\int \frac{0}{y^2}\,dy=\int 0\,dy=C,$$

where C is a constant. It is natural to assume that one is least likely to get into trouble by choosing $C=0$, however this turns out not to be the case.

For $x\neq 0$, we use the standard formula

$$\int \frac{dy}{a^2+y^2}=\frac{1}{a}\tan^{-1}\frac{y}{a}.$$

This yields (for $x\neq 0$),

$$f(x,y)=\int \frac{x\,dy}{x^2+y^2}=\tan^{-1}\frac{y}{x}\,.$$

Now this example was chosen because it has a zinger, showing how the geometry of the domain of a vector field can screw up the calculation of a potential function. And yet when we check the function $f(x,y)$, we see that it does in fact work.

$$\frac{\partial f}{\partial x}=\frac{\partial}{\partial x}\tan^{-1}\frac{y}{x}=\frac{-y}{x^2}\left(\frac{1}{1+(y/x)^2}\right)=\frac{-y}{x^2+y^2}=P$$

$$\frac{\partial f}{\partial y}=\frac{1}{x}\left(\frac{1}{1+(y/x)^2}\right)=\frac{x}{x^2+y^2}=Q.$$

This seems just fine, and in fact it is just fine, as far as it goes. Except that it doesn't really deal with the special case $x=0$. In fact, we decided to choose $f(0,y)=0$. But this makes the function $f(x,y)$ discontinuous, since for points close to the y-axis, $\tan^{-1}\dfrac{y}{x}$ is not close to 0. In fact, assuming that y and x are both positive, and that x approaches 0 but y does not, then y/x approaches $+\infty$, and it's not hard to see that $\tan^{-1}(y/x)$ approaches $\pi/2$. (We don't need to worry about x and y simultaneously approaching 0, because we knew from the beginning that the origin was a singularity for the vector field, so it's unrealistic to expect things to work well there.)

What this shows is that the problem of doing the indefinite integral in a consistent way to produce a function that is nice with respect to both x and y is not always completely trivial. What we have just seen suggests that to get the function to be continuous, we need to compute

$$f(0, y) = \int 0\, dy = \frac{\pi}{2}.$$

This is not a serious problem, though. It's just a matter of choosing a different constant of integration.

What *is* serious, though, is that this fix doesn't work. If we choose $f(0, y) = \pi/2$, then we have a problem when we look at points (x, y) in the second quadrant close to the y axis. In this case, y is positive and x is negative and close to 0, so that y/x is close to $-\infty$, and $\tan^{-1}(y/x)$ is close to $-\pi/2$. In other words the function $f(x, y)$ as we have defined it is discontinuous when (x, y) cross the y-axis from the first quadrant into the second quadrant, since it jumps from $\pi/2$ to $-\pi/2$. (The same thing happens when (x, y) crosses from the third quadrant into the fourth.)

This is a serious problem, but there is still a possible fix. In fact, if one defines

$$f(x, y) = \begin{cases} \tan^{-1} \dfrac{y}{x} & \text{if } x > 0 \\[2ex] \dfrac{\pi}{2} & \text{if } x = 0 \\[2ex] \pi + \tan^{-1} \dfrac{y}{x} & \text{if } x < 0 \end{cases}$$

then, weird as this definition may seem, the resulting function is actually continuous on the positive part of the y-axis, since for points near the y-axis, either to the left or to the right of it, with y-positive, $f(x, y)$ is close to $\dfrac{\pi}{2}$.

This fix fails, though, on the negative y-axis (although it works just fine everywhere else). For if x is positive and approaches 0 and y is negative, then y/x approaches $-\infty$ and $f(x, y) = \tan^{-1}(y/x)$ approaches $-\pi/2$. But if x is negative, then by definition $f(x, y) = \pi + \tan^{-1}(y/x)$, and as x approaches 0, y/x approaches $+\infty$ (both x and y are negative), and $f(x, y)$ approaches $\pi + \pi/2 = 3\pi/2$. Thus we get different limits when we approach the negative y-axis from the right and from the left. (Furthermore, the actual value of $f(0, y)$ has been assigned as $\pi/2$, which is not the same as either of these two limits!) (It helps a whole lot to draw a picture here. Unfortunately, I haven't yet discovered how to do that in TeX.)

There is no completely happy ending to this story. No matter what one tries, there is no way of defining a function $f(x,y)$ in the entire plane (even with the origin deleted) that satisfied the conditions $\partial P/\partial x = P$ and $\partial Q/\partial y = Q$.

In fact, if one thinks in terms of polar coordinates, one can see is that basically what one wants is to set $f(x,y) = \theta$. But of course there is no way of defining θ as a continuous well-defined variable in the whole plane. If one circles the origin, then either there must be a jump somewhere in the value of θ, or one winds up with an inconsistency — trying to give θ two different values at the same point.

One could have predicted these behavior by looking at the original vector field. If one thinks in terms of polar coordinates, one has

$$P\,\mathbf{i} + Q\,\mathbf{j} = \frac{-y\,\mathbf{i}}{x^2+y^2} + \frac{x\,\mathbf{j}}{x^2+y^2} \quad = \quad \frac{-1}{r}(-\sin\theta\,\mathbf{i} + \cos\theta\,\mathbf{j}).$$

Now if $\mathbf{r} = x\,\mathbf{i} + y\,\mathbf{j}$ is the radius vector, one sees that $P\,\mathbf{i} + Q\,\mathbf{j}$ is perpendicular to $\mathbf{r}$. And if one draws the vector field, one sees that the drawing seems to describe a flow moving in circles around the origin counter-clockwise. In fact, if one moves around a circle with radius one centered at the origin, one will observe that the field $P\,\mathbf{i} + Q\,\mathbf{j} = -\sin\theta\,\mathbf{i} + \cos\theta\,\mathbf{j}$ (since $r = 1$ on this circle) is the same as the tangent vector to this circle. This means that if there were a function $f(x,y)$ such that $\nabla f(x,y) = P\,\mathbf{i} + Q\,\mathbf{j}$, then this function would be strictly increasing all the way around the circle. But this is not possible, since when one had gone all the way around the circle and returned to the original point, the function would no longer have the same value.

The correct theorem (which we have not actually proved) is as follows:

If Ω is a **simply connected** region in the plane and $P(x,y)$ and $Q(x,y)$ are functions which have continuous partial derivatives in all of Ω, then there exists a function $f(x,y)$ such that $\nabla f = P\,\mathbf{i} + Q\,\mathbf{j}$ if and only if

$$\frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x}$$

at all points in Ω.

Even the region Ω is not simply connected, there cannot exist such a function $f(x,y)$ if $\dfrac{\partial P}{\partial y} \neq \dfrac{\partial Q}{\partial x}$ at any point in Ω.