Multivariable Calculus

Toby Bartels

Math-2080-es31

2017 Winter

Welcome to multivariable Calculus! Here are my supplemental notes for this course, giving alternative ways to think about some things, practical advice, and sometimes more theoretical detail.

This does not cover everything that you need to know (although it covers a lot); you should also have the official course textbook, which is the 3rd Edition of *University Calculus: Early Transcendentals* by Hass et al published by Addison–Wesley (Pearson). There are also some references in these notes to that textbook.

Vectors

The first part of this course involves working with points and vectors in multidimensional space. There are not really any new ideas of Calculus itself here, but the setting may be new.

Points

In this class, we look at spaces with up to 3 dimensions, but most of the ideas in this course (and the next) continue to make sense in spaces with any whole number of dimensions. Although spaces with more than 3 dimensions are difficult to visualize, since we're used to living in a 3-dimensional space, they make perfect sense mathematically. Furthermore, whenever you're trying to keep track of 4 or more independent quantities at once, then you need the mathematics of a space with 4 or more dimensions, whether or not you choose to visualize that space geometrically.

If we assign rectangular coordinates to a space of n dimensions, then the result is called $\mathbf{R}^n$ (or $\mathbb{R}^n$); in particular, a coordinate space of 1 dimension is $\mathbf{R}^1$ or simply $\mathbf{R}$, which is the set of real numbers, or (thinking geometrically) the real number line. You can call the coordinates whatever you like, but it's most common to use x (or sometimes t) as the coordinate in $\mathbf{R}$; then to use x and y as the coordinates in $\mathbf{R}^2$; then x, y, and z in $\mathbf{R}^3$; and finally x_1, x_2, ..., and x_n in $\mathbf{R}^n$ generally. But there are other systems; as long as you list n independent variables in a row, then you have a valid list of coordinates for $\mathbf{R}^n$.

A **point** in $\mathbf{R}^n$ may be denoted by listing the values of its coordinates in order, separated by commas and optionally surrounded by grouping parentheses. Thus, (x) or (more commonly) x gives a point in the real line $\mathbf{R}$, while (x, y) gives a point in the coordinate plane $\mathbf{R}^2$, (x, y, z) gives a point in the coordinate space $\mathbf{R}^3$, and $(x_1, x_2, \ldots, x_n)$ gives a point in $\mathbf{R}^n$ (which is the most general case).

Sometimes it's nice to have a way to refer to a point in any number of dimensions without having to write a long list with dots in it; then I usually write P for the point. Thus, in 1 dimension, $P = x$; in 2 dimensions, $P = (x, y)$; in 3 dimensions, $P = (x, y, z)$; and in n dimensions, $P = (x_1, x_2, \ldots, x_n)$. So for example, if I say that $P = (2, 3, 5)$, then this is the same as saying that $x = 2$, $y = 3$, and $z = 5$.

It's traditional to use uppercase letters to name points, as I just did. Another tradition is to leave out the equality sign when naming points; so instead of writing $P = (2, 3, 5)$ as I did above, people often just write $P(2, 3, 5)$. I think that this is a terrible convention, so I won't follow it, but you will see it sometimes, even in the textbook.

Vectors

A **vector** is a movement between points. For example, to move in the plane from the point $(2, 3)$ to the point $(3, 1)$, you move 1 unit to the right (in the positive x direction) and 2 units downwards (in the negative y direction). This movement —1 unit to the right and 2 units downwards— is a vector.

A vector in $\mathbf{R}^n$ has the same amount of information as a point there: n real numbers. For this reason, people sometimes write a vector using the same notation as they use to write a point. For example, the vector from the previous paragraph could be written as $(1, -2)$, the same notation as used for the point $(1, -2)$. When referring to a vector, $(1, -2)$ means a movement 1 unit to the right and 2 units downards; when referring to a point, $(1, -2)$ means the point that lies 1 unit to the right and 2 units downwards from the origin.

However, a vector is not the same thing as a point, and so people often use different notation instead. Common notations for the vector that I've been talking about include $[1, -2]$ and $\langle 1, -2 \rangle$. I will use the last of these, since that is used in the textbook. (There is another notation, which the book uses even more often than $\langle 1, -2 \rangle$, and that is $\mathbf{i} - 2\mathbf{j}$. However, I'll save that for later.) The terminology for these numbers is also different; while 1 and -2 are the *coordinates* of the point $(1, -2)$, we say that 1 and -2 are the **components** of the vector $\langle 1, -2 \rangle$.

Whereas a point tells you a location, a vector tells you only about the motion and nothing about the location. So the vector from $(2, 3)$ to $(3, 1)$ is the same vector as, say, the vector from $(-2, 7)$ to $(-1, 5)$. In both cases, the motion is 1 unit to the right and 2 units downwards, so the vector is $\langle 1, -2 \rangle$.

Motion on a number line corresponds arithmetically to addition. For example, if you start at the number 2 on a number line and move 4 units to the right, then you end up at the number 6, and we represent this fact in arithmetic as $2 + 4 = 6$. Similarly, if you start at $(2, 3)$ and move according to the vector $\langle 1, -2 \rangle$, then you end up at $(3, 1)$, and we represent this fact in arithmetic as $(2, 3) + \langle 1, -2 \rangle = (3, 1)$. So you can add a point and a vector to get another point. Or from another perspective, we could write $6 - 2 = 4$, and similarly $(3, 1) - (2, 3) = \langle 1, -2 \rangle$. So one way to describe a vector is to say that it's what you get when you subtract two points. The textbook doesn't talk about arithmetic with points and vectors like this; it does talk about calculating the vector from one point to another or calculating the point reached from another point by following a given vector, but it doesn't refer to these operations as subtraction and addition. Nonetheless, that's exactly what they are.

The rules for these calculations are very straightforward: you add or subtract corresponding coordinates and components. That is, to get the first coordinate of the sum, you add the first coordinate of the original point and the first component of the vector, and similarly for the second coordinate; or when you subtract two vectors, you subtract the first coordinates of the two points to get the first component of the difference, and similarly for the second component. So you can write out the calculations in full thus:

$$(2, 3) + \langle 1, -2 \rangle = (2 + 1, 3 - 2) = (3, 1);$$
$$(3, 1) - (2, 3) = \langle 3 - 2, 1 - 3 \rangle = \langle 1, -2 \rangle.$$

Here are general formulas for this rule in any number of dimensions:

$$(a_1, a_2, \ldots, a_n) + \langle v_1, v_2, \ldots, v_n \rangle = (a_1 + v_1, a_2 + v_2, \ldots, a_n + v_n);$$
$$(b_1, b_2, \ldots, b_n) - (a_1, a_2, \ldots, a_n) = \langle b_1 - a_1, b_2 - a_2, \ldots, b_n - a_n \rangle.$$

When I use P to denote a generic point, I'll use ΔP to denote a generic vector. Here, the uppercase Greek letter Delta, 'Δ', which stands for 'difference', is commonly used to indicate the amount by which the value of some quantity changes. (Think of $\Delta y / \Delta x$ for the slope of a line.) That is,

$$\Delta P = P_1 - P_0,$$

or

$$\Delta P = \langle \Delta x_1, \Delta x_2, \ldots, \Delta x_n \rangle.$$

When you give a vector a name of its own, however, it's common to use a boldface lowercase letter, such as $\mathbf{u}$ or $\mathbf{v}$. Thus, if I use $\mathbf{v}$ to refer to the vector that I've been using as an example throughout this section, then I would write $\mathbf{v} = \langle 1, -2 \rangle$. In handwriting, you can write a little arrow over the letter instead, to produce something like $\vec{v}$; other common conventions are to underline or overline vectors, producing symbols such as $\underline{v}$ or $\overline{v}$. On the other hand, it's OK to just write v if you want. The meaning of any symbol that you use should be clear from the context that you provide; in particular, the context should make clear whether a symbol refers to a number, function, point, vector, or whatever, regardless of whatever fancy fonts or decorations you may or may not use.

Arithmetic with vectors

Besides adding vectors to points and subtracting points to get a vector, you can also do arithmetic within the world of vectors itself. If $\mathbf{u}$ and $\mathbf{v}$ are vectors in n dimensions, both representing some motion within $\mathbf{R}^n$, then $\mathbf{u} + \mathbf{v}$ represents the motion of $\mathbf{u}$ followed by the motion of $\mathbf{v}$. This is consistent with how addition of motions works on a number line; for example, if you move 4 units to the right and then move 3 units to the right, then overall you're moving $4 + 3 = 7$ units to the right.

If $\mathbf{v}$ is a vector, then $-\mathbf{v}$ is the vector representing the opposite motion. Again, this matches arithmetic on a number line; the opposite of moving 4 units to the right is moving 4 units to the left, which is represented by the number -4. Then $\mathbf{u} - \mathbf{v}$ just means $\mathbf{u} + (-\mathbf{v})$.

You calculate these by the same principles as arithmetic between points and vectors. For example, to add $\langle 1, -2\rangle$ and $\langle 3, 5\rangle$, you simply add the corresponding components:

$$\langle 1, -2\rangle + \langle 3, 5\rangle = \langle 1 + 3, -2 + 5\rangle = \langle 4, 3\rangle.$$

And this should make sense; if you move 1 unit to the right and 2 units downwards, then move 3 units to the right and 5 units upwards, then overall you're moving 4 units to the right and 3 units upwards. Similarly,

$$\langle 1, -2\rangle - \langle 3, 5\rangle = \langle 1 - 3, -2 - 5\rangle = \langle -2, -7\rangle.$$

That is, if you move 1 unit to the right and 2 units downards and then move the opposite of 3 units to the right and 5 units upwards (which is 3 units to the left and 5 units downwards), then overall you're moving 2 units to the left and 7 units downwards. Here are the general formulas in $\mathbf{R}^n$:

$$\langle u_1, u_2, \ldots, u_n\rangle + \langle v_1, v_2, \ldots, v_n\rangle = \langle u_1 + v_1, u_2 + v_2, \ldots, u_n + v_n\rangle;$$
$$\langle u_1, u_2, \ldots, u_n\rangle - \langle v_1, v_2, \ldots, v_n\rangle = \langle u_1 - v_1, u_2 - v_2, \ldots, u_n - v_n\rangle.$$

Besides adding and subtracting vectors, you can multiply or divide them by real numbers. For example, if $\mathbf{v}$ is a vector representing some motion, then $2\mathbf{v}$ represents doing that motion twice, $1/2\,\mathbf{v}$ or $\mathbf{v}/2$ represents performing half of that motion, $-2\mathbf{v}$ represents making the opposite motion twice, and so on. You calculate these by multiplying each component by that same real number; for example,

$$2\langle 1, -2\rangle = \langle 2(1), 2(-2)\rangle = \langle 2, -4\rangle,$$
$$\frac{1}{2}\langle 1, -2\rangle = \left\langle \frac{1}{2}(1), \frac{1}{2}(-2)\right\rangle = \left\langle \frac{1}{2}, -1\right\rangle \text{ or}$$
$$\frac{\langle 1, -2\rangle}{2} = \left\langle \frac{1}{2}, \frac{-2}{2}\right\rangle = \left\langle \frac{1}{2}, -1\right\rangle, \text{ and}$$
$$-2\langle 1, -2\rangle = \langle -2(1), -2(-2)\rangle = \langle -2, 4\rangle.$$

Here are the general formulas in $\mathbf{R}^n$:

$$a\langle v_1, v_2, \ldots, v_n\rangle = \langle av_1, av_2, \ldots, av_n\rangle;$$
$$\frac{\langle v_1, v_2, \ldots, v_n\rangle}{a} = \left\langle \frac{v_1}{a}, \frac{v_2}{a}, \ldots, \frac{v_n}{a}\right\rangle \text{ for } a \neq 0.$$

This operation is called **scalar multiplication** (or *scalar division*), because geometrically it amounts to changing the scale used to measure the vector (at least when the real number in question is positive). As a result of this, numbers are often called **scalars** when working with vectors, even though the word 'number' would work perfectly well.

More generally, you can take any homogeneous linear expression (that is a linear expression without a constant term) in any number of variables, replace the variables with vectors, and get a legitimate operation on vectors. Such an operation is called, in general, a **linear combination**. For example, $2\mathbf{u} + 3\mathbf{v} - 5\mathbf{w}$ is a linear combination of the vectors $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$. Geometrically, this represents moving twice according to $\mathbf{u}$, then moving 3 times according to $\mathbf{v}$, and moving 5 times the reverse of the motion given by $\mathbf{w}$.

Still more generally, you can replace the variables with points or vectors; if the sum of the coefficients on the points is 0, then the result is a vector, and if the sum of the coefficients on the points is 1, then the result is a point. For example, if A, B, and C are points, while $\mathbf{u}$ and $\mathbf{v}$ are vectors, then $2A - 3B + 2C + 4\mathbf{u} - 5\mathbf{v}$ is a point (because $2 - 3 + 2 = 1$), while $2A - 3B + C + 4\mathbf{u} - 5\mathbf{v}$ is a vector (because $2 - 3 + 1 = 0$). Geometrically, $2A - 3B + 2C + 4\mathbf{u} - 5\mathbf{v}$ means the point that you reach by starting at A, moving as you would move to get to A from B, then moving twice as you would move to get to C from B, then moving 4 times according to $\mathbf{u}$, and moving 5 times the reverse of the motion given by $\mathbf{v}$. (That is, think of it as $A + (A - B) + 2(C - B) + 4\mathbf{u} - 5\mathbf{v}$.) Similarly, $2A - 3B + C + 4\mathbf{u} - 5\mathbf{v}$ is the motion consisting of moving twice as you would move to get to A from B, then moving as you would move to get to C from B,

then moving 4 times according to $\mathbf{u}$, and moving 5 times the reverse of the motion given by $\mathbf{v}$. (That is, think of it as $2(A-B)+(C-B)+4\mathbf{u}-5\mathbf{v}$.)

Another example of a point is $1/3\,A+1/3\,B+1/3\,C$, which is the average of the 3 points. If you think of this as $A+2/3\,(B-A)+1/3\,(C-B)$, then you can describe this in terms similar to those of the previous examples, but in this case it's probably better to think of it directly as an average.

If the sum of the coefficients on the points is neither 1 nor 0, then there is no direct geometric interpretation of the linear combination, but you can still perform calculations with such things; they basically represent internal parts of a larger calculation, such as the $2A-3B$ that begins some of the examples above.

All of the usual algebraic identities apply to linear combinations of points and vectors. For example, $\mathbf{u}+\mathbf{v}=\mathbf{v}+\mathbf{u}$, $(A+\mathbf{u})+\mathbf{v}=A+(\mathbf{u}+\mathbf{v})$, $2(\mathbf{u}+\mathbf{v})=2\mathbf{u}+2\mathbf{v}$, and so on. Although you can prove these geometrically, the simplest way to verify them is to do so component by component; then they reduce to identities about real numbers.

You could try multiplying and dividing vectors by each other using the same method of calculation as you use for adding and subtracting them, component by component. People do this sometimes, but there's no geometric interpretation of these operations, neither directly nor as part of a larger calculation with a geometric interpretation. So we won't be doing that. Instead, we'll see some other methods of multiplying vectors later on.

The **zero vector**, denoted $\mathbf{0}$, represents no motion at all. It's general formula in $\mathbf{R}^n$ is

$$\mathbf{0}=\langle 0,0,\ldots,0\rangle.$$

It obeys algebraic rules analogous to those obeyed by the real number 0, such as $\mathbf{0}+\mathbf{v}=\mathbf{v}$, $\mathbf{v}-\mathbf{v}=\mathbf{0}$, and $A+\mathbf{0}=A$. (The last of these demonstrates what it means to say that $\mathbf{0}$ represents no motion at all; you start at the point A, do nothing, and wind up still at A.)

The standard basis vectors

There are some other special symbols for special vectors, and these lead to another general system of notation for vectors (and points).

In $\mathbf{R}^2$, there are 2 **standard basis vectors**, $\mathbf{i}$ and $\mathbf{j}$:

$$\mathbf{i}=\langle 1,0\rangle,\ \mathbf{j}=\langle 0,1\rangle.$$

In $\mathbf{R}^3$, there are 3 of them:

$$\mathbf{i}=\langle 1,0,0\rangle,\ \mathbf{j}=\langle 0,1,0\rangle,\ \mathbf{k}=\langle 0,0,1\rangle.$$

In $\mathbf{R}^n$, there is a shift in the usual notation:

$$\mathbf{e}_1=\langle 1,0,0,\ldots,0\rangle,\ \mathbf{e}_2=\langle 0,1,0,0,\ldots,0\rangle,\ \ldots,\ \mathbf{e}_n=\langle 0,0,\ldots,0,0,1\rangle.$$

The value of this is that any vector can be written as a unique linear combination of the standard basis vectors:

$$\begin{aligned}\langle a,b\rangle &= a\mathbf{i}+b\mathbf{j};\\ \langle a,b,c\rangle &= a\mathbf{i}+b\mathbf{j}+c\mathbf{k};\\ \langle a_1,a_2,\ldots,a_n\rangle &= a_1\mathbf{e}_1+a_2\mathbf{e}_2+\cdots+a_n\mathbf{e}_n.\end{aligned}$$

Work out the right-hand sides of these and see for yourself that you get the left-hand side. (It's a little annoying that $\mathbf{i}$ and $\mathbf{j}$ are ambiguous, but as long as you know whether they're supposed to be in $\mathbf{R}^2$ or in $\mathbf{R}^3$, then you know what they mean.)

If a component of a vector happens to be 1, then you can leave it out of the expression in the standard basis vectors; if the component is negative, then you use subtraction instead of addition; if the component is 0, then you leave that term out entirely. For example, $\langle 1,-2\rangle=1\mathbf{i}+(-2)\mathbf{j}=\mathbf{i}-2\mathbf{j}$. In $\mathbf{R}^3$, $\langle 1,-2,0\rangle$ is also written $\mathbf{i}-2\mathbf{j}$, because the component on $\mathbf{k}$ is 0.

You can now do arithmetic with vectors by following the ordinary rules of algebra and leaving the symbols for the standard basis vectors alone. For example, instead of $\langle 1, -2\rangle + \langle 3, 5\rangle = \langle 4, 3\rangle$, you calculate

$$(\mathbf{i} - 2\mathbf{j}) + (3\mathbf{i} + 5\mathbf{j}) = (1 + 3)\mathbf{i} + (-2 + 5)\mathbf{j} = 4\mathbf{i} + 3\mathbf{j}.$$

Similarly, instead of $2\langle 1, -2\rangle = \langle 2, -4\rangle$, you calculate

$$2(\mathbf{i} - 2\mathbf{j}) = 2\mathbf{i} - 2(2\mathbf{j}) = 2\mathbf{i} - 4\mathbf{j}.$$

You can even extend this notation to points by introducing O for the origin of the coordinate system. That is,

$$\mathrm{O} = (0, 0, \ldots, 0)$$

in $\mathbf{R}^n$. Then any point can be described by starting at the origin and moving along a vector whose components are the coordinates of that point; for example, $(2, 3) = \mathrm{O} + \langle 2, 3\rangle = \mathrm{O} + 2\mathbf{i} + 3\mathbf{j}$. Then you can again do calculations using the rules of algebra; for example, instead of $(2, 3) + \langle 1, -2\rangle = (3, 1)$, you calculate

$$(\mathrm{O} + 2\mathbf{i} + 3\mathbf{j}) + (\mathbf{i} - 2\mathbf{j}) = \mathrm{O} + (2 + 1)\mathbf{i} + (3 - 2)\mathbf{j} = \mathrm{O} + 3\mathbf{i} + \mathbf{j}.$$

The textbook uses this notation for vectors most of the time, although it continues to use a list of coordinates with commas for points, which it has to do since it never refers directly to addition of points and vectors.

Lengths and angles

In many situations, we want to refer to the distance between two points, or equivalently to the length of a vector. This goes by several names; in general, the **length**, **magnitude**, or **norm** of a vector in $\mathbf{R}^n$ is

$$\|\langle v_1, v_2, \ldots, v_n\rangle\| = \sqrt{{v_1}^2 + {v_2}^2 + \cdots + {v_n}^2}.$$

Here I've denoted the length of a vector $\mathbf{v}$ as $\|\mathbf{v}\|$, although the textbook writes this as simply $|\mathbf{v}|$ instead. As a statement about distances, this is the n-dimensional generalization of the Pythagorean Theorem.

One basic algebraic property of lengths is

$$\|a\mathbf{v}\| = |a|\,\|\mathbf{v}\|.$$

(Note that you must write $|a|$ when a is a scalar, even if you choose to use the notation $\|\mathbf{v}\|$ when $\mathbf{v}$ is a vector.) You can check this from the general formula by factoring inside the square root; remember the identity $\sqrt{a^2} = |a|$ for arbitrary real numbers. (It's a common algebra mistake to think that $\sqrt{a^2} = a$; this is correct when $a \geq 0$ but not otherwise.) In particular,

$$\|-\mathbf{v}\| = \|\mathbf{v}\|.$$

Also,

$$\|\mathbf{0}\| = 0;$$

conversely, if $\|\mathbf{v}\| = 0$, then it must be that $\mathbf{v} = \mathbf{0}$. (Ultimately this is because a sum of squares of real numbers can only be zero if all of the original numbers are zero.)

There is no general formula for $\|\mathbf{u} + \mathbf{v}\|$; however, we can say

$$\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|.$$

This is called the **triangle inequality**, since if you draw a triangle whose sides are $\mathbf{u}$, $\mathbf{v}$, and their sum $\mathbf{u} + \mathbf{v}$, then this expresses the fact that the length of the last side is the shortest distance between its two endpoints. You can check this from the formula by squaring both sides, cancelling some common terms,

squaring again, subtracting the two sides, and factoring the result as a perfect square. You can then argue that this perfect square is greater than or equal to zero, so the right-hand side just before the subtraction is greater than or equal to the left-hand side at that stage, and this remains so upon taking principal square roots, adding some common terms, and taking principal square roots again. I'll skip the details.

If $\mathbf{v} \neq \mathbf{0}$ (so that you can divide by $\|\mathbf{v}\|$), then $\mathbf{v}/\|\mathbf{v}\|$ is a vector whose own magnitude is 1. (This is because

$$\left\| \frac{\mathbf{v}}{\|\mathbf{v}\|} \right\| = \frac{\|\mathbf{v}\|}{|\|\mathbf{v}\||} = \frac{\|\mathbf{v}\|}{\|\mathbf{v}\|} = 1,$$

using that $\|\mathbf{v}\| \geq 0$.) This is called the **unit vector** in the direction of $\mathbf{v}$, or simply the **direction** of $\mathbf{v}$. The usual notation for this is $\hat{\mathbf{v}}$:

$$\hat{\mathbf{v}} = \frac{\mathbf{v}}{\|\mathbf{v}\|}.$$

For some reason, the book never introduces this notation (or any other notation for this concept), but it refers to the idea itself quite often. Notice that you can write $\mathbf{v} = \|\mathbf{v}\|\hat{\mathbf{v}}$; this expresses the common slogan that a vector has both a length and a direction. (However, the zero vector has only a length, of 0, and no way to pick out any unit vector as its direction.)

If you perform some algebraic tricks with the triangle inequality and assume that neither $\mathbf{u}$ nor $\mathbf{v}$ is the zero vector $\mathbf{0}$ (so that you can divide by their norms), then you can also derive the compound inequality

$$-1 \leq \frac{\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - \|\mathbf{u} - \mathbf{v}\|^2}{2\,\|\mathbf{u}\|\,\|\mathbf{v}\|} \leq 1.$$

(I'll skip this derivation, but it's based on first replacing $\mathbf{v}$ with $-\mathbf{v}$, squaring both sides, and rearranging terms to derive one half of this result, then going back to the beginning and replacing $\mathbf{u}$ with $\mathbf{u} - \mathbf{v}$, squaring both sides again, and rearranging terms to derive the other half of the result.) If you draw a triangle whose sides are $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{u} - \mathbf{v}$ (so that $\mathbf{u}$ and $\mathbf{v}$ are both starting from the same point), then the Law of Cosines says that the expression in the middle of the compound inequality above is the cosine of the angle between the sides $\mathbf{u}$ and $\mathbf{v}$, and the inequality verifies that this lies within the possible range of values for a cosine. (If either $\mathbf{u}$ or $\mathbf{v}$ is the zero vector, then you don't really have a triangle, and this angle doesn't make sense.)

If you have two rays emanating from the same point in a multidimensional space, then the only way to describe the angle between them is with an angle between 0 and π (or 180°), which is the range of possible values of an arccosine (or inverse cosine), so taking the arccosine of the expression above gives you this angle:

$$\angle(\mathbf{u}, \mathbf{v}) = \operatorname{acos}\left(\frac{\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - \|\mathbf{u} - \mathbf{v}\|^2}{2\,\|\mathbf{u}\|\,\|\mathbf{v}\|} \right).$$

(In $\mathbf{R}^2$, and only in $\mathbf{R}^2$, it's possible to distinguish clockwise and counterclockwise angles, which I'll come back to when I discuss the cross product below.) Thus, it's possible to describe both lengths and angles using vectors, through the concept of the magnitude of a vector. (There's a more efficient way to calculate this cosine, which we'll see later on using the dot product, but it's important that angles can be calculated from lengths alone.)

Two vectors $\mathbf{u}$ and $\mathbf{v}$ are **perpendicular** or **orthogonal** if the angle between them is a right angle ($\pi/2$, or 90°), whose cosine is 0; the symbol for this is $\mathbf{u} \perp \mathbf{v}$. Similarly, $\mathbf{u}$ and $\mathbf{v}$ are **parallel** if the angle between them is the zero angle (0, or 0°), whose cosine is 1; the symbol for this is $\mathbf{u} \parallel \mathbf{v}$. However, people sometimes use this symbol (or even the word 'parallel') to include the case where $\mathbf{u}$ and $\mathbf{v}$ are **antiparallel**, meaning that the angle between them is a straight angle (π, or 180°), whose cosine is -1.

However, for many applications of vectors, the concept of length or magnitude really doesn't make sense! This is because vectors describe motion within any space with any coordinates, and those coordinates might refer to incompatible quantities. For example, if x measures time and y measures something that changes with time but is not itself a time (the height of a falling object, the price of a stock, the population of the world, or nearly any other quantity of interest), then it really doesn't make sense to talk about the magnitude

$$\|\Delta P\| = \|\langle \Delta x, \Delta y \rangle\| = \sqrt{\Delta x^2 + \Delta y^2}.$$

You can see this if you imagine what units of measurement you might use for such a magnitude; if x is measured in seconds and y is measured in metres (as one might do when talking about the height of a falling object, for example), then which unit is $\|\Delta P\|$ in? Neither one makes sense, nor does any combination of them.

So while lengths of vectors and angles between them always exist in the realm of mathematical abstraction, they can only be meaningful when all of the coordinates measure the same type of quantity. (Even then, these concepts may or may not really be meaningful, but at least they have a chance.) The exception to this is that we can say whether two nonzero vectors are parallel (or antiparallel) without reference to angles: $\mathbf{u}$ and $\mathbf{v}$ are parallel if there is a scalar $k > 0$ such that $\mathbf{u} = k\mathbf{v}$; they're antiparallel if there is a scalar $k < 0$ such that $\mathbf{u} = k\mathbf{v}$.

Projections

If you have two vectors $\mathbf{u}$ and $\mathbf{v}$, and assuming that neither of them is $\mathbf{0}$, place them so that they both start at the same point A and then draw a line from $A + \mathbf{v}$ to the line through A and $A + \mathbf{u}$ so that these lines intersect at a right angle. Let B be the point where these lines intersect; the vector $B - A$ is the **projection** of $\mathbf{v}$ onto $\mathbf{u}$, denoted $\operatorname{proj}_{\mathbf{u}} \mathbf{v}$. Sometimes people also consider the projection of $\mathbf{v}$ *perpendicular* to $\mathbf{u}$; this is the vector from $A + \mathbf{u}$ to B:

$$\operatorname{proj}_{\mathbf{u}}^{\perp} \mathbf{v} = \mathbf{v} - \operatorname{proj}_{\mathbf{u}} \mathbf{v}.$$

(In general, the symbol '$\perp$' is used when talking about perpendicular things, which the shape of the symbol is supposed to remind you of.)

A related concept is the **component** of $\mathbf{v}$ in the direction of $\mathbf{u}$, denoted $\operatorname{comp}_{\mathbf{u}} \mathbf{v}$; this is a scalar chosen so that

$$\operatorname{proj}_{\mathbf{u}} \mathbf{v} = \operatorname{comp}_{\mathbf{u}} \mathbf{v}\, \hat{\mathbf{u}}.$$

It's a common mistake to think that $\operatorname{proj}_{\mathbf{u}} \mathbf{v}$ has the same direction as $\mathbf{u}$, so that consequently $\operatorname{comp}_{\mathbf{u}} \mathbf{v} = \|\operatorname{proj}_{\mathbf{u}} \mathbf{v}\|$. But in fact, $\operatorname{proj}_{\mathbf{u}} \mathbf{v}$ can just as easily have the opposite direction, so the general rule is

$$|\operatorname{comp}_{\mathbf{u}} \mathbf{v}| = \|\operatorname{proj}_{\mathbf{u}} \mathbf{v}\|.$$

The component of $\mathbf{v}$ in the direction of $\mathbf{u}$ is positive if $\mathbf{u}$ and $\mathbf{v}$ have roughly the same direction but negative if they have roughly opposite directions. (It's also possible that this component is zero, when $\mathbf{u}$ and $\mathbf{v}$ are perpendicular.)

I have not allowed $\mathbf{v}$ to be the zero vector, because then $A + \mathbf{v}$ is simply A, right on the line through A and $A + \mathbf{u}$, so it makes no sense to draw anything from that point perpendicular to that line. However, since we're already on the line, we can simply take B to be A as well, so that $\operatorname{proj}_{\mathbf{u}} \mathbf{v}$, which is $B - A$, is also $\mathbf{0}$. Thus, we have these results:

$$\operatorname{proj}_{\mathbf{u}} \mathbf{0} = \mathbf{0},\ \operatorname{comp}_{\mathbf{u}} \mathbf{0} = 0.$$

Now $\operatorname{proj}_{\mathbf{u}} \mathbf{v}$ and $\operatorname{comp}_{\mathbf{u}} \mathbf{v}$ exist no matter what $\mathbf{v}$ is (although it's still necessary that $\mathbf{u} \neq \mathbf{0}$). Once we have that, you can verify these facts by drawing the relevant pictures:

$$\operatorname{proj}_{\mathbf{u}} (\mathbf{v} + \mathbf{w}) = \operatorname{proj}_{\mathbf{u}} \mathbf{v} + \operatorname{proj}_{\mathbf{u}} \mathbf{w},\ \text{so}\ \ \operatorname{comp}_{\mathbf{u}} (\mathbf{v} + \mathbf{w}) = \operatorname{comp}_{\mathbf{u}} \mathbf{v} + \operatorname{comp}_{\mathbf{u}} \mathbf{w};$$
$$\operatorname{proj}_{\mathbf{u}} (a\mathbf{v}) = a \operatorname{proj}_{\mathbf{u}} \mathbf{v},\ \text{so}\ \ \operatorname{comp}_{\mathbf{u}} (a\mathbf{v}) = a \operatorname{comp}_{\mathbf{u}} \mathbf{v}.$$

This is all well and good, but if you know a little trigonometry, then you can get a nice formula for this component. This is because $\mathbf{v}$ forms the hypotenuse of a right triangle, one of whose legs is $\operatorname{proj}_{\mathbf{u}} \mathbf{v}$, and whose angle next to that leg is $\angle(\mathbf{u}, \mathbf{v})$ if $\mathbf{u}$ and $\mathbf{v}$ have roughly the same direction or $\pi - \angle(\mathbf{u}, \mathbf{v})$ if they have roughly opposite directions. In the first case,

$$\cos \angle(\mathbf{u}, \mathbf{v}) = \frac{\|\operatorname{proj}_{\mathbf{u}} \mathbf{v}\|}{\|\mathbf{v}\|} = \frac{\operatorname{comp}_{\mathbf{u}} \mathbf{v}}{\|\mathbf{v}\|};$$

in the other case,

$$\cos\angle(\mathbf{u},\mathbf{v}) = -\cos\big(\pi - \angle(\mathbf{u},\mathbf{v})\big) = -\frac{\|\mathrm{proj}_{\mathbf{u}}\,\mathbf{v}\|}{\|\mathbf{v}\|} = -\frac{-\,\mathrm{comp}_{\mathbf{u}}\,\mathbf{v}}{\|\mathbf{v}\|} = \frac{\mathrm{comp}_{\mathbf{u}}\,\mathbf{v}}{\|\mathbf{v}\|}.$$

In the middle, when $\mathbf{u}$ and $\mathbf{v}$ are perpendicular, then $\cos\angle(\mathbf{u},\mathbf{v})$ and $\mathrm{comp}_{\mathbf{u}}\,\mathbf{v}$ are both 0. So in any case,

$$\mathrm{comp}_{\mathbf{u}}\,\mathbf{v} = \|\mathbf{v}\|\cos\angle(\mathbf{u},\mathbf{v})$$

as long as $\mathbf{v} \neq \mathbf{0}$. (If $\mathbf{v} = \mathbf{0}$, then the angle $\angle(\mathbf{u},\mathbf{v})$ doesn't make sense, but the equation is still true in a way, since it becomes the true statement $0 = 0$ no matter what value you use for the angle.) We saw earlier how to express this cosine using only $\|\mathbf{u}\|$, $\|\mathbf{v}\|$, and $\|\mathbf{u}-\mathbf{v}\|$, but for now, let's just leave it as $\cos\angle(\mathbf{u},\mathbf{v})$.

The dot product

This now suggests that we'll get a very nice operation if we define

$$\mathbf{u}\cdot\mathbf{v} = \|\mathbf{u}\|\,\mathrm{comp}_{\mathbf{u}}\,\mathbf{v} = \|\mathbf{u}\|\,\|\mathbf{v}\|\cos\angle(\mathbf{u},\mathbf{v}).$$

This has many nice properties; for example, these follow from the corresponding properties for components:

$$\mathbf{u}\cdot(\mathbf{v}+\mathbf{w}) = (\mathbf{u}\cdot\mathbf{v}) + (\mathbf{u}\cdot\mathbf{w}),$$
$$\mathbf{u}\cdot(a\mathbf{v}) = a(\mathbf{u}\cdot\mathbf{v}).$$

However, since $\mathbf{u}$ and $\mathbf{v}$ appear symmetrically in the formula with the cosine, we have

$$\mathbf{u}\cdot\mathbf{v} = \mathbf{v}\cdot\mathbf{u},$$

and then these properties also follow:

$$(\mathbf{u}+\mathbf{v})\cdot\mathbf{w} = (\mathbf{u}\cdot\mathbf{w}) + (\mathbf{v}\cdot\mathbf{w}),$$
$$(a\mathbf{u})\cdot\mathbf{v} = a(\mathbf{u}\cdot\mathbf{v}).$$

The definition $\|\mathbf{u}\|\,\mathrm{comp}_{\mathbf{u}}\,\mathbf{v}$ allows $\mathbf{v}$ to be $\mathbf{0}$, but not $\mathbf{u}$. However, since the operation is symmetric when the vectors are nonzero, we can define it so that it continues to be symmetric, so that $\mathbf{0}\cdot\mathbf{v} = 0$ as well as $\mathbf{v}\cdot\mathbf{0} = 0$. In particular, we define $\mathbf{0}\cdot\mathbf{0}$ to be 0. (Thus, it remains true in a way that $\mathbf{u}\cdot\mathbf{v} = \|\mathbf{u}\|\,\|\mathbf{v}\|\cos\angle(\mathbf{u},\mathbf{v})$, even when $\angle(\mathbf{u},\mathbf{v})$ doesn't make sense, because in that case the equation becomes $0 = 0$ no matter what value you use for the angle.) Then the properties listed above continue to be true.

By this point, you should see where the notation comes from; this operation has a lot of the same properties as multiplication. It's variously called **inner multiplication** (for the operation) or the **inner product** (for the result of the operation), the **scalar product** (because the result is a scalar), or (naming it after its notation) the **dot product**. (Don't confuse *scalar multiplication*, describing the operation for $a\mathbf{v}$, with the *scalar product*, describing the result of the operation $\mathbf{u}\cdot\mathbf{v}$.) The properties above state that the dot product distrubutes over addition, that it's commutative, associative with scalar multiplication, etc.

Since angles can be expressed in terms of lengths, so can the dot product; you get

$$\mathbf{u}\cdot\mathbf{v} = \frac{\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - \|\mathbf{u}-\mathbf{v}\|^2}{2},$$

an expression that works regardless of whether $\mathbf{u}$ and $\mathbf{v}$ are nonzero. An important special case is when $\mathbf{u}$ and $\mathbf{v}$ are the same vector; then this simplifies to

$$\mathbf{v}\cdot\mathbf{v} = \|\mathbf{v}\|^2.$$

(Another way to see this is that the angle between a vector and itself is 0, the cosine of which is 1, so $\mathbf{v} \cdot \mathbf{v} = \|\mathbf{v}\| \, \|\mathbf{v}\| \cos 0 = \|\mathbf{v}\|^2$.)

However, as a practical matter, there is a better way to calculate this. Because the dot product distributes over addition and associates with scalar multiplication, we only need to know $\mathbf{i} \cdot \mathbf{i}$, $\mathbf{i} \cdot \mathbf{j}$, and so on; that is, we only need to know what it does to the standard basis vectors. Since these vectors are all perpendicular to one another, so the cosine between any two different ones is 0, these dot products are almost all 0. The exception is the dot product of one of these with itself; since these vectors all have a magnitude of 1, the dot product of any one with itself is $1^2 = 1$. So in 2 dimensions,

$$\langle a, b \rangle \cdot \langle c, d \rangle = (a\mathbf{i} + b\mathbf{j}) \cdot (c\mathbf{i} + d\mathbf{j}) = ac\,\mathbf{i} \cdot \mathbf{i} + ad\,\mathbf{i} \cdot \mathbf{j} + bc\,\mathbf{j} \cdot \mathbf{i} + bd\,\mathbf{j} \cdot \mathbf{j} = ac\,1 + ad\,0 + bc\,0 + bd\,1 = ac + bd;$$

in 3 dimensions,

$$\langle a, b, c \rangle \cdot \langle d, e, f \rangle = ad + be + cf$$

by a similar calculation, and most generally in n dimensions,

$$\langle a_1, a_2, \ldots, a_n \rangle \cdot \langle b_1, b_2, \ldots, b_n \rangle = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n.$$

That is, you multiply corresponding components of the vectors and add these all up. For example,

$$\langle 1, -2 \rangle \cdot \langle 3, 5 \rangle = (1)(3) + (-2)(5) = 3 - 10 = -7.$$

Now its best to give formulas for angles, projections, and components in terms of the dot product, rather than the other way around. So:

$$\begin{aligned}
\operatorname{comp}_{\mathbf{u}} \mathbf{v} &= \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|}; \\
\operatorname{proj}_{\mathbf{u}} \mathbf{v} &= \operatorname{comp}_{\mathbf{u}} \mathbf{v}\, \hat{\mathbf{u}} = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|^2} \mathbf{u} = \frac{\mathbf{u} \cdot \mathbf{v}}{\mathbf{u} \cdot \mathbf{u}} \mathbf{u}; \\
\angle(\mathbf{u}, \mathbf{v}) &= \operatorname{acos} \frac{\operatorname{comp}_{\mathbf{u}} \mathbf{v}}{\|\mathbf{v}\|} = \operatorname{acos} \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \, \|\mathbf{v}\|}.
\end{aligned}$$

Even lengths can be expressed using the dot product:

$$\|\mathbf{v}\| = \sqrt{\mathbf{v} \cdot \mathbf{v}}.$$

Row vectors

I developed the dot product geometrically, and we've seen that it's closely related to lengths and angles. I remarked before that lengths and angles don't always make sense, and the same goes for the the dot product (as well as projections and components onto a given vector). For example, if x is measured in seconds (s) and y is measured in metres (m), then $\langle 1\,\mathrm{s}, -2\,\mathrm{m} \rangle \cdot \langle 3\,\mathrm{s}, 5\,\mathrm{m} \rangle = 3\,\mathrm{s}^2 - 10\,\mathrm{m}^2$ doesn't really make sense.

On the other hand, sometimes dot products can make sense in a context like this. For example, suppose that x represents the time at which something occurs and y represents its location, so that the vector $\Delta P = \langle \Delta x, \Delta y \rangle$ represents a passage of time together with a change in location, like the vectors above might do; then if the object in question is a missile that's going to explode at some unkown time and distance and you think that it's going to move slowly while I think that it's going to move quickly, then we might make a bet where I pay you \$1 for every second that it lasts until it explodes but you pay me \$2 for every metre that it travels. If it travels 5 metres in 3 seconds before exploding, then you'll get $(1)(3) - (2)(5) = -7$ dollars, or put another way, you'll owe me \$7. This can be represented as the dot product

$$\langle \$1/\mathrm{s}, -\$2/\mathrm{m} \rangle \cdot \langle 3\,\mathrm{s}, 5\,\mathrm{m} \rangle = (\$1/\mathrm{s})(3\,\mathrm{s}) + (-\$2/\mathrm{m})(5\,\mathrm{m}) = \$3 - \$10 = -\$7,$$

where the first vector is determined by the nature of our bet (you get \$1 per second and pay \$2 per metre), while the second vector is determined by the behaviour of the missile (it lasts 3 seconds and travels 5 metres).

Now, while the vector $\langle 3\,\mathrm{s}, 5\,\mathrm{m}\rangle$ really does describe a change in x and a change in y, where x and y represent time and position as I stated above, the vector $\langle \$1/\mathrm{s}, -\$2/\mathrm{m}\rangle$ does not. In the context of measuring time and position, this vector is a different kind of vector, one for which a dot product with an ordinary vector makes sense, even though lengths and angles don't make sense for any of these vectors. A vector like this is variously called a **dual vector**, a **covector**, or a **row vector**; in the last case, an ordinary vector may be called a **column vector**. I'll use the terminology of row and column vectors, which ultimately comes from matrix theory.

Row vectors obey the same rules of arithmetic as column vectors; here is a list of operations with these that make sense:

- Addition: adding a column vector to a point to get another point, adding two column vectors together to get another column vector, adding two row vectors together to get another row vector;
- Subtraction: subtracting a column vector from a point to get another point, subtracting one column vector from another to get another column vector, subtracting one row vector from another to get another row vector;
- Multiplication: multiplying a column vector by a scalar to get another column vector, multiplying a row vector by a scalar to get another row vector, multiplying a row vector and a column vector to get a scalar.

In particular, there is (in general) no notion of 'row point' that can interact with row vectors in the way that points interact with column vectors.

Area

Now let's go back to a geometric conception of vectors. If you take two vectors $\mathbf{u}$ and $\mathbf{v}$ and place them to start at a point A, then you can connect their endpoints to make a triangle and then ask what the area of that triangle is. It's actually a bit nicer to think of that triangle as half of a parallelogram: two opposite sides of the parallelogram are $\mathbf{u}$, one running from A to $A+\mathbf{u}$, the other running from $A+\mathbf{v}$ to $A+\mathbf{v}+\mathbf{u}$; the other two opposite sides are $\mathbf{v}$, one running from A to $A+\mathbf{v}$, the other running from $A+\mathbf{u}$ to $A+\mathbf{u}+\mathbf{v}$ (which of course is the same as $A+\mathbf{v}+\mathbf{u}$).

This question can be asked in any number of dimensions, and the answer may be written $\|\mathbf{u}\times\mathbf{v}\|$. This notation suggests that this area will be the magnitude of something more fundamental, which is $\mathbf{u}\times\mathbf{v}$ itself, and this is true to an extent, but exactly how that works depends on how many dimensions we're in. So for now, I'm just going to stick with $\|\mathbf{u}\times\mathbf{v}\|$. However, I can give you the terminology: whatever $\mathbf{u}\times\mathbf{v}$ is, the operation may be called **outer multiplication**, and the result may be called the **outer product** or the **cross product**, and in 3 dimensions (where it is best known), it's also called the **vector product**.

With the help of trigonometry,

$$\|\mathbf{u}\times\mathbf{v}\| = \|\mathbf{u}\|\,\|\mathbf{v}\|\sin\angle(\mathbf{u},\mathbf{v}).$$

Notice that this sine is always positive, since the angle lies between 0 and π. For such an angle θ, $\sin\theta = \sqrt{1-\cos^2\theta}$; with the help of the dot product, this means that

$$\|\mathbf{u}\times\mathbf{v}\| = \sqrt{\|\mathbf{u}\|^2\,\|\mathbf{v}\|^2 - (\mathbf{u}\cdot\mathbf{v})^2}.$$

(This formula makes sense even if $\mathbf{u}$ or $\mathbf{v}$ is the zero vector, in which case the result is zero.) If you write out $\mathbf{u}\cdot\mathbf{v}$ in this expression in terms of the lengths $\|\mathbf{u}\|$, $\|\mathbf{v}\|$, and $\|\mathbf{u}-\mathbf{v}\|$, then the formula factors as

$$\|\mathbf{u}\times\mathbf{v}\| = \frac{\sqrt{-(\|\mathbf{u}\|+\|\mathbf{v}\|+\|\mathbf{u}-\mathbf{v}\|)(\|\mathbf{u}\|+\|\mathbf{v}\|-\|\mathbf{u}-\mathbf{v}\|)(\|\mathbf{u}\|-\|\mathbf{v}\|+\|\mathbf{u}-\mathbf{v}\|)(\|\mathbf{u}\|-\|\mathbf{v}\|-\|\mathbf{u}-\mathbf{v}\|)}}{2}.$$

(Despite the initial minus sign, the expression inside the square root is positive, since the last factor is negative.) This result was known to the ancient Greek–Egyptian mathematician and inventor Hero (or Heron) of Alexandria. (He invented the steam engine, the windmill, and the vending machine, although none of those caught on at the time.)

If $\mathbf{u}$ and $\mathbf{v}$ are parallel (or antiparallel), or if either (or both) of them is the zero vector $\mathbf{0}$, then $|\mathbf{u} \cdot \mathbf{v}| = \|\mathbf{u}\| \, \|\mathbf{v}\|$, so $\|\mathbf{u} \times \mathbf{v}\| = 0$. From another perspective, if $\mathbf{u}$ and $\mathbf{v}$ are parallel, then the angle between them is 0, whose sine is 0; if they're antiparallel, then the sine is still $\sin \pi = 0$. In this case, you don't really have a parallelogram, but a simple line segment (or a point if $\mathbf{u}$ and $\mathbf{v}$ are both $\mathbf{0}$), whose area is indeed zero.

Here are some important algebraic properties of $\|\mathbf{u} \times \mathbf{v}\|$:

$$\|\mathbf{u} \times \mathbf{v}\| = \|\mathbf{v} \times \mathbf{u}\|;$$
$$\|\mathbf{u} \times a\mathbf{v}\| = |a| \, \|\mathbf{u} \times \mathbf{v}\|;$$
$$\|\mathbf{u} \times \mathbf{v}\| = \left\|\mathbf{u} \times \operatorname{proj}_{\mathbf{u}}^{\perp} \mathbf{v}\right\| = \|\mathbf{u}\| \left\|\operatorname{proj}_{\mathbf{u}}^{\perp} \mathbf{v}\right\|.$$

(The last of these assumes that $\mathbf{u} \neq \mathbf{0}$, so that projection perpendicular to $\mathbf{u}$ makes sense.) These should be obvious geometrically; in particular, the last of these states that the area of a parallelogram is the same as the area of a rectangle with the same base and height.

The cross product in three dimensions

For vectors in $\mathbf{R}^3$, we can interpret $\mathbf{u} \times \mathbf{v}$ as a vector. The magnitude $\|\mathbf{u} \times \mathbf{v}\|$ is the area from the previous section, so we only need to describe the direction of $\mathbf{u} \times \mathbf{v}$: it will be perpendicular to both $\mathbf{u}$ and $\mathbf{v}$.

Most of the time, there are precisely two directions perpendicular to two vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbf{R}^3$. To decide which of these is the direction of $\mathbf{u} \times \mathbf{v}$, we use the *right-hand rule*: if you start by pointing the fingers of your right hand in the direction of $\mathbf{u}$, curl them to point in the direction of $\mathbf{v}$, and then stick out your thumb, then your thumb will point roughly in the direction of $\mathbf{u} \times \mathbf{v}$. (This should be used together with a right-handed coordinate system: if you point your fingers along the positive x-axis, curl them to point along the positive y-axis, and then stick out your thumb, then your thumb will point roughly along the positive z-axis.) If $\mathbf{u}$ and $\mathbf{v}$ happen to be parallel (or antiparallel), or if either (or both) of them is the zero vector $\mathbf{0}$, then this won't work; however, in that case, $\|\mathbf{u} \times \mathbf{v}\| = 0$, so then $\mathbf{u} \times \mathbf{v}$ must be $\mathbf{0}$, which has no direction.

Like the dot product, this operation distributes over addition and associates with scalar multiplication:

$$\mathbf{u} \times (\mathbf{v} + \mathbf{w}) = \mathbf{u} \times \mathbf{v} + \mathbf{u} \times \mathbf{w},$$
$$\mathbf{u} \times a\mathbf{v} = a(\mathbf{u} \times \mathbf{v}).$$

The latter fact is easy to see, since we have a corresponding fact for $\|\mathbf{u} \times a\mathbf{v}\|$ and the direction of $\mathbf{u} \times a\mathbf{v}$ reverses when a is negative. The first of these is more difficult; it uses the result for $\|\mathbf{u} \times \mathbf{v}\|$ in terms of $\operatorname{proj}_{\mathbf{u}}^{\perp} \mathbf{v}$. This allows you to draw everything in the plane perpendicular to $\mathbf{u}$; if you look in the direction of $\mathbf{u}$ when looking at this plane, then $\mathbf{u} \times \mathbf{v}$ rotates $\operatorname{proj}_{\mathbf{u}}^{\perp} \mathbf{v}$ (which is in this plane) clockwise through a right angle and scales it by $\|\mathbf{v}\|$; since both this operation and projection distribute over addition, so does the cross product itself.

However, there is one important difference between the properties of the dot and cross products:

$$\mathbf{u} \times \mathbf{v} = -\mathbf{v} \times \mathbf{u}.$$

This is because, while the magnitudes are the same, the directions are reversed, since you're curling your fingers the other way.

For practical calculations, it's again enough to know what happens to the standard basis vectors:

$$\mathbf{i} \times \mathbf{i} = \mathbf{0}, \ \mathbf{i} \times \mathbf{j} = \mathbf{k}, \ \mathbf{i} \times \mathbf{k} = -\mathbf{j},$$
$$\mathbf{j} \times \mathbf{i} = -\mathbf{k}, \ \mathbf{j} \times \mathbf{j} = \mathbf{0}, \ \mathbf{j} \times \mathbf{k} = \mathbf{i},$$
$$\mathbf{k} \times \mathbf{i} = \mathbf{j}, \ \mathbf{k} \times \mathbf{j} = -\mathbf{i}, \ \mathbf{k} \times \mathbf{k} = \mathbf{0}.$$

Based on this,

$$\begin{aligned}\langle a,b,c\rangle \times \langle d,e,f\rangle &= (a\mathbf{i}+b\mathbf{j}+c\mathbf{k}) \times (d\mathbf{i}+e\mathbf{j}+f\mathbf{k}) = (bf-ce)\mathbf{i}+(cd-af)\mathbf{j}+(ae-bd)\mathbf{k} \\ &= \langle bf-ce, cd-af, ae-bd\rangle.\end{aligned}$$

For example,

$$\langle 1,-2,0\rangle \times \langle 2,2,1\rangle = \langle (-2)(1)-(0)(2),(0)(2)-(1)(1),(1)(2)-(-2)(2)\rangle = \langle -2-0,0-1,2+4\rangle = \langle -2,-1,6\rangle.$$

If you know about determinants, then you can think of

$$\langle a,b,c\rangle \times \langle d,e,f\rangle = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a & b & c \\ d & e & f \end{vmatrix};$$

the value of this determinant is the value of the cross product above.

Along with the cross product, people often look at the so-called *triple scalar product* of three vectors in $\mathbf{R}^3$; this is simply

$$\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w}).$$

This can be calculated with determinants as well; then the top row of the determinant, instead of consisting of the standard basis vectors, now consists of the components of $\mathbf{u}$ to go with the components of $\mathbf{v}$ and $\mathbf{w}$ in the other rows. Geometrically, this represents a volume; more precisely, $|\mathbf{u} \cdot \mathbf{v} \times \mathbf{w}|$ is the volume of a parallelepiped whose edges are $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$, and $\mathbf{u} \cdot \mathbf{v} \times \mathbf{w}$ is positive if you can curl the fingers of your right hand from $\mathbf{u}$ to $\mathbf{v}$ and stick out your thumb along $\mathbf{w}$ but negative if your thumb points the wrong way.

The cross product in two dimensions

For vectors in $\mathbf{R}^2$, we can interpret $\mathbf{u} \times \mathbf{v}$ as a scalar, so this is sometimes callled the *scalar cross product*. The absolute value $|\mathbf{u} \times \mathbf{v}|$ is the $\|\mathbf{u} \times \mathbf{v}\|$ from above; $\mathbf{u} \times \mathbf{v}$ itself is positive if you turn counterclockwise to go from $\mathbf{u}$ to $\mathbf{v}$ but negative if you turn clockwise. (Here I'm assuming a counterclockwise coordinate system: the rotation from the positive x-axis to the positive y-axis is counterclockwise.) If $\mathbf{u}$ and $\mathbf{v}$ are parallel (or antiparallel), or if either of them is the zero vector $\mathbf{0}$, then $\mathbf{u} \times \mathbf{v}$ is just 0.

The cross product in 2 dimensions follows the same algebraic rules as in 3 dimensions:

$$\begin{aligned}\mathbf{u} \times (\mathbf{v}+\mathbf{w}) &= \mathbf{u} \times \mathbf{v} + \mathbf{u} \times \mathbf{w}, \\ \mathbf{u} \times a\mathbf{v} &= a(\mathbf{u} \times \mathbf{v}), \\ \mathbf{u} \times \mathbf{v} &= -\mathbf{v} \times \mathbf{u}.\end{aligned}$$

If anything, these are easier to establish geometrically than the corresponding properties in $\mathbf{R}^3$.

Another way to think of the scalar cross product is to embed $\mathbf{R}^2$ within $\mathbf{R}^3$; that is, we take the z-coordinate of every point to be fixed (typically $z=0$), so that the z-component of every vector is $\Delta z = 0$. Then instead of the scalar cross product $\mathbf{u} \times \mathbf{v}$, you can speak of the triple scalar product $\mathbf{k} \cdot \mathbf{u} \times \mathbf{v}$. Yet another way to think of it is as a dot product; much as $a-b$ is the sum of a and $-b$, so $\mathbf{u} \times \mathbf{v}$ is the dot product of $\mathbf{u}$ and $\times\mathbf{v}$, where $\times\mathbf{v}$ is simply $\mathbf{v}$ rotated clockwise through a right angle. You can also speak of signed angles in 2 dimensions; if you treat a counterclockwise angle as positive and a clockwise angle as negative, then

$$\mathbf{u} \times \mathbf{v} = \|\mathbf{u}\| \, \|\mathbf{v}\| \sin \bar{\angle}(\mathbf{u},\mathbf{v}),$$

where the bar over the angle symbol indicates a signed angle.

For practical calculations, since $\mathbf{i} \times \mathbf{i} = 0$, $\mathbf{i} \times \mathbf{j} = 1$, $\mathbf{j} \times \mathbf{i} = -1$, and $\mathbf{j} \times \mathbf{j} = 0$, the formula is

$$\langle a,b\rangle \times \langle c,d\rangle = ad - bc.$$

For example,

$$\langle 1, -2\rangle \times \langle 3, 5\rangle = (1)(5) - (-2)(3) = 5 + 6 = 11.$$

If you know about determinants, then

$$\langle a, b\rangle \times \langle c, d\rangle = \begin{vmatrix} a & b \\ c & d \end{vmatrix}.$$

Similarly,

$$\times\langle a, b\rangle = \begin{vmatrix} \mathbf{i} & \mathbf{j} \\ a & b \end{vmatrix} = \langle b, -a\rangle.$$

Cross products in more than 3 dimensions can also be done, but in that case the result is neither a scalar nor a vector but a more general concept called a *tensor*. We will not be using these.

Orientation

The dot and cross products both rely on the geometric notion of length, but the cross product additionally depends on an **orientation**; this is the choice between the right-hand and left-hand rules (in 3 dimensions) or between counterclockwise and clockwise angles (in 2 dimensions). While our physical space really does have lengths and angles, the choice of orientation is arbitrary, so results that apply to geometry shouldn't depend on it.

Just as we can distinguish row vectors from column vectors in situations where lengths don't make sense, so we can distinguish axial vectors from polar vectors in situations where orientation is arbitrary. So, a **polar vector** is an ordinary vector representing a change in position, but an **axial vector** or **pseudovector** is a vector together with a choice of orientation, where we may reverse our choice of orientation as we please so long as we replace the vector with its opposite when we do so. For example, while a polar vector in $\mathbf{R}^3$ may be fully described as $\langle -2, -1, 6\rangle$, an axial vector in $\mathbf{R}^3$ might be described as $\langle -2, -1, 6\rangle$ right-handed, or (for the *same* axial vector) as $\langle 2, 1, -6\rangle$ left-handed. Thus you can say, for example,

$$\langle 1, -2, 0\rangle \times \langle 2, 2, 1\rangle = \langle -2, -1, 6\rangle \text{ right-handed} = \langle 2, 1, -6\rangle \text{ left-handed}.$$

Similarly, a **pseudoscalar** is a scalar together with a choice of orientation, where again we may reverse our choice of orientation as we please so long as we replace the scalar with its opposite. In $\mathbf{R}^2$, the cross product of two vectors is a pseudoscalar; in $\mathbf{R}^3$, the triple scalar product of three vectors is a pseudoscalar. For example,

$$\langle 1, -2\rangle \times \langle 3, 5\rangle = 11 \text{ counterclockwise} = -11 \text{ clockwise},$$

and

$$\langle 1, -2, 0\rangle \cdot \langle 2, 2, 1\rangle \times \langle 0, 3, 5\rangle = 27 \text{ right-handed} = -27 \text{ left-handed}.$$

Axial vectors obey the same rules of arithmetic as polar vectors; here is a list of operations with these that make sense in $\mathbf{R}^3$:

- Addition: adding a polar vector to a point to get another point, adding two polar vectors together to get another polar vector, adding two axial vectors together to get another axial vector;
- Subtraction: subtracting a polar vector from a point to get another point, subtracting one polar vector from another to get another polar vector, subtracting one axial vector from another to get another axial vector;
- Scalar multiplication: multiplying a polar vector by a scalar to get another polar vector, multiplying an axial vector by a scalar to get another axial vector, multiplying a polar vector by a pseudoscalar to get an axial vector, multiplying an axial vector by a pseudoscalar to get a polar vector;
- Inner multiplication (dot product): multiplying two polar vectors to get a scalar, multiplying a polar vector and an axial vector to get a pseudoscalar, multiplying two axial vectors to get a scalar;
- Outer multiplication (cross product): multiplying two polar vectors to get an axial vector, multiplying a polar vector and an axial vector to get a polar vector, multiplying two axial vectors to get an axial vector.

Similarly, pseudoscalars can be added or subtracted to produce more pseudoscalars and can be multiplied together to produce an ordinary scalar, or you can multiply a scalar and a pseudoscalar to produce another pseudoscalar. In $\mathbf{R}^2$, the list of operations is the same, except that the result of a cross product is a scalar or a pseudoscalar rather than a vector (a polar vector) or a pseudovector (an axial vector).

The rule of thumb for all of this is that you can only add or subtract things that are alike in every way, but you can multiply anything together; the result is 'pseudo' if you multiplied together an odd number of pseudothings (so pseudos cancel, like minus signs, in pairs), where the cross product introduces an extra pseudo.

In the most general case, where you don't have a good notion of length and also don't have any way to prefer one orientation over another, you have polar column vectors (the ordinary notion of vector), axial column vectors, polar row vectors, and axial row vectors. In general, only polar column vectors can interact with points. None of this affects calculations when properly done, but like keeping track of units, keeping track of these can prevent you from accidentally doing meaningless calculations.

Linear geometry

Here I'll summarize the formulas in Section 11.5 of the textbook that can be made simpler by doing arithmetic with points and vectors (instead of just with vectors as the book does) or by using the two-dimensional cross product (instead of only the three-dimensional cross product as the book does).

A parametric equation for the line through a point P_0 in the direction of a nonzero vector $\mathbf{v}$ is

$$P = P_0 + t\mathbf{v},$$

where t is the parameter and $P = (x, y)$ or $P = (x, y, z)$ is a point on the line. Similarly, a parametric equation for the line through points P_1 and P_2 is

$$P = P_1 + t(P_2 - P_1).$$

A nonparametric equation for the line through P_0 in the direction of $\mathbf{v}$ in 2 dimensions is

$$(P - P_0) \times \mathbf{v} = 0.$$

Similarly, a system of equations for the line through P_0 in the direction of $\mathbf{v}$ in 3 dimensions is

$$(P - P_0) \times \mathbf{v} = \mathbf{0}.$$

(The only difference is whether the zero on the right-hand side is the scalar 0 or the vector $\mathbf{0}$.)

The distance from a point S to the line through P_0 in the direction of $\mathbf{v}$ is

$$\|(S - P_0) \times \hat{\mathbf{v}}\| = \frac{\|(S - P_0) \times \mathbf{v}\|}{\|\mathbf{v}\|}.$$

Similarly, the distance from S to the line through P_1 and P_2 is

$$\frac{\|(S - P_1) \times (P_2 - P_1)\|}{\|P_2 - P_1\|}.$$

An equation for the line (in 2 dimensions) or plane (in 3 dimensions) through P_0 and perpendicular to a vector $\mathbf{n}$ is

$$(P - P_0) \cdot \mathbf{n} = 0.$$

Finally, the distance from S to the line or plane through P_0 and perpendicular to $\mathbf{n}$ is

$$\|(S - P_0) \cdot \hat{\mathbf{n}}\| = \frac{\|(S - P_0) \cdot \mathbf{n}\|}{\|\mathbf{n}\|}.$$

Point- and vector-valued functions

Besides individual points and vectors, one can also consider variable points and vectors, which are the outputs of point- and vector-valued functions. A **point-valued function** in $\mathbf{R}^n$ consists of n ordinary functions, all with the same domain. For example, a point-valued function in $\mathbf{R}^2$ consists of 2 functions with the same domain, say $f(t) = t^2$ and $g(t) = t^3$. We put these together into a single function (f, g), which takes a real-number t as input and produces the point

$$(f, g)(t) = \Big(f(t), g(t)\Big) = (t^2, t^3) = \mathrm{O} + t^2\mathbf{i} + t^3\mathbf{j}$$

(in this case) as output. A **vector-valued function** in $\mathbf{R}^n$ also consists of n ordinary functions, all with the same domain. But now we think of the output as a vector:

$$\langle f, g\rangle(t) = \langle f(t), g(t)\rangle = \langle t^2, t^3\rangle = t^2\mathbf{i} + t^3\mathbf{j}$$

(for example). If we want to know whether one of these functions is continuous or differentiable, then we just look at each of its components separately. For example, since the functions f and g above are continuous and differentiable everywhere, so are (f, g) and $\langle f, g\rangle$.

The textbook often doesn't distinguish between a point P and its position vector $\mathbf{r} = P - \mathrm{O}$, where O is the origin of a coordinate system. Conceptually, they're very different, since you can talk about points and vectors geometrically without bringing coordinates into it, so the concepts are meaningful even if there is no inherent point O to equivocate them. On the other hand, when doing calculations, it's easy to conflate them; since the coordinates of O are all 0, when you do the subtraction, the components of $\mathbf{r}$ are exactly the same as the coordinates of P. Still, you should always keep in mind whether a given expression really refers to a point or to a vector. In particular, a point-valued function can be viewed as a **parametrized curve**; each value of the input t (which in this context is called a *parameter*) gives a point, and all of these points together make up a curve. A vector-valued function only defines a curve by interpreting each vector with reference to point O deemed to be the origin, but that is how the textbook insists on doing it starting in Chapter 12.

If P is a point, then the difference ΔP is a vector (because it's the result of subtracting two points), and then the differential $\mathrm{d}P$ is an infinitesimal vector. If P is a function of some scalar quantity t, then $\mathrm{d}P/\mathrm{d}t$ makes sense, because it's a vector divided by a scalar, but now it's no longer infinitesimal (unless it happens to be zero). In other words, *the derivative of a point with respect to a scalar is a vector.* Another way to see this is that if F is a point-valued function, then its derivative F' is a vector-valued function:

$$F'(t) = \lim_{h\to 0}\left(\frac{F(t+h) - F(t)}{h}\right);$$

first subtract two points to get a vector, then divide by the scalar h to get another vector, and finally take the limit of these vectors to get a vector. Of course, the derivative of a *vector* with respect to a scalar is *also* a vector; in other words, the derivative of a vector-valued function is also a vector-valued function.

For example, if P gives the position of some object at time t, then P is a point, but $\mathrm{d}P/\mathrm{d}t$, the *velocity* of the object, is a vector. (Note that the magnitude of this vector is the object's *speed*.) If we write $\mathbf{v}$ for $\mathrm{d}P/\mathrm{d}t$ (which can also be written as $\mathrm{d}\mathbf{r}/\mathrm{d}t$), then $\mathrm{d}\mathbf{v}/\mathrm{d}t$ is the acceleration of the object, which is also a vector. (Physicists and mechanical engineers use the word 'acceleration' like this, to indicate any change in velocity —speed or direction— over time. In everyday language, this word means something more like $\mathrm{d}\|\mathbf{v}\|/\mathrm{d}t$, the derivative of speed with respect to time, which is the same as the scalar component of the acceleration in the direction of the velocity. This is positive if the object is speeding up and negative if the object is slowing down, or decelerating. Section 12.5 of the textbook discusses all of this in detail.)

Reversing this, if you take the indefinite integral of a vector, then the result may be either a point *or* a vector, because differentiating either of these yields a vector. This ambiguity is packaged into the constant of integration. For example, $\int\langle 2t, 3\rangle\,\mathrm{d}t = \langle t^2, 3t\rangle + C$, which is a point if C is a point and a vector if C is a vector. (If C is a vector, then you may want to call it $\mathbf{C}$ instead, but that is just a convention, not a requirement.) The definite integral of a vector, however, is always a vector: fundamentally, you get

it by adding up infinitely many infinitesimal vectors (or approximate it by adding up a large number of small vectors), and adding up vectors yields a vector; in practice, you usually calculate it by subtracting indefinite integrals, and regardless of whether you view the indefinite integrals as points or as vectors, subtracting them yields a vector. For example, both $\int_{t=0}^{1}\langle 2t,3\rangle\,\mathrm{d}t = \langle t^2,3t\rangle|_{t=0}^{1} = \langle 1,3\rangle - \langle 0,0\rangle = \langle 1,3\rangle$, and $\int_{t=0}^{1}\langle 2t,3\rangle\,\mathrm{d}t = (t^2,3t)|_{t=0}^{1} = (1,3) - (0,0) = \langle 1,3\rangle$ give the same result. In fact, either of them could be packaged up as

$$\int_{t=0}^{1}\langle 2t,3\rangle\,\mathrm{d}t = \left\langle \int_{t=0}^{1}2t\,\mathrm{d}t, \int_{t=0}^{1}3\,\mathrm{d}t\right\rangle = \left\langle t^2|_{t=0}^{1}, 3t|_{t=0}^{1}\right\rangle = \langle 1-0,3-0\rangle = \langle 1,3\rangle.$$

Putting this all together, consider the initial-value problem in which the acceleration of an object is $-32\mathbf{k} = \langle 0,0,-32\rangle$ (which is the acceleration of a freely falling object near Earth's surface, if we use units of feet and seconds), the object's initial velocity is $\langle 3,0,4\rangle$ (so a speed of 5 ft/s eastward and upward with a slope of 4/3), and the object's initial position is $(0,0,100)$ (so 100 feet above the origin on the ground). Then you can calculate a general formula for the object's position P as a function of the elapsed time t by integrating:

$$\begin{aligned}
\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} &= \langle 0,0,-32\rangle;\\
\mathrm{d}\mathbf{v} &= \langle 0,0,-32\rangle\,\mathrm{d}t;\\
\int_{\mathbf{v}=\langle 3,0,4\rangle}\mathrm{d}\mathbf{v} &= \int_{t=0}\langle 0,0,-32\rangle\,\mathrm{d}t;\\
\mathbf{v} - \langle 3,0,4\rangle &= \langle 0,0,-32t\rangle - \langle 0,0,-32(0)\rangle;\\
\mathbf{v} &= \langle 3,0,4\rangle + \langle 0,0,-32t\rangle;\\
\frac{\mathrm{d}P}{\mathrm{d}t} &= \langle 3,0,4-32t\rangle;\\
\mathrm{d}P &= \langle 3,0,4-32t\rangle\,\mathrm{d}t;\\
\int_{P=(0,0,100)}\mathrm{d}P &= \int_{t=0}\langle 3,0,4-32t\rangle\,\mathrm{d}t;\\
P-(0,0,100) &= \langle 3t,0,4t-16t^2\rangle - \left\langle 3(0),0,4(0)-16(0)^2\right\rangle;\\
P &= (0,0,100) + \langle 3t,0,4t-16t^2\rangle;\\
P &= (3t,0,100+4t-16t^2).
\end{aligned}$$

In other words, the position after t seconds is $3t$ feet east of the origin at a height of $100+4t-16t^2$ feet.

(In the course of solving this, I've used the *semidefinite integral*:

$$\int_{t=a}f(t)\,\mathrm{d}t = \int_{\tau=a}^{t}f(\tau)\,\mathrm{d}\tau.$$

The Fundamental Theorem of Calculus allows us to calculate these integrals easily:

$$\int_{t=a}F'(t)\,\mathrm{d}t = F(t)-F(a).$$

This is very handy when solving initial-value problems. Since $\mathbf{v} = \langle 3,0,4\rangle$ when $t=0$, the first step in which I introduced integrals is doing the same operation to both sides of the equation; similarly, the second introduction of integrals is valid because $P=(0,0,100)$ when $t=0$. To solve this problem using indefinite integrals instead requires two extra steps —one for each integration— to find the constants associated with the indefinite integrals, but using semidefinite integrals avoids that.)

Arclength

When finding the length of a curve by integration, you are ultimately integrating an expression such as $\sqrt{\mathrm{d}x^2 + \mathrm{d}y^2}$. This particular expression applies in 2 dimensions; in words, it is the principal square root of the sum of the square of the differential of x and the square of the differential of y. An expression like this, containing differentials, is called a *differential form*; the textbook mentions differential forms briefly in Section 15.3, but they are really all over the place in this multivariable Calculus, sometimes hidden just under the surface, sometimes out in the open without being acknowledged. I'll be pointing them out whenever they appear.

This particular differential form is called the **arclength element** and is traditionally written $\mathrm{d}s$ (although that notation is misleading for reasons that I will return to later). A simpler way to think of $\mathrm{d}s$, which works in *any* number of dimensions, is as $\|\mathrm{d}P\|$, the magnitude of the differential of the position P. Remember that $\mathrm{d}P$ is a vector when P is a point, so it has a magnitude; in fact, $\mathrm{d}P$ is the same as $\mathrm{d}\mathbf{r}$, so you can also think of $\mathrm{d}s$ as $\|\mathrm{d}\mathbf{r}\|$, the magnitude of the differential of the position vector $\mathbf{r}$. In 2 dimensions, where $P = (x, y)$ and $\mathbf{r} = \langle x, y\rangle$, $\mathrm{d}\mathbf{r} = \mathrm{d}P = \langle \mathrm{d}x, \mathrm{d}y\rangle$, whose magnitude is the arclength element that I talked about above. In 3 dimensions, $DP = \langle \mathrm{d}x, \mathrm{d}y, \mathrm{d}z\rangle$, whose magnitude is $\mathrm{d}s = \sqrt{\mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2}$.

When working with a parametrized curve, every variable (x and y, and z if it exists, whether individually or combined into P or $\mathbf{r}$) is given as a function of some parameter t. By differentiating these, their differentials come to be expressed using t and $\mathrm{d}t$. The absolute value $|\mathrm{d}t|$ will naturally appear in the integrand; but if you set up the integral so that t is increasing, then $\mathrm{d}t$ is positive, so $|\mathrm{d}t| = \mathrm{d}t$. Then you can write $\|\mathrm{d}P\|$ as $\|\mathbf{v}\|\ |\mathrm{d}t| = \|\mathbf{v}\|\,\mathrm{d}t$, where $\mathbf{v} = \mathrm{d}P/\mathrm{d}t = \mathrm{d}\mathbf{r}/\mathrm{d}t$ is the velocity, as given in the textbook. More explicitly, this is

$$\mathrm{d}s = \sqrt{\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2 + \left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2}\,\mathrm{d}t$$

(in 2 dimensions), which is also given in the textbook. But while you might integrate this in practice to perform a specific calculation, you are most fundamentally integrating a differential form in which t does not appear. This is why the result ultimately does not depend on how you parametrize the curve. (Later on, I'll discuss what it means, in general, to integrate a differential form along a curve, including why and to what extent this is independent of the parametrization.)

Functions of several variables

While a parametrized curve is given by a point-valued function, that is a function that takes a scalar (a number) as input and gives a point as output, the main object of study in this class is the reverse: a function that takes a point as input and gives a number as output. Since a point is given by a list of numbers (its coordinates), a function of this sort can also be viewed as taking a list of numbers as input; for this reason, we call it a **function of several variables** (the variables in question being those that stand for the input numbers).

The hierarchy of functions and relations

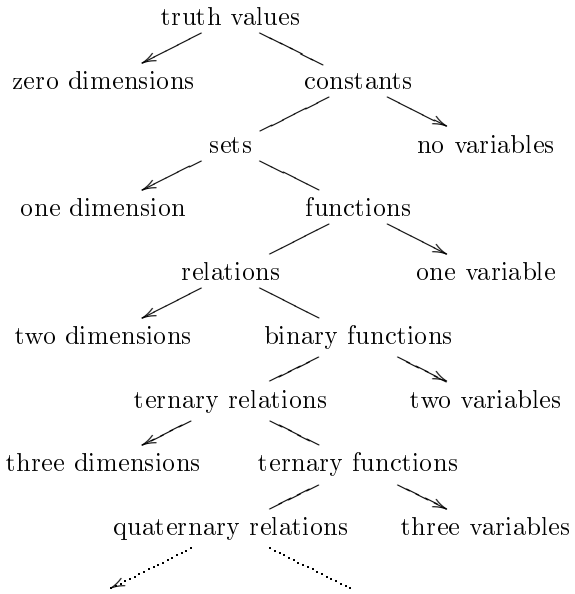
There are many different types of mathematical objects that we could study in this class. Some of them are relation-like objects:

- truth values,
- sets,
- relations,
- ternary relations,
- quaternary relations,
- etc;

some of them are function-like objects:

- constants,
- functions,
- binary functions,
- ternary functions,
- etc.

As you go along these lists, both the number of variables and the number of dimensions needed for graphing increase, as in the following diagram:



A **truth value** is either true or false; any statement with no variables in it, such as the statement that $0 < 2$, should evaluate to true or false (in this case, true). To indicate that you are talking about the truth value of this statement, rather than asserting the statement itself, you can put curly braces around it (although there are several other notations used for this); for example, $\{0 < 2\}$ is the truth value that 0 is less than 2, which is the true truth value rather than the false one. You can also use a variable to give

a name to a truth value, so maybe p stands for $\{0 < 2\}$; we won't need to do this in this course, but you'll do this constantly if you take a course on Logic.

A **constant** is, in this class, a *real number*, such as -2. Any expression with no variables should evaluate to a constant, but we use one dimension to graph a constant on a number line. Again, you can use a variable to stand for a constant, so maybe a stands for -2; in other words, $a = -2$.

A **set** is, in the simplest case, a *set of real numbers*. A statement with one variable defines a set, such as $\{x \mid x < 2\}$, the set of real numbers that are less than 2. We again use one dimension to graph a set. If A stands for the set $\{x \mid x < 2\}$, then these two statements mean the same thing:

- $x \in A$, usually pronounced 'x in A';
- $x < 2$.

The first of these says that x *belongs* to the set A, while the second uses the definition of A to say precisely what that means about x.

A **function**, or *unary function* for emphasis, is a rule for taking a number (the *input*) and using it to calculate a number (the *output*). An example is $(x \mapsto x - 2)$, the rule which subtracts 2 from any number. To graph a function, we need two dimensions, one for the input and one for the output. If f stands for the function $(x \mapsto x - 2)$, then these two expressions mean the same thing:

- $f(x)$, usually pronounced 'f of x';
- $x - 2$.

The first of these is the *value* of the function f at the *argument* x, while the second uses the definition of f to say precisely what that means in terms of x.

A **relation**, or *binary relation* for emphasis, is a set of ordered pairs instead of a set of individual numbers. An example is $\{x, y \mid x + y < 2\}$. We again use two dimensions to graph a relation. If R stands for the relation $\{x, y \mid x + y < 2\}$, then these two statements mean the same thing:

- $(x, y) \in R$;
- $x + y < 2$.

The first of these says that x and y are *related* by the relation R, while the second uses the definition of R to say precisely what that means in terms of x and y.

A **binary function**, or *function of two variables*, is a rule for taking an ordered pair of two inputs and using it to calculate an output. An example is $(x, y \mapsto x + y - 2)$, the rule which subtracts 2 from the sum of the two inputs. To graph a binary function, we need three dimensions, two for the inputs and one for the output. If g stands for the function $(x, y \mapsto x + y - 2)$, then these two expressions mean the same thing:

- $g(x, y)$;
- $x + y - 2$.

A **ternary relation**, or *relation between three variables*, is a set of ordered triples instead of a set of ordered pairs. An example is $\{x, y, z \mid x + y + z < 2\}$. We again use three dimensions to graph a ternary relation.

A **ternary function**, or *function of three variables*, is a rule for taking an ordered triple of three inputs and using it to calculate an output. An example is $(x, y, z \mapsto x + y + z - 2)$, the rule which subtracts 2 from the sum of the three inputs. To graph a ternary function, we need four dimensions, three for the inputs and one for the output.

A **quaternary relation**, or *relation between four variables*, is a set of ordered quadruples. An example is $\{x_1, x_2, x_3, x_4 \mid x_1 + x_2 + x_3 + x_4 < 2\}$. We again use four dimensions to graph a quaternary relation.

We can continue with quaternary functions, quinary functions, etc, which are functions of four or more variables; and we can continue with quinary relations, senary relations, etc, which are relations between five or more variables. (But around this point, most people stop using the '-ary' terms, because few people can remember them.)

There are various relationships between these different kinds of objects:

- The domain of a function of n variables is a relation between n variables (the same n variables).
- The range of a function of any number of variables is a set (a relation with 1 variable, the output).
- The graph of a function of n variables is the graph of a relation between $n+1$ variables (the n input variables plus the 1 output variable), which contains all of the information in the function.

For example, a binary function (a function of 2 variables) has a relation (a binary relation, a relation between 2 variables) as its domain, a set (a unary relation, a relation with 1 variable) as its range, and a ternary relation (a relation between 3 variables) as its graph. In particular, if $f(a,b) = c$, then $(a,b) \in \operatorname{dom} f$ (where $\operatorname{dom} f$ is the domain of f), $c \in \operatorname{ran} f$ (where $\operatorname{ran} f$ is the range of f), and $(a,b,c) \in \operatorname{gr} f$ (where $\operatorname{gr} f$ is the graph of f).

Definitions for functions of several variables

In order to form precise definitions of various concepts related to functions of several variables, it's handy to piggyback on the definitions for functions of one variable. This is *not* the way that the book writes its definitions, but it's the way that I prefer. So here are my definitions.

General principles

Recall that a *parametrized curve*, or *point-valued function*, takes a number to a point (in however many dimensions we're dealing with, typically 2 or 3 dimensions). That is, if C is a parametrized curve and t is a real number, then $C(t)$ is a point $P = (x,y)$, $P = (x,y,z)$, etc. Meanwhile, a *function of several variables* (however many variables we're dealing with, typically 2 or 3 variables) takes a point to a number; that is, if f is a function of 2 or 3 variables and $P = (x,y)$ or $P = (x,y,z)$ is a point in 2 or 3 dimensions, then $f(P) = f(x,y)$ or $f(P) = f(x,y,z)$ is a real number c. If we combine these by composition of functions, then $f \circ C$ is an ordinary function; that is, if t is a real number, then so is $(f \circ C)(t)$:

$$(f \circ C)(t) = f\Big(C(t)\Big) = f(P) = c.$$

From one-variable Calculus, you should know how to define various concepts (continuity, limits, differentiability, derivatives, differentials) for ordinary functions. It's easy to extend these concepts to vector- and point-valued functions (parametrized curves), since these simply consist of several ordinary functions (the coordinates or components). So to define these concepts for functions of several variables, we typically use a formula like this:

> If $f \circ C$ has a certain property whenever C does, no matter what C might be (as long as it has the property), then that's what it means for f to have that property.

This formula doesn't always work perfectly; for one thing, we often want to say more than just a Yes/No property, and it may not be obvious what matters about C or how to extract the appropriate information from the composites. Besides that, even when this formula would make perfect sense, sometimes some of the nice theorems that we would expect aren't always true, which means that we should look for a modifed definition that makes the theorems work. (That's what mathematicians really want from a definition; they're not handed down from on high but developed for the purpose of getting correct results.) Nevertheless, all of the definitions here will be based on something like this formula.

Continuity

Continuity follows the general formula precisely. A function f of several variables is **continuous** if, whenever C is a continuous parametrized curve, the composite $f \circ C$ is a continuous function. (It wouldn't be fair to expect $f \circ C$ to be continuous unless C is continuous as well as f, but if both C and f are continuous, then their composite ought to be as well.)

Sometimes we want to look at continuity in more detail; in general, to say that a function is continuous really means that it's continuous at every number in its domain. So for a function of several variables, we want to talk about continuity at particular points in its domain. A function f is **continuous** at a point P_0 in the domain of f if, whenever C is a parametrized curve and t_0 is a number such that

$C(t_0) = P_0$ and C is continuous at t_0, then $f \circ C$ is also continuous at t_0. Again, it wouldn't be fair to demand more than this if we're only asking f to be continuous at P_0.

An equivalent definition is to say that f is continuous at P_0 if f is defined at P_0 and, for every positive number ϵ, there is some positive number δ such that, whenever $\|P - P_0\| < \delta$ and f is defined at P, then $|f(P) - f(P_0)| < \epsilon$. This is essentially how it is defined in the textbook. However, this ϵ-δ stuff is rather less fun to work with. Ultimately, you have to say something like this some time, but I prefer to say it once, when giving the first definition in one-variable Calculus, and then never again.

Any function with a formula that is built out of the coordinate variables using only the usual operations is continuous wherever it is defined. (To be definite, the usual operations are addition, subtraction, multiplication, division, taking opposites, taking reciprocals, taking absolute values, raising to powers with constant exponents and/or positive bases, extracting roots with constant indexes and/or positive radicands, logarithms, the six trigonometric operations, and the six inverse trigonometric operations. Some notable operations *not* on this list are piecewise definitions and powers where the exponent varies and the base may be zero or negative.)

To prove this, you use the continuity of each component of a continuous parameterized curve and the one-variable theorem that any function built out of continuous functions using these operations is continuous. For example, if f and g are continuous at P_0 and I want to prove that $f + g$ is continuous at P_0, consider a parametrized curve C and a number t_0 such that $C(t_0) = P_0$ and C is continuous at t_0; by definition, $f + g$ is continuous at P_0 if, for each such C and t_0, $(f + g) \circ C$ is continuous at t_0. Since f is continuous at P_0, this means (by definition) that $f \circ C$ is continuous at t_0; similarly, since g is continuous at P_0, this means that $f \circ C$ is continuous at t_0. By a theorem in one-variable Calculus, since $f \circ C$ and $g \circ C$ are both continuous at t_0, so is their sum $(f \circ C) + (g \circ C)$. But $(f \circ C) + (g \circ C)$ is the same function as $(f + g) \circ C$, since they do the same thing to any input t:

$$\Big((f \circ C) + (g \circ C)\Big)(t) = (f \circ C)(t) + (g \circ C)(t) = f\Big(C(t)\Big) + g\Big(C(t)\Big);$$
$$\Big((f + g) \circ C\Big)(t) = (f + g)\Big(C(t)\Big) = f\Big(C(t)\Big) + g\Big(C(t)\Big).$$

Therefore, $(f + g) \circ C$ is continuous at t_0. Since this argument works for any relevant C and t_0, this proves that $f + g$ is continuous at P_0, as desired. (Similar arguments work for all of the other operations.)

Limits

To keep things simple, we'll only look at finite limits approaching a finite value; none of our limits will involve infinity in any role. (Things will become more complicated in another way shortly!)

There is a technicality about limits that's often ignored in one-variable Calculus, which is that the expression whose limit you're taking must be defined at numbers arbitrarily close to the number that the variable is approaching. It's often treated as a big deal that the function doesn't have to be defined at that number precisely, which is certainly true and important, but it still has to be defined *near* that number. For example (and assuming that we're only working with real numbers), you can't talk about the limit of $\sqrt{t}$ as $t \to -1$, because t can't get very close to -1 while $\sqrt{t}$ is defined. On the other hand, it's fine to talk about the limit as $t \to 0$, because even though $\sqrt{t}$ is undefined when $t < 0$, still $\sqrt{t}$ is defined when $t > 0$, which allows t to get arbitrarily close to 0. (But on the other other hand, you can't talk about the limit as $t \to 0^-$, because now this requires $t < 0$, which leaves $\sqrt{t}$ undefined again.)

A number t_0 is a **limit point** of a set D if it makes sense to talk about a function defined on D as having a limit approaching t_0, in other words if there exists a function whose domain is D (a constant function will do) that has a limit approaching t_0. (The term 'limit point' is traditional even in one dimension, even though I would normally call t_0 a number rather than a point.) This is equivalent to saying that there are numbers in D (other than possibly t_0 itself) that are arbitrarily close to t_0, in other words if, given any positive distance $\delta > 0$, there is at least some number t in the set D such that $0 < |t - t_0| < \delta$. But I prefer to think of the definition that has no δ (or ϵ) in it.

Keeping this technicality in mind, the **limit** approaching a point P_0 of a function f of several variables (which in symbols we can write as

$$\lim_{P\to P_0} f(P),$$

that is

$$\lim_{(x,y)\to P_0} f(x,y)$$

in 2 dimensions or

$$\lim_{(x,y,z)\to P_0} f(x,y,z)$$

in 3 dimensions) is the unique number L (if this exists) such that, whenever C is a parametrized curve and t_0 is a number, if $C(t) = P_0$ when and only when $t = t_0$, and if C is continuous at t_0, and if t_0 is a limit point of the domain of $f \circ C$, then L is the limit of $f \circ C$ approaching t_0. In other words (ignoring the fine print),

$$\lim_{P\to P_0} f(P) = L$$

if

$$\lim_{t\to t_0} f\Big(C(t)\Big) = L$$

whenever

$$\lim_{t\to t_0} C(t) = P_0.$$

The point of all of that is this: the limit of one of these composites is basically the limit of the function along a particular curve. If the function is undefined along the curve, then we don't expect its limit to exist, and this is what the clause about limit points takes care of. We also don't want to worry about $f(P_0)$ itself, since f might not be continuous at P_0, which is why $C(t)$ is not allowed to be P_0 except when $t = t_0$. So we're only looking at certain curves that are *appropriate* to the problem. Then, in order for the limit to exist overall, the limit must exist along each appropriate curve and be the same along all of them.

If for any appropriate curve, there is no limit along that curve, then the limit overall does not exist. Besides that, if there are two appropriate curves such that the limits along them are different, then again the limit does not exist overall. It is in this way that one generally proves that a limit does not exist, when it doesn't. When the limit does exist, however, then you usually need to find a general argument to show that it does and what it is, because you can't actually check every individual curve. Fortunately, we have a theorem that

$$\lim_{P\to P_0} f(P) = f(P_0)$$

whenever f is continuous at P_0 (assuming that P_0 is a limit point of the domain of f), as in one-variable Calculus.

One often talks about limits with restrictions on the manner of approaching the point. For example, instead of saying (x, y) approaches $(2, 3)$, we might say that (x, y) approaches $(2, 3)$ *while* $x \neq y$. (An analogue in one-variable Calculus is, for example, $t \to 0^-$; that is, $t \to 0$ while $t < 0$.) Technically, this is handled by modifying the function so that it is defined only when the given restriction is met (so in this example, the function would be undefined when $x = y$). That is,

$$\lim_{\substack{(x,y)\to(2,3)\\ x\neq y}} f(x,y) = \lim_{(x,y)\to(2,3)} (f(x,y) \text{ for } x \neq y),$$

where by '$f(x, y)$ for $x \neq y$' I mean $f(x, y)$ if $x \neq y$ but something undefined if $x = y$.

Differentiability

The way that differentiability fits in with composition of functions is the chain rule

$$(f \circ g)'(t) = f'\Big(g(t)\Big)g'(t).$$

Following the general principle, we replace g with a parametrized curve C, and the values of the derivatives of this (replacing $g'(t)$) are vectors. However, the composite is an ordinary function, so the derivative of f should multiply by a vector to get a scalar. One way to do this is to multiply a vector by a vector with the dot product, so the derivative of a function of several variables should also be a vector. (Since we want this concept to make sense even when lengths and angles don't apply, this vector is going to have to be a *row* vector; see page 9 of the handout on vectors.) There are actually several sorts of derivatives in higher dimensions, and we'll come back to this subject later; but the one which is a vector will provide the definition of differentiability.

We say that the function f is **differentiable** at some point P_0 if there exists a (row) vector $\mathbf{v}$ such that, whenever C is a parametrized curve and t_0 is a number such that $C(t_0) = P_0$ and C is differentiable at t_0, then $f \circ C$ is also differentiable at t_0 and furthermore $(f \circ C)'(t_0) = \mathbf{v} \cdot C'(t_0)$. If f is differentiable at every point P_0 in its domain, then f is simply *differentiable*.

This vector $\mathbf{v}$ is called the **gradient** of f at P_0 and may be written as $\nabla f(P_0)$ (although $f'(P_0)$ would make a lot of sense), so the basic rule is

$$(f \circ C)'(t) = \nabla f\Big(C(t)\Big) \cdot C'(t).$$

Higher differentiability

Where a function f is differentiable, the components of its gradients define some more functions, called the **partial derivatives** of f. (We will do more with these partial derivatives later on.) Wherever the partial derivatives are themselves continuous, the original function is **continuously differentiable**. Where the partial derivatives are themselves differentiable, the original function is **twice differentiable**. Where the partial derivatives are continuously differentiable, the original function is **twice continuously differentiable**. Etc etc etc. (As in one-variable Calculus, there is a theorem that a differentiable function must be continuous, so a twice-differentiable function must be continuously differentiable, etc.)

Where this goes on forever, the function is **infinitely differentiable**: it is differentiable, its partial derivatives are differentiable, their partial derivatives are differentiable, etc. Any function built out of the usual operations is infinitely differentiable except at certain exceptional places where a derivative fails to exist, such as when taking the absolute value or square root of zero. But to prove this, it's best to look at how to calculate the derivatives, which I'll get to next.

Differentials and differential forms

Differential forms are, broadly speaking, expressions that may have *differentials* in them. They have many uses in modern science and engineering, even though they are not traditionally covered explicitly in math class. They are covered somewhat, however, and they are there whenever you differentiate or integrate, even if you don't recognize them. They are especially prominent in multivariable Calculus, and I want to bring them to your attention; you'll find that symbols that otherwise seem meaningless or merely mnemonic can be understood literally (sometimes with slight changes) as differential forms.

Examples

The most basic examples of differential forms are differentials such as $\mathrm{d}x$ and $\mathrm{d}y$. In general, if u is any quantity that might change, then $\mathrm{d}u$ is intended to be a related quantity whose value is an infinitely small change in u, or rather the amount by which the value of u changes when an infinitely small (or arbitrarily small) change is made. (I will make this precise later on.)

Besides the differentials themselves, differential forms can be constructed by applying arithmetic operations, so $\mathrm{d}x + \mathrm{d}y$, $\mathrm{d}x\,\mathrm{d}y$, and $\sqrt{\mathrm{d}x}$ are all differential forms. In all of these expressions, we adopt an order of operations in which the differential operator d is applied before any arithmetic operator; for example, $\mathrm{d}x^2$ means $(\mathrm{d}x)^2$, not $\mathrm{d}(x^2)$ (which is $\mathrm{d}u$ when $u = x^2$ and turns out to equal $2x\,\mathrm{d}x$). Additionally, we can include ordinary quantities in these expressions, so $x + \mathrm{d}x$, $3\,\mathrm{d}x + x^2\,\mathrm{d}y + \mathrm{e}^y\,\mathrm{d}z$, and $x \ln(y/\mathrm{d}z)$ are also differential forms. We can also use differentials of differentials, such as d^2x (which means $\mathrm{d}(\mathrm{d}x)$, the differential of $\mathrm{d}x$), although we won't need such *higher-order* differentials in this course. Besides all of this, any ordinary expression counts as a differential form in a degenerate way; thus, x, y^2, and $2xy^3$ are also differential forms (of order zero).

Some differential forms are more useful than others. Of those listed above, besides the differentials and the non-differential quantities, the ones most likely to appear in a real problem are $\mathrm{d}x + \mathrm{d}y$ and $3\,\mathrm{d}x + x^2\,\mathrm{d}y + \mathrm{e}^y\,\mathrm{d}z$. These consist of any number of terms, each of which is the product of an ordinary quantity (possibly the constant 1) and the differential of an ordinary quantity. Differential forms with this property are most commonly found in practice. We will use other differential forms, such as $3x\,|\mathrm{d}y|$ and $\sqrt{\mathrm{d}x^2 + \mathrm{d}y^2}$; however, you might be able to see how even these forms are differential of *degree* 1 in a sense similar to the degree of a polynomial.

All of the examples so far are differential forms of *rank* 1; there are also differential forms of higher rank, such as $\mathrm{d}x \wedge \mathrm{d}y$, which are written using a new operation, the *wedge product*. We will not use these until later; these notes are only about differential forms of rank 1, or 1-forms for short. (Ordinary quantities have rank 0.)

Evaluating differential forms

In this class, we generally assume that any ordinary quantity (that is any 0-form) is a function of 2 or 3 ordinary variables, $P = (x, y)$ or $P = (x, y, z)$. Thus, we evaluate ordinary quantites (0-forms) by specifying specific values for the variables that comprise P. For example, to evaluate $u = x^2 + xy$ when $x = 2$ and $y = 3$, we may write

$$u|_{P=(2,3)} = (x^2 + xy)|_{(x,y)=(2,3)} = (2)^2 + (2)(3) = 10.$$

To evaluate a differential form (of order 1), we need not only a point (a value of P) but also a vector (a value of $\mathrm{d}P = \langle \mathrm{d}x, \mathrm{d}y \rangle$ or $\mathrm{d}P = \langle \mathrm{d}x, \mathrm{d}y, \mathrm{d}z \rangle$). So for example, to evaluate $\alpha = 3\,\mathrm{d}x + x^2\,\mathrm{d}y + \mathrm{e}^y\,\mathrm{d}z$ when $x = 2$, $y = 3$, $z = 4$, $\mathrm{d}x = 0.05$, $\mathrm{d}y = -0.01$, and $\mathrm{d}z = 0$, we may write

$$\begin{aligned}\alpha\big|_{\substack{P=(2,3,4),\\ \mathrm{d}P=\langle 0.05,-0.01,0\rangle}} &= (3\,\mathrm{d}x + x^2\,\mathrm{d}y + \mathrm{e}^y\,\mathrm{d}z)\big|_{\substack{(x,y,z)=(2,3,4),\\ \langle \mathrm{d}x,\mathrm{d}y,\mathrm{d}z\rangle=\langle 0.05,-0.01,0\rangle}} \\ &= 3(0.05) + (2)^2(-0.01) + \mathrm{e}^{(3)}(0) = 0.11.\end{aligned}$$

(Differential forms are often denoted with Greek letters such as 'α', although they don't have to be.) We say that α has been evaluated *at* the point $P = (2, 3, 4)$ *along* the vector $\mathrm{d}P = \langle 0.05, -0.01, 0\rangle$. (The components of $\mathrm{d}P$ don't need to be small, since the definition makes sense in any case, but in applications that's usually what matters; after all, $\mathrm{d}P$ is supposed to be a *small* change in position.)

(To evaluate higher-order differential forms (those that involve higher-order differentials), we need to specify additional vectors such as $\mathrm{d}^2P = \langle \mathrm{d}^2x, \mathrm{d}^2y, \mathrm{d}^2z\rangle$, etc. However, we won't need that level of generality in this course.)

Differential forms as vectors

A differential form $\alpha = M\,\mathrm{d}x + N\,\mathrm{d}y + O\,\mathrm{d}z$ may be viewed as a dot product $\alpha = \langle M, N, O\rangle \cdot \langle \mathrm{d}x, \mathrm{d}y, \mathrm{d}z\rangle = \mathbf{V} \cdot \mathrm{d}P$. For example, if $\alpha = 3\,\mathrm{d}x + x^2\,\mathrm{d}y + \mathrm{e}^y\,\mathrm{d}z$, then $\alpha = \langle 3, x^2, \mathrm{e}^y\rangle \cdot \mathrm{d}P$; conversely, if $\mathbf{V} = \langle 3, x^2, \mathrm{e}^y\rangle$, then

$$\mathbf{V} \cdot \mathrm{d}P = \langle 3, x^2, \mathrm{e}^y\rangle \cdot \langle \mathrm{d}x, \mathrm{d}y, \mathrm{d}z\rangle = 3\,\mathrm{d}x + x^2\,\mathrm{d}y + \mathrm{e}^y\,\mathrm{d}z.$$

(We can recover $\mathbf{V}$ from α formally by evaluating α when $\mathrm{d}P$ is $\langle \mathbf{i}, \mathbf{j}\rangle$ or $\langle \mathbf{i}, \mathbf{j}, \mathbf{k}\rangle$, but there's probably no need to think about that explicitly.)

Even in circumstances where it makes no sense to interpret a change in the values of (x, y, z) as a vector in the geometric sense (with length and direction), in which case dot products involving them generally have no meaning, it is traditional to write differential forms in this way and to focus on $\mathbf{V}$ rather than on α as the object of study. In this case, we need to think of $\mathbf{V}$ as a *row* vector. Regardless of whether $\mathbf{V}$ has geometric significance as a vector, it can be helpful to visualize it as one.

When calculations with a row vector need to be performed, ultimately it is the differential form $\alpha = \mathbf{V} \cdot \mathrm{d}P$ that matters. It's more common to see $\mathbf{V} \cdot \mathrm{d}\mathbf{r}$, where as usual the vector $\mathbf{r} = P - \mathrm{O}$ (where O is $(0, 0)$ or $(0, 0, 0)$) satisfies $\mathrm{d}\mathbf{r} = \mathrm{d}P$. Sometimes $\mathbf{V} \cdot \mathrm{d}\mathbf{r}$ is even regarded as merely a mnemonic notation (especially in the context of defining integrals such as those in Section 15.2 of the textbook), but it can be taken literally, just as $\mathrm{d}y/\mathrm{d}x$ (which is also sometimes regarded as merely mnemonic) can be taken literally as the result of a division of differentials. In any case, people do write $\mathbf{V} \cdot \mathrm{d}\mathbf{r}$ (even in the textbook), so it can be nice to know what it means!

In the textbook, they also sometimes write $\mathrm{d}\mathbf{r} = \mathbf{T}\,\mathrm{d}s$, where $\mathrm{d}s$ (which is not really the differential of anything in space as a whole) is the magnitude $\mathrm{d}s = |\mathrm{d}\mathbf{r}|$ and $\mathbf{T} = \widehat{\mathrm{d}\mathbf{r}}$, the unit vector in the direction of $\mathrm{d}\mathbf{r}$. This is sometimes useful when thinking about things geometrically, but it's not necessary for purposes of calculation. In 2 dimensions, we can also take cross products (using the rule $\langle a, b\rangle \times \langle c, d\rangle = ad - bc$); for example, if $\mathbf{V} = \langle 3, x^2\rangle$, then

$$\mathbf{V} \times \mathrm{d}\mathbf{r} = \langle 3, x^2\rangle \times \langle \mathrm{d}x, \mathrm{d}y\rangle = 3\,\mathrm{d}y - x^2\,\mathrm{d}x.$$

(This requires that changes in x and y make sense as having a geometric length even when $\mathbf{V}$ is regarded as merely a row vector, so it doesn't come up as often.) If you use $\times\langle c, d\rangle = \langle d, -c\rangle$, so that $\mathbf{u} \times \mathbf{v} = \mathbf{u} \cdot \times\mathbf{v}$, then you can write $\mathbf{V} \times \mathrm{d}\mathbf{r}$ as $\mathbf{V} \cdot \times\mathrm{d}\mathbf{r}$; the book sometimes writes this as $\mathbf{V} \cdot \mathbf{n}\,\mathrm{d}s$, where $\mathrm{d}s = |\times\mathrm{d}\mathbf{r}| = |\mathrm{d}\mathbf{r}|$ again, and now $\mathbf{n} = \widehat{\times\mathrm{d}\mathbf{r}} = \times\mathbf{T}$ is the direction perpendicular and clockwise from $\mathrm{d}\mathbf{r}$. Again, sometimes this is useful when thinking about the geometry, but you don't need it for doing calculations.

This is all especially common when $\mathbf{V}$ is the output of a *vector field*, that is a vector-valued function of several variables. For example, if $\mathbf{F}(x, y) = \langle 3, x^2\rangle$, then

$$\mathbf{F}(x, y) \cdot \mathrm{d}\mathbf{r} = \langle 3, x^2\rangle \cdot \langle \mathrm{d}x, \mathrm{d}y\rangle = 3\,\mathrm{d}x + x^2\,\mathrm{d}y,$$

and

$$\mathbf{F}(x, y) \times \mathrm{d}\mathbf{r} = \langle 3, x^2\rangle \times \langle \mathrm{d}x, \mathrm{d}y\rangle = 3\,\mathrm{d}y - x^2\,\mathrm{d}x.$$

So in Section 15.2, which is really about integrating differential 1-forms along curves, the book spends most of its time talking about integrating vector fields along curves (and occasionally integrating them across curves in 2 dimensions). What's really going on is that you integrate the vector field $\mathbf{F}$ by integrating one of the two differential forms above (usually the first one).

Differentials and the rules of differentiation

In one-variable Calculus, one sometimes sees the Chain Rule expressed as

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\mathrm{d}y}{\mathrm{d}u} \cdot \frac{\mathrm{d}u}{\mathrm{d}x},$$

but the Chain Rule is a nontrivial fact that cannot be proved by simply cancelling factors. I prefer to state the Chain Rule as

$$\mathrm{d}\Big(f(u)\Big) = f'(u)\,\mathrm{d}u.$$

the point is that the *same* function f' appears regardless of which argument u we use.

Even this is more abstract than how the Chain Rule is applied. For example, suppose that you have discovered (say from the definition as a limit) that the derivative of $f(x) = \sin x$ is $f'(x) = \cos x$. Since $f'(x)$ may be defined as $\dfrac{\mathrm{d}\big(f(x)\big)}{\mathrm{d}x}$, this derivative can be expressed in differential form without even bothering to name the functions involved:

$$\mathrm{d}(\sin x) = \cos x\,\mathrm{d}x.$$

Once you know this, you know something even more general:

$$\mathrm{d}(\sin u) = \cos u\,\mathrm{d}u$$

for any other differentiable quantity u; the Chain Rule is the power to derive this equation from the previous one. Thus, using $u = x^2$ (to continue the example),

$$\mathrm{d}\Big(\sin\,(x^2)\Big) = \cos\,(x^2)\,\mathrm{d}(x^2) = \cos\,(x^2)(2x\,\mathrm{d}x) = 2x\cos\,(x^2)\,\mathrm{d}x.$$

You may now divide both sides of this equation by $\mathrm{d}x$ if you wish, but the basic calculation involves only rules for differentials.

For the record, here are the rules for differentiation that you should already know, expressed using differentials:

- The Constant Rule: $\mathrm{d}(K) = 0$ if K is constant.
- The Sum Rule: $\mathrm{d}(u + v) = \mathrm{d}u + \mathrm{d}v$.
- The Translate Rule: $\mathrm{d}(u + C) = \mathrm{d}u$ if C is constant.
- The Difference Rule: $\mathrm{d}(u - v) = \mathrm{d}u - \mathrm{d}v$.
- The Product Rule: $\mathrm{d}(uv) = v\,\mathrm{d}u + u\,\mathrm{d}v$.
- The Multiple Rule: $\mathrm{d}(ku) = k\,\mathrm{d}u$ if k is constant.
- The Quotient Rule: $\mathrm{d}\left(\dfrac{u}{v}\right) = \dfrac{v\,\mathrm{d}u - u\,\mathrm{d}v}{v^2}$.
- The Power Rule: $\mathrm{d}(u^n) = nu^{n-1}\,\mathrm{d}u$ if n is constant.
- The Exponentiation Rule: $\mathrm{d}(\exp u) = \exp u\,\mathrm{d}u$ (where $\exp u$ means e^u).
- The Logarithm Rule: $\mathrm{d}(\ln u) = \dfrac{\mathrm{d}u}{u}$.
- The Sine Rule: $\mathrm{d}(\sin u) = \cos u\,\mathrm{d}u$.
- The Cosine Rule: $\mathrm{d}(\cos u) = -\sin u\,\mathrm{d}u$.
- The Tangent Rule: $\mathrm{d}(\tan u) = \sec^2 u\,\mathrm{d}u$.
- The Cotangent Rule: $\mathrm{d}(\cot u) = -\csc^2 u\,\mathrm{d}u$.
- The Secant Rule: $\mathrm{d}(\sec u) = \tan u \sec u\,\mathrm{d}u$.
- The Cosecant Rule: $\mathrm{d}(\csc u) = -\cot u \csc u\,\mathrm{d}u$.
- The Arcsine Rule: $\mathrm{d}(\mathrm{asin}\,u) = \dfrac{\mathrm{d}u}{\sqrt{1-u^2}}$ (where $\mathrm{asin}\,u$ means $\sin^{-1} u$).
- The Arccosine Rule: $\mathrm{d}(\mathrm{acos}\,u) = -\dfrac{\mathrm{d}u}{\sqrt{1-u^2}}$.

- The Arctangent Rule: $\mathrm{d}(\operatorname{atan} u) = \dfrac{\mathrm{d}u}{u^2+1}$.
- The Arccotangent Rule: $\mathrm{d}(\operatorname{acot} u) = -\dfrac{\mathrm{d}u}{u^2+1}$.
- The Arcsecant Rule: $\mathrm{d}(\operatorname{asec} u) = \dfrac{\mathrm{d}u}{|u|\sqrt{u^2-1}}$.
- The Arccosecant Rule: $\mathrm{d}(\operatorname{acsc} u) = -\dfrac{\mathrm{d}u}{|u|\sqrt{u^2-1}}$.
- The Chain Rule: $\mathrm{d}(f(u)) = f'(u)\,\mathrm{d}u$ if f is a function of one variable that's differentiable at u.
- The First Fundamental Theorem of Calculus: $\mathrm{d}\Big(\int_{t=u}^{v} f(t)\,\mathrm{d}t\Big) = f(v)\,\mathrm{d}v - f(u)\,\mathrm{d}u$ if f is a function of one variable that's continuous between u and v.

(The last one might not be familiar to you in such a general form, but it can be handy.)

Notice that every one of the rules above turns the differential on the left into a sum of terms (possibly only one term, or none in the case of the Constant Rule), each of which is an ordinary expression multiplied by a differential (or something algebraically equivalent to this). This is a kind of differential form; more precisely, these are *linear differential* 1*-forms* (which are also called *exterior differential* 1*-forms*).

Here is an example of how to use the rules, step by step, to find a differential. Specifically, I'll find the differential of $x^2y + \sin(z^2)$. (In one-variable Calculus, you might consider this if x, y, and z all happen to be functions of some other variable t; but in multivariable Calculus, the same calculation will apply even when the variables x, y, and z are all independent.)

$$\begin{aligned}
\mathrm{d}\Big(x^2y + \sin(z^2)\Big) &= \mathrm{d}(x^2y) + \mathrm{d}\Big(\sin(z^2)\Big) \\
&= y\,\mathrm{d}(x^2) + x^2\,\mathrm{d}y + \cos(z^2)\,\mathrm{d}(z^2) \\
&= y(2x^{2-1}\,\mathrm{d}x) + x^2\,\mathrm{d}y + \cos(z^2)(2z^{2-1}\,\mathrm{d}z) \\
&= 2xy\,\mathrm{d}x + x^2\,\mathrm{d}y + 2z\cos(z^2)\,\mathrm{d}z.
\end{aligned}$$

Here I've used, in turn, the sum rule, the product and sine rules (one in one term and the other in the other term), the power rule (in two places), and finally some algebra to simplify. Of course, you can usually do this much faster; with practice, you can jump immediately to the second-to-last line by applying the next rule whenever one rule results in a differential; then you only need one more step to simplify it algebraically. Often you can even do some of the algebra in your head immediately (like simplifying x^{2-1} to x, so that $\mathrm{d}(x^2)$ immediately becomes $2x\,\mathrm{d}x$).

Partial derivatives

If $f(x,y,z)$ (for example) can be expressed using the usual operations (and possibly even if it cannot), then its differential will come out as

$$\mathrm{d}\Big(f(x,y,z)\Big) = f_1(x,y,z)\,\mathrm{d}x + f_2(x,y,z)\,\mathrm{d}y + f_3(x,y,z)\,\mathrm{d}z$$

for some functions f_1, f_2, and f_3. These functions are the **partial derivatives** of f. Since subscripts can be used for many things, a better notation for f_1, f_2, and f_3 is D_1f, D_2f, and D_3f (respectively); compare the notation Df for f' that is sometimes used in single-variable Calculus. For example, if $f(x,y,z) = x^2y + \sin(z^2)$, then

$$\mathrm{d}\Big(f(x,y,z)\Big) = \mathrm{d}\Big(x^2y + \sin(z^2)\Big) = 2xy\,\mathrm{d}x + x^2\,\mathrm{d}y + 2z\cos(z^2)\,\mathrm{d}z$$

(as I calcluated earlier), so

$$\begin{aligned}
D_1f(x,y,z) &= 2xy, \\
D_2f(x,y,z) &= x^2, \text{ and} \\
D_3f(x,y,z) &= 2z\cos(z^2).
\end{aligned}$$

If instead we write u for $f(x, y, z)$, then we have a different notation for the coefficients on the differentials:

$$\mathrm{d}u = \left(\frac{\partial u}{\partial x}\right)_{y,z} \mathrm{d}x + \left(\frac{\partial u}{\partial y}\right)_{x,z} \mathrm{d}y + \left(\frac{\partial u}{\partial z}\right)_{x,y} \mathrm{d}z.$$

(The symbol '∂' is a variation on the lowercase Greek Delta, 'δ'. It is usually not pronounced directly; instead, you read the entire expression as described below.) So for example, if $u = x^2y + \sin(z^2)$, then

$$\mathrm{d}u = \mathrm{d}\Big(x^2y + \sin(z^2)\Big) = 2xy\,\mathrm{d}x + x^2\,\mathrm{d}y + 2z\cos(z^2)\,\mathrm{d}z$$

again, so

$$\begin{aligned}\left(\frac{\partial u}{\partial x}\right)_{y,z} &= 2xy,\\ \left(\frac{\partial u}{\partial y}\right)_{x,z} &= x^2, \text{ and}\\ \left(\frac{\partial u}{\partial z}\right)_{x,y} &= 2z\cos(z^2).\end{aligned}$$

This $\left(\dfrac{\partial u}{\partial x}\right)_{y,z}$ is the **partial derivative** of u with respect to x, fixing y and z, which tells you how much u changes relative to the change in x as long as y and z remain the same. All of the information in this notation is necessary to avoid ambiguity, but in practice people usually write simply $\dfrac{\partial u}{\partial x}$, call this simply the partial derivative of u with respect to x, and expect you to guess from context what other variables are remaining fixed.

Of course, people also mix notation for f with notation for u, writing $D_x f$, f_x, $\dfrac{\partial f}{\partial x}$, and so on, as well as u_x, u_1, $D_1 u$, and so on. Technically, notation with numbers makes sense only when applied to the name of a function, because the arguments of that function come in a specific order; while notation referring to the variables used does *not* make sense when applied to the name of a function, since one could use any variables as the arguments of the function (although it does make sense when applied to an expression such as $f(x, y, z)$, in which these variables have been specified). In practice, however, people usually use the variables x, y, z in that order; then there is no confusion.

Defining differentials

Recall from the handout on definitions for functions of several variables that the function f is **differentiable** at the point P_0 if there exists a row vector $\nabla f(P_0)$ such that, for every differentiable parametrized curve C and real number t_0, if $C(t_0)$ exists and equals P_0, then the composite function $f \circ C$ is differentiable at t_0 and furthermore $(f \circ C)'(t_0) = \nabla f(P_0) \cdot C'(t_0)$. This makes ∇f a vector field, called the **gradient** of f, that is defined wherever f is differentiable. (The symbol '∇' is variously pronounced 'Atled', 'Nabla', and 'Del'; people also write $\operatorname{grad} f$ for ∇f.)

If $u = f(P)$ and f is differentiable, then we write

$$\mathrm{d}u = \nabla f(P) \cdot \mathrm{d}P = \nabla f(P) \cdot \mathrm{d}\mathbf{r},$$

where $\mathbf{r}$ is $P - \mathrm{O}$ (P minus the origin), as usual. If you think of ∇f as a derivative of f, then this is simply taking the Chain Rule as a definition. There are two good things about this definition of $\mathrm{d}u$. First of all, all of the usual rules of differentiation are actually true of it; because the definition ultimately refers to ordinary functions, we can prove each rule in the list on pages 3 and 4 by using the corresponding result for ordinary functions. The other good thing about this definition is that when we evaluate a differential at a given point and vector, then the result is one of the derivatives $(f \circ C)'(t_0)$ that appear in the definition above.

Specifically, fixing a point P_0 and a vector $\mathbf{v}_0$, let $C(t) = P_0 + t\mathbf{v}_0$; then C is a differentiable curve with $C(0) = P_0$ and $C'(0) = \mathbf{v}_0$, so

$$\mathrm{d}u\big|_{\substack{P=P_0,\\ \mathrm{d}P=\mathbf{v}_0}} = \nabla f(P_0)\cdot\mathbf{v}_0 = \nabla f\Big(C(0)\Big)\cdot C'(0) = (f\circ C)'(0)$$

when $u = f(P)$. If $\mathbf{v}_0$ happens to be a unit vector (a *direction*), then $\nabla f(P_0)\cdot\mathbf{v}_0$ is called the **directional derivative** of f at P_0 in the direction of $\mathbf{v}_0$. In general, the directional derivative in the direction of $\mathbf{v}_0$ is $\nabla f(P_0)\cdot\hat{\mathbf{v}}_0$ (where $\hat{\mathbf{v}} = \mathbf{v}/|\mathbf{v}|$ is the unit vector in the direction of $\mathbf{v}$); however, be careful, because some people use the term 'directional derivative' for $\nabla f(P_0)\cdot\mathbf{v}_0$ in the general case (since it's important but there is no standard name for it). In particular, the directional derivatives parallel to the coordinate axes —that is $\nabla f(P_0)\cdot\mathbf{i}$, $\nabla f(P_0)\cdot\mathbf{j}$, and (in 3 dimensions) $\nabla f(P_0)\cdot\mathbf{k}$— are simply the partial derivatives of f at P_0.

Because $\mathrm{d}\Big(f(P)\Big) = \nabla f(P)\cdot\mathrm{d}P = \nabla f(P)\cdot\mathrm{d}\mathbf{r}$, the value of the gradient may also be written as $\nabla f(P) = \mathrm{d}\Big(f(P)\Big)\Big/\mathrm{d}P = \mathrm{d}\Big(f(P)\Big)\Big/\mathrm{d}\mathbf{r}$ (although we cannot define division by a vector in general). An even simpler notation for $\nabla f(P)$ would be $f'(P)$, but this is traditionally not used, because there are many notions of derivative of f (such as the directional derivatives and the partial derivatives); even though the gradient is the most general derivative, it is commonly thought that f' would be ambiguous in this context. (When we start differentiating vector fields near the end of this course, there will be another reason that it's convenient to have a symbol ∇ that we can manipulate more easily than the tiny tick mark on f'.)

Gradients

If f is a function of (say) 3 variables, then the definition of differential above states that

$$\mathrm{d}\Big(f(x,y,z)\Big) = \nabla f(x,y,z)\cdot\mathrm{d}(x,y,z) = \nabla f(x,y,z)\cdot\langle\mathrm{d}x,\mathrm{d}y,\mathrm{d}z\rangle;$$

meanwhile, the definition of partial derivatives states that

$$\begin{aligned}\mathrm{d}\Big(f(x,y,z)\Big) &= D_1f(x,y,z)\,\mathrm{d}x + D_2f(x,y,z)\,\mathrm{d}y + D_3f(x,y,z)\,\mathrm{d}z\\ &= \langle D_1f(x,y,z), D_2f(x,y,z), D_3f(x,y,z)\rangle\cdot\langle\mathrm{d}x,\mathrm{d}y,\mathrm{d}z\rangle.\end{aligned}$$

In other words,

$$\nabla f(x,y,z) = \langle D_1f(x,y,z), D_2f(x,y,z), D_3f(x,y,z)\rangle = \left\langle \frac{\partial\Big(f(x,y,z)\Big)}{\partial x}, \frac{\partial\Big(f(x,y,z)\Big)}{\partial y}, \frac{\partial\Big(f(x,y,z)\Big)}{\partial z}\right\rangle.$$

Put more simply,

$$\nabla f = \langle D_1f, D_2f, D_3f\rangle,$$

or even

$$\nabla = \langle D_1, D_2, D_3\rangle.$$

The gradient has the same information as the differential, and the partial derivatives are the componenets of the gradient, so any one of these (the gradient, the partial derivatives, or the differential) may be used to solve any problem. The differential is usually the most useful for direct calculation, which is one reason why I use it heavily. However, if we have a geometric notion of length available to allow us to think of row vectors (such as the gradient) as the same as column vectors (the usual ones, going between points), then the gradient is easier to visualize.

For reference, here are a bunch of relationships between differentials, partial derivatives, and gradients, assuming that $u = f(x, y, z)$:

$$\mathrm{d}u = \left(\frac{\partial u}{\partial x}\right)_{y,z} \mathrm{d}x + \left(\frac{\partial u}{\partial y}\right)_{x,z} \mathrm{d}y + \left(\frac{\partial u}{\partial z}\right)_{x,y} \mathrm{d}z;$$

$$\mathrm{d}u = D_1 f(x, y, z)\,\mathrm{d}x + D_2 f(x, y, z)\,\mathrm{d}y + D_3 f(x, y, z)\,\mathrm{d}z;$$

$$D_1 f(x, y, z) = \left(\frac{\partial u}{\partial x}\right)_{y,z},\ D_2 f(x, y, z) = \left(\frac{\partial u}{\partial y}\right)_{x,z},\ D_3 f(x, y, z) = \left(\frac{\partial u}{\partial z}\right)_{x,y};$$

$$\nabla f(x, y, z) = \langle D_1 f(x, y, z), D_2 f(x, y, z), D_3 f(x, y, z)\rangle;$$

$$\nabla f(x, y, z) = \left\langle \left(\frac{\partial u}{\partial x}\right)_{y,z}, \left(\frac{\partial u}{\partial y}\right)_{x,z}, \left(\frac{\partial u}{\partial z}\right)_{x,y} \right\rangle;$$

$$\mathrm{d}u = \nabla f(x, y, z) \cdot \langle \mathrm{d}x, \mathrm{d}y, \mathrm{d}z\rangle;$$

$$\mathrm{d}u|_{\langle \mathrm{d}x, \mathrm{d}y, \mathrm{d}z\rangle = \mathbf{v}} = \nabla f(x, y, z) \cdot \mathbf{v}.$$

Applications of differentiation

There are various applications of differentiation of functions of several variables, some analogous to the applications for functions of one variable, some new.

Tangents and normal lines

If f is a function of 2 (or 3) variables and P_0 is a point in 2 (or 3) dimensions, then the level curve (or surface) of f through P_0 is given by the equation $f(P) = f(P_0)$, where $P = (x, y)$ (or (x, y, z), as usual). (The function f and the point P_0 have already been fixed, but the point P is allowed to vary, so this is an equation in our 2 (or 3) variables, as it should be.) If f is differentiable at P_0 and the gradient of f is nonzero at P_0, then this level curve (or surface) has a **tangent** line (or plane) through P_0, given by the equation $\nabla f(P_0) \cdot (P - P_0) = 0$. Finally, perpendicular to this tangent line (or plane), there is a **normal** line (always a line!) through P_0, with parametrization $P = P_0 + t\,\nabla f(P_0)$ in the parameter t.

Writing u for $f(P)$, the equation for the level curve (or surface) is $u = u|_{P=P_0}$. Writing Δu for $f(P + \Delta P) - f(P)$, a quantity that depends on both a point P and a vector ΔP, another equation for the level curve (or surface) is $\Delta u|_{\substack{P=P_0,\\ \Delta P=P-P_0}} = 0$. That is, you take the expression for Δu, which says how much u changes between two points, put P_0 in for the starting point P, and then put $P - P_0$ in for the difference ΔP between the two points. Since the value of u shouldn't change on the level curve (or surface), this difference Δu should be zero. (Notice that the meaning of P changes over the course of this substitution; originally it refers to the starting point, which we set to P_0, but afterwards it refers to another point on the level curve (or surface), so we set the displacement ΔP between the two points to $P - P_0$.)

The tangent line (or plane) is given by a very similar equation, except that now we look at how the curve (or surface) is changing infinitesimally at P_0 and extend this out to arbitrary distances. Thus, the equation $\Delta u = 0$ for the level curve (or surface) becomes $\mathrm{d}u = 0$ for the tangent line (or plane). However, we're still looking for the values of u in the same place, so the full equation is $\mathrm{d}u|_{\substack{P=P_0,\\ \mathrm{d}P=P-P_0}} = 0$. If you follow the definition of differential from page 5 and 6 above, then you'll see that this means precisely $\nabla f(P_0) \cdot (P - P_0) = 0$.

For example, if $u = xy$ and $P_0 = (2, 3)$, then the level curve is $xy = (2)(3)$, or simply $xy = 6$. (Replace x with 2 and y with 3 on the right-hand side.) Alternatively, $\Delta u = (x + \Delta x)(y + \Delta y) - xy = y\,\Delta x + x\,\Delta y + \Delta x\,\Delta y$, so the level curve is $(3)(x - 2) + (2)(y - 3) + (x - 2)(y - 3) = 0$. (Replace x with 2, y with 3, Δx with $x - 2$, and Δy with $y - 3$.) This also simplifies to $xy = 6$.

That was obviously more work than necessary for the level curve, but now apply the same technique to the differential to get the tangent line: $\mathrm{d}u = y\,\mathrm{d}x + x\,\mathrm{d}y$, so the tangent line is $(3)(x - 2) + (2)(y - 3) = 0$. (Replace x with 2, y with 3, $\mathrm{d}x$ with $x - 2$, and $\mathrm{d}y$ with $y - 3$.) This simplifies to $3x + 2y = 12$, and now we learnt something that we didn't know before.

Because the normal line depends on the geometric notion of angle (to tell you what's perpendicular to what), this can't be done as slickly using only differentials. Now we really do want to think of the gradient vector. All the same, since this can be read off of the differential so easily, you can still start with $\mathrm{d}u = y\,\mathrm{d}x + x\,\mathrm{d}y$. First, replace only x with 2 and y with 3 to get $3\,\mathrm{d}x + 2\,\mathrm{d}y$, then read off the gradient vector $\langle 3, 2\rangle$. Since we started at the point $(2, 3)$, the parametric equation is $P = (2, 3) + t\langle 3, 2\rangle$, or $(x, y) = (3t + 2, 2t + 3)$ in more detail.

None of this (beyond the level curve (or surface) itself) works right if the gradient $\nabla f(P_0)$ is zero or undefined. If the gradient is undefined, then of course we can't say anything using it; but if the gradient is zero, then these equations say that every point belongs to the tangent line (or plane) and only the point P_0 belongs to the normal line. Of course, that would mean that they're not lines (or a plane and a line) at all! When the gradient is zero, the truth may be that there is no tangent or that there is a tangent but it really does consist of everything, or there may be an honest tangent line (or plane) after all; but in any case, these formulas won't help you know that!

Taylor's Theorem in several variables

One version of Taylor's Theorem in one-variable Calculus is

$$f(a+h) = \sum_{n=0}^{k} \frac{1}{n!} f^{(n)}(a) h^n + \frac{1}{k!} \int_{t=0}^{1} (1-t)^k f^{(k+1)}(a+th) h^{k+1} \, \mathrm{d}t.$$

To be more explicit, here is the statement for the first few values of k:

$$\begin{aligned} f(a+h) &= f(a) + \int_{t=0}^{1} f'(a+th) h \, \mathrm{d}t \\ &= f(a) + f'(a)h + \int_{t=0}^{1} (1-t) f''(a+th) h^2 \, \mathrm{d}t \\ &= f(a) + f'(a)h + \frac{1}{2} f''(a) h^2 + \frac{1}{2} \int_{t=0}^{1} (1-t)^2 f'''(a+th) h^3 \, \mathrm{d}t \\ &\vdots \end{aligned}$$

Here, a and h are real numbers, k is a whole number, and f is a function that is continuously differentiable $k+1$ times (at least) between a and $a+h$. These statements may be proved by repeated application of integration by parts (and the Fundamental Theorem of Calculus, which is why $f^{(k+1)}$ must not only exist but also be continuous).

To write down the general statement in several variables requires more advanced notation than we use in this class, but I will write down the first few statements when f is a function of 2 variables:

$$\begin{aligned} f(a+h,b+i) &= f(a,b) + \int_{t=0}^{1} D_1 f(a+th,b+ti) h \, \mathrm{d}t + \int_{t=0}^{1} D_2 f(a+th,b+ti) i \, \mathrm{d}t \\ &= f(a,b) + D_1 f(a,b) h + D_2 f(a,b) i \\ &\quad + \int_{t=0}^{1} (1-t) D_{1,1} f(a+th,b+ti) h^2 \, \mathrm{d}t + \int_{t=0}^{1} (1-t) D_{1,2} f(a+th,b+ti) hi \, \mathrm{d}t \\ &\quad + \int_{t=0}^{1} (1-t) D_{2,1} f(a+th,b+ti) ih \, \mathrm{d}t + \int_{t=0}^{1} (1-t) D_{2,2} f(a+th,b+ti) i^2 \, \mathrm{d}t \\ &= f(a,b) + D_1 f(a,b) h + D_2 f(a,b) i \\ &\quad + \frac{1}{2} D_{1,1} f(a,b) h^2 + \frac{1}{2} D_{1,2} f(a,b) hi + \frac{1}{2} D_{2,1} f(a,b) ih + \frac{1}{2} D_{2,2} f(a,b) i^2 \\ &\quad + \frac{1}{2} \int_{t=0}^{1} (1-t)^2 D_{1,1,1} f(a+th,b+ti) h^3 \, \mathrm{d}t + \frac{1}{2} \int_{t=0}^{1} (1-t)^2 D_{1,1,2} f(a+th,b+ti) h^2 i \, \mathrm{d}t \\ &\quad + \frac{1}{2} \int_{t=0}^{1} (1-t)^2 D_{1,2,1} f(a+th,b+ti) hih \, \mathrm{d}t + \frac{1}{2} \int_{t=0}^{1} (1-t)^2 D_{1,2,2} f(a+th,b+ti) hi^2 \, \mathrm{d}t \\ &\quad + \frac{1}{2} \int_{t=0}^{1} (1-t)^2 D_{2,1,1} f(a+th,b+ti) ih^2 \, \mathrm{d}t + \frac{1}{2} \int_{t=0}^{1} (1-t)^2 D_{2,1,2} f(a+th,b+ti) ihi \, \mathrm{d}t \\ &\quad + \frac{1}{2} \int_{t=0}^{1} (1-t)^2 D_{2,2,1} f(a+th,b+ti) i^2 h \, \mathrm{d}t + \frac{1}{2} \int_{t=0}^{1} (1-t)^2 D_{2,2,2} f(a+th,b+ti) i^3 \, \mathrm{d}t \\ &\vdots \end{aligned}$$

These may again be proved by using integration by parts. In fact, by doing the integration by parts in slightly different ways, we can rearrange the order of the mixed partial derivatives (such as $D_{1,2}f$ and

$D_{2,1}f$); this both proves the theorem that the mixed partial derivatives are the same in either order (when they are continuous) but also allows us to simplify the formulas slightly:

$$\begin{aligned}
f(a+h,b+i) &= f(a,b) + \int_{t=0}^{1} D_1 f(a+th,b+ti)h\,\mathrm{d}t + \int_{t=0}^{1} D_2 f(a+th,b+ti)i\,\mathrm{d}t \\
&= f(a,b) + D_1 f(a,b)h + D_2 f(a,b)i + \int_{t=0}^{1} (1-t) D_{1,1} f(a+th,b+ti)h^2\,\mathrm{d}t \\
&\quad + 2\int_{t=0}^{1} (1-t) D_{1,2} f(a+th,b+ti)hi\,\mathrm{d}t + \int_{t=0}^{1} (1-t) D_{2,2} f(a+th,b+ti)i^2\,\mathrm{d}t \\
&= f(a,b) + D_1 f(a,b)h + D_2 f(a,b)i + \frac{1}{2} D_{1,1} f(a,b)h^2 + D_{1,2} f(a,b)hi + \frac{1}{2} D_{2,2} f(a,b)i^2 \\
&\quad + \frac{1}{2}\int_{t=0}^{1} (1-t)^2 D_{1,1,1} f(a+th,b+ti)h^3\,\mathrm{d}t + \frac{3}{2}\int_{t=0}^{1} (1-t)^2 D_{1,1,2} f(a+th,b+ti)h^2 i\,\mathrm{d}t \\
&\quad + \frac{3}{2}\int_{t=0}^{1} (1-t)^2 D_{1,2,2} f(a+th,b+ti)hi^2\,\mathrm{d}t + \frac{1}{2}\int_{t=0}^{1} (1-t)^2 D_{2,2,2} f(a+th,b+ti)i^3\,\mathrm{d}t \\
&\quad \vdots
\end{aligned}$$

However, in my opinion, the pattern is not so clear when it's put this way.

For purposes of approximation, it's useless to actually work out the integrals that appear here; if you knew the exact value of the derivatives of f at all the points between (a,b) and $(a+h,b+i)$, then you could probably just evaluate f at $(a+h,b+i)$ directly. However, if there is a value M such that you know that none of the derivatives of f of order $k+1$ have an absolute value greater than M at any point between (a,b) and $(a+h,b+i)$, then you can leave off the integrals to get an approximation of $f(a+h,b+i)$ and then use M to get an estimate of the error of this approximation:

$f(a+h,b+i) \approx f(a,b)$, a constant approximation, if f is continuous;

$f(a+h,b+i) \approx f(a,b) + D_1 f(a,b)h + D_2 f(a,b)i$, a linear approximation, if f is differentiable;

$f(a+h,b+i) \approx f(a,b) + D_1 f(a,b)h + D_2 f(a,b)i + \frac{1}{2} D_{1,1} f(a,b)h^2 + D_{1,2} f(a,b)hi + \frac{1}{2} D_{2,2} f(a,b)i^2$,
a quadratic approximation, if f is twice differentiable;

$\vdots$

with

$$|f(a+h,b+i) - f(a,b)| \le M_1(|h| + |i|)$$

if $|D_1 f|$ and $|D_2 f|$ are never greater than M_1 between (a,b) and $(a+h,b+i)$,

$$\left| f(a+h,b+i) - \Big(f(a,b) + D_1 f(a,b)h + D_2 f(a,b)i \Big) \right| \le \frac{1}{2} M_2 (|h| + |i|)^2$$

if $|D_{1,1} f|$, $|D_{1,2} f|$, and $|D_{2,2} f|$ are never greater than M_2 between (a,b) and $(a+h,b+i)$,

$$\left| f(a+h,b+i) - \left(f(a,b) + D_1 f(a,b)h + D_2 f(a,b)i + \frac{1}{2} D_{1,1} f(a,b)h^2 + D_{1,2} f(a,b)hi + \frac{1}{2} D_{2,2} f(a,b)i^2 \right) \right| \le \frac{1}{6} M_3 (|h| + |i|)^3$$

if $|D_{1,1,1} f|$, $|D_{1,1,2} f|$, $|D_{1,2,2} f|$, and $|D_{2,2,2} f|$ are never greater than M_3 between (a,b) and $(a+h,b+i)$, etc.

Using vectors, we can write the first approximation and its error in any number of variables:

$$f(P_0 + \mathbf{v}) \approx f(P_0),$$
$$|f(P_0 + \mathbf{v}) - f(P_0)| \leq M_1 |\mathbf{v}|_1,$$

where $|\mathbf{v}|_1$ is the so-called 1-norm of $\mathbf{v}$, found by adding up the absolute values of its components. (The usual magnitude is then called the 2-norm, because these absolute values are raised to the power of 2 before they are added and then the principal root of index 2 is extracted; in general, you can consider the p-norm $|\mathbf{v}|_p$ for any positive real number p, or even other values of p if you're sufficiently clever.) We can also write the second approximation and its error using vectors:

$$f(P_0 + \mathbf{v}) \approx f(P_0) + \nabla f(P_0) \cdot \mathbf{v},$$
$$\left| f(P_0 + \mathbf{v}) - \left(f(P_0) + \nabla f(P_0) \cdot \mathbf{v} \right) \right| \leq \frac{1}{2} M_2 |\mathbf{v}|_1^2.$$

The next approximation, however, requires dyads to write down, which are more complicated than vectors; to write down the general case to any order involves a massive generalization of vectors called tensors. However, you can always write it down in any specific dimension by writing a lot of terms according to the appropriate pattern, as I did on the first page; there is also a technique, called multi-index notation, to encode these patterns, which you can see (for example) on the English Wikipedia article on Taylor's Theorem (as of today).

It's handy to describe these approximations in terms of differentials and differences. While a differential represents an infinitesimal (infinitely small) change, a **difference** represents an appreciable or finitesimal (not infinitely small) change. As $P = (x, y)$ (or (x, y, z) etc) changes from P_0 to $P_0 + \mathbf{v}$, we say that the difference in P is

$$\Delta P = (P_0 + \mathbf{v}) - P_0 = \mathbf{v}.$$

Meanwhile, if $u = f(P)$, then the difference in u is

$$\Delta u|_{\substack{P=P_0, \\ \Delta P = \mathbf{v}}} = f(P_0 + \mathbf{v}) - f(P_0).$$

Then the constant approximation says

$$\Delta u|_{\substack{P=P_0, \\ \Delta P = \mathbf{v}}} \approx 0,$$

while the linear approximation says (more precisely)

$$\Delta u|_{\substack{P=P_0, \\ \Delta P = \mathbf{v}}} \approx \mathrm{d}u|_{\substack{P=P_0, \\ \mathrm{d}P = \mathbf{v}}}.$$

So in the end, the linear approximation replaces differences with differentials. The next (quadratic) approximation can be written using the second differential $\mathrm{d}^2 u$, and so on, but we won't cover that in this class. The error estimates are

$$\left| \Delta u|_{\substack{P=P_0, \\ \Delta P = \mathbf{v}}} \right| \leq M_1 |\mathbf{v}|_1$$

and

$$\left| \Delta u|_{\substack{P=P_0, \\ \Delta P = \mathbf{v}}} - \mathrm{d}u|_{\substack{P=P_0, \\ \mathrm{d}P = \mathbf{v}}} \right| \leq \frac{1}{2} M_2 |\mathbf{v}|_1^2.$$

Optimization

Literally, *optimization* is making something the best, but we use it in math to mean *maximization*, which is making something the biggest. (You can imagine that the thing that you're maximizing is a numerical measure of how good the thing that you're optimizing is.) Essentially the same principles apply to *minimization*, which is making something the smallest. (And *pessimization* is making something the worst, although people don't use that term very much, because who would want to do that?) A generic term for making something the largest or smallest is *extremization*.

The key principle of optimization is this:

> A quantity u can only take a maximum (or minimum) value when its differential $\mathrm{d}u$ is zero or undefined.

If you write u as $f(x, y)$, where f is a fixed differentiable function of (say) 2 variables, and x and y are quantities whose range of possible values you already understand (typically intervals), then $\mathrm{d}u = D_1 f(x, y)\,\mathrm{d}x + D_2 f(x, y)\,\mathrm{d}y$, or equivalently, $\mathrm{d}u = \frac{\partial u}{\partial x}\,\mathrm{d}x + \frac{\partial u}{\partial y}\,\mathrm{d}y$.

So one way that u might conceivably take an extreme value is if either (or both) of its partial derivatives are undefined. Another way is if both (not just one) of its partial derivatives are zero. If you can vary x and y smoothly however you please (essentially, if you are in the interior of the domain of f and you are free to access the entire domain), then these are the only possibilities. However, if you cannot vary them smoothly (essentially, if you are on the boundary of the domain of f or if the sitaution is otherwise constrained so that you cannot access the entire domain of f), then there are more possibilities!

If your constraint (or constraints) can be written as an equation $g(x, y) = 0$ (or really, with any constant on the right-hand side), then as long as the gradient ∇g is never zero on the solution set of the constraint equations, then you can use the method of *Lagrange multipliers*. Here, you set up an equation $\nabla f(x, y) = \lambda \nabla g(x, y)$, combine this with the equation $g(x, y) = 0$, and try to solve for x, y, and λ. (Since a vector equation is equivalent to 2 scalar equations, this amounts to a system of 3 equations in 3 variables, so there is hope to solve for it.) If you're working in 3 variables, then you might need two equations to specify the constraint, in which case there are two functions in the place of g and two Lagrange multipliers. (But you can also have just one g even in 3 dimensions; it's a question of whether the boundary in question is a surface or a curve.) While λ ultimately doesn't matter, the solutions that you get for the original variables give you additional critical points to check for extreme values.

On the other hand, you don't actually need Lagrange multipliers! Writing v for $g(x, y)$, if the constraint is $v = 0$ (or any constant), then differentiate this to get $\mathrm{d}v = 0$. (In fact, you could take any equation and just differentiate both sides.) Then if you try to solve the system of equations consisting of $\mathrm{d}u = 0$ and $\mathrm{d}v = 0$ for the differentials $\mathrm{d}x$ and $\mathrm{d}y$, you should immediately see that $\mathrm{d}x = 0$ and $\mathrm{d}y = 0$ is a solution. However, if you actually go through the steps of solving this as a system of linear equations (which you can always do because differentials are always linear in the differentials of the independent variables), you'll find that at some point you need to divide by some quantity involving x and y, which is invalid if that quantity is zero! So, setting whatever you divide by to zero and combining that with the constraint equation $v = 0$, you get two equations to solve for the two variables x and y. (With this method, λ never enters into it.) This will give you the other critical points to check for extreme values.

Be careful, because u might not have a maximum or minimum value! Assuming that u varies continuously (which it must if Calclulus is to be useful at all), then it must have a maximum and minimum value whenever the domain of the function (including any constraints) is both closed and bounded (which is called *compact*); this means that if you pass continuously through the possibilities in any way, then you are always approaching some limiting possibility. However, if the range of possibilities heads off to infinity in some way, then you also have to take a limit to see what value u is approaching, which can be very difficult to do in more than one dimension. Or if there is a boundary that's not included in the domain, then you have to take a limit approaching that boundary, although in that case you can hope that you can check the boundary as if it were included, the same way as above. If any such limit is larger than every value that u actually reaches (which includes the possibility that a limit is ∞), then u has no maximum value; if any such limit is smaller than every value that u actually reaches (which includes the possibility that a limit is $-\infty$), then u has no minimum value.

So in the end, you look at these possibilities to optimize u:

- when any partial derivative of u is undefined,
- when all partial derivatives of u are zero,
- any boundary possibilities given by a constraint,
- any corners (boundaries of the boundaries) given by two constraints,
- any corners of corners given by three constraints (not possible with only two independent variables),
- etc (in more than 3 dimensions), and
- the limits approaching impossible limiting cases.

Whichever of these has the largest value of u gives you the maximum, and whichever has the smallest value of u gives you the minimum; but if the largest or smallest value is only approached in the limit, then the maximum or minimum technically does not exist.

Here is a typical problem: The hypotenuse of a right triangle (maybe it's a ladder leaning against a wall) is fixed at 20 feet, but the other two sides of the triangle could be anything. Still, since it's a right triangle, we know that $l^2 + h^2 = 20^2$, where l and h (length and height) are the lengths of legs of the triangle. (If we think of l and h as independent variables, then this equation is our constraint.) Differentiating this, $2l\,\mathrm{d}l + 2h\,\mathrm{d}h = 0$. Now suppose that we want to maximize or minimize the area of this triangle. Since it's a right triangle, the area is $A = \frac{1}{2}lh$, so $\mathrm{d}A = \frac{1}{2}h\,\mathrm{d}l + \frac{1}{2}l\,\mathrm{d}h$. If this is zero, then $\frac{1}{2}h\,\mathrm{d}l + \frac{1}{2}l\,\mathrm{d}h = 0$, to go along with the other equation $2l\,\mathrm{d}l + 2h\,\mathrm{d}h = 0$.

The equations at this point are linear in the differentials (as they always must be), so think of this is a system of linear equations in the variables $\mathrm{d}l$ and $\mathrm{d}h$. There are various methods for solving systems of linear equations; I'll use the method of addition aka elimination, but any other method should work just as well. So $\frac{1}{2}h\,\mathrm{d}l + \frac{1}{2}l\,\mathrm{d}h = 0$ becomes $2lh\,\mathrm{d}l + 2l^2\,\mathrm{d}h = 0$ (multiplying both sides by $4l$), while $2l\,\mathrm{d}l + 2h\,\mathrm{d}h = 0$ becomes $2lh\,\mathrm{d}l + 2h^2\,\mathrm{d}h = 0$ (multiplying both sides by h). Subtracting these equations gives $(2l^2 - 2h^2)\,\mathrm{d}h = 0$, so either $\mathrm{d}h = 0$ or $l^2 = h^2$. Now, l and h can change freely as long as they're positive, but we have limiting cases: $l \to 0^+$ and $h \to 0^+$. Since $l^2 + h^2 = 400$, we see that $l^2 \to 400$, so $l \to 20$, as $h \to 0$. Similarly, $h \to 20$ as $l \to 0$. In those cases, $A = \frac{1}{2}lh \to 0$. On the other hand, if $l^2 = h^2$, then $l = h$, so $l, h = 10\sqrt{2}$, since $l^2 + h^2 = 400$. In that case, $A = \frac{1}{2}lh = 100$.

So the largest area is 100 square feet, and while there is no smallest area, the area can get arbitrarily small with a limit of 0.

Integration on curves

Differential 1-forms (that is differential forms without the wedge product that we will get to later) can be integrated along curves. To a large extent, that is what they are for. Since differential forms are made of differentials and the definition of the differential of an expression (at least the one that I gave in the handout from January 18) is ultimately about curves, this is a very natural operation.

The definition

Like the textbook does for one-variable Calculus, I'll define the Riemann integral as a limit of Riemann sums, although there are more general notions of integration that can handle more expressions. The Riemann integral will be sufficient for *piecewise continuous* differential forms (those defined in one or more pieces using continuous operations applied to continuous quantities and the differentials of continuously differentiable quantities) along *piecewise continuously differentiable curves* (those with parametrizations defined in one more pieces using continuously differentiable operations applied to the parameter).

So, suppose that we have a differential form α written using the variables $P = (x, y, \ldots)$ and their differentials, and a curve in the same number of dimensions, given by some parametrization function C whose domain is a closed interval $[a, b]$. Then we can try to integrate α along the curve where $P = C(t)$, by defining the integral

$$\int_{P=C(t)} \alpha.$$

Given any way of dividing the interval $[a, b]$ into a partition $a = t_0 \leq t_1 \leq \cdots \leq t_{n-1} \leq t_n = b$ (with n subintervals) and tagging this partition with n values c_k with $t_{k-1} \leq c_k \leq t_k$ for k from 1 to n (this is exactly the kind of partition considered in one-variable Calculus, as on pages 297–299 of the textbook), there is a **Riemann sum**

$$\sum_{k=1}^{n} \alpha\big|_{\substack{P=C(c_k),\\ \mathrm{d}P=C(t_k)-C(t_{k-1})}}.$$

That is, on the kth subinterval, we evaluate the form α at the point through which the curve passes at time c_k within that subinterval along the vector from where the curve is at the beginning of the subinterval to where it is at the end of the subinterval. If we require that the magnitude of this vector be less than δ and take the limit as $\delta \to 0^+$, then this limit (if it exists) is the value of the integral. And there is a theorem that it does exist, at least if α is piecewise continuous and C is piecewise continuously differentiable (and sometimes otherwise); I don't know a nice proof of this directly, but you can prove that it exists because the practical calculation method on page 2 works.

There is now another nice theorem, that the value of this integral does not depend on the parametrization of the curve, at least not very much. That is, if ϕ is a function in the ordinary sense (a real-valued function of one real variable), then $C \circ \phi$ is another parametrized curve; if ϕ is one-to-one and increasing (so that we travel along the curve in the same direction without repetition) and its range lies entirely within the domain of C (so that we cover the entire curve), then the theorem is that $\int_{P=C(t)} \alpha = \int_{P=(C\circ\phi)(t)} \alpha$. The proof is that any Riemann sum for C is also a Riemann sum for $C \circ \phi$; the same points $C(t_k)$ and $C(c_k)$ occur in the same order, just at different values of the parameter. So the Riemann integrals, which are the limits of these Riemann sums, must also be the same.

For this reason, we usually don't specify a parametrized curve in the notation at all. Instead, we specify an **oriented curve**, which is anything that *could* be given as a parametrized curve, keeping track of which direction we travel along the curve (this is the **orientation** of the curve) but otherwise ignoring the parametrization.

Evaluating integrals along curves

The practical method of evaluating integrals along curves is to pick any convenient parametrization (preferably one that is continuously differentiable) and put everything in terms of that parameter. For example, to integrate $2x\,\mathrm{d}x + 3xy\,\mathrm{d}y$ along the top half of the circle $x^2 + y^2 = 4$, oriented counterclockwise, try the parametrization where $x = 2\cos t$, $y = 2\sin t$, and $0 \le t \le \pi$. Then $\mathrm{d}x = -2\sin t\,\mathrm{d}t$ and $\mathrm{d}y = 2\cos t\,\mathrm{d}t$, so the value of the integral is

$$\begin{aligned}\int_{\substack{x^2+y^2=4,\ y\ge 0\\ \mathrm{d}x\le 0}} (2x\,\mathrm{d}x + 3xy\,\mathrm{d}y) &= \int_{t=0}^{\pi} \Big(2(2\cos t)(-2\sin t\,\mathrm{d}t) + 3(2\cos t)(2\sin t)(2\cos t\,\mathrm{d}t)\Big) \\ &= \int_{t=0}^{\pi} (-8\sin t\cos t + 24\sin t\cos^2 t)\,\mathrm{d}t = 16.\end{aligned}$$

(You can do this last integral with the substitution $u = \cos t$.) I've described the curve of integration with an equation (of a circle) and an inequality (to get the top half only) and oriented it by saying that x is always decreasing (so that $\mathrm{d}x$ is always negative), but usually people write that all out to the side somewhere, call the resulting oriented curve C (for example), and write $\int_C (2x\,\mathrm{d}x + 3xy\,\mathrm{d}y)$.

The reason why this gives the correct result is that any Riemann sum for the integral involving t involves almost the same calculations as a Riemann sum for the integral along the curve. The only difference is that the integral involving t looks at the point from within of each subinterval to handle the differentials, whereas as the integral of the curve looks at the points on each end of the subinterval. But in the limit, all of these points approach each other, and the result is the same. (There is another slight complication because the integral involving t takes a limit as the change in t goes to 0, while the integral along the curve takes a limit as the magnitude of the change in position goes to 0. However, these are the same because the parametrization is continuous. If you can calculate $\mathrm{d}x$ and $\mathrm{d}y$ at all, then the parametrization must be differentiable and so definitely continuous.)

You should be able to visualize this example geometrically well enough to see that the answer would have to be positive. The term $2x\,\mathrm{d}x$ should completely cancel, because the right half of the curve exactly mirrors the left half, with $\mathrm{d}x$ the same on both halves (always negative because of movement to the left) but x being the opposite on the two halves (first positive, then negative). On the other hand, the term $3xy\,\mathrm{d}y$ will be negative on both sides; while y is always positive (above the horizontal axis), x and $\mathrm{d}y$ are both positive on the right half (right of the vertical axis and moving upwards) and both negative on the left half (left of the axis and moving downwards), making for a positive product everywhere.

If you are asked to integrate a vector field $\mathbf{F}$ along an oriented curve, then they really want you to integrate the differential form $\mathbf{F}(x,y) \cdot \langle \mathrm{d}x, \mathrm{d}y\rangle$, or more generally $\mathbf{F}(P) \cdot \mathrm{d}P$, where P is (x,y) or (x,y,z). If you write $\mathbf{r}$ for the vector $P - \mathrm{O}$ (where O is the origin $(0,0)$ or $(0,0,0)$), then $\mathrm{d}P = \mathrm{d}\mathbf{r}$, and this is the reason for the traditional notation $\int_C \mathbf{F}\cdot\mathrm{d}\mathbf{r}$, which is used in the textbook. (You may also see $\int_C \mathbf{F}\cdot\mathbf{T}\,\mathrm{d}s$, where $\mathrm{d}s$ is the $\mathrm{d}s$ that appears at the very bottom of this page and $\mathbf{T}$ is defined to be $\mathrm{d}\mathbf{r}/\mathrm{d}s$. This is usually completely pointless; if you see $\mathbf{T}\,\mathrm{d}s$, just think of it as $\mathrm{d}\mathbf{r}$.)

For example, to integrate $\langle 2x, 3xy\rangle$ along the same semicircle as in the previous example (with the same orientation), you do exactly the same integral as in the previous example. This is because

$$\langle 2x, 3xy\rangle \cdot \langle \mathrm{d}x, \mathrm{d}y\rangle = 2x\,\mathrm{d}x + 3xy\,\mathrm{d}y,$$

so

$$\int_C \langle 2x, 3xy\rangle \cdot \mathrm{d}\mathbf{r} = \int_C (2x\,\mathrm{d}x + 3xy\,\mathrm{d}y) = 16$$

as before. Since the vector $\langle 2x, 3xy\rangle$ points to the right on the right side and to the left on the left side, while we move along the curve consistently to the left, this suggests that the horizontal component should cancel. However, since this vector points upwards where we move upwards along the curve (on the right side) and points downwards where we move downards along the curve (on the left side), this suggests a positive contribution from the vertical component. So as in the first example, you should expect a positive result even before doing the calculation.

If you are asked to integrate a function f along a curve, then they really want you to integrate the differential form $f(x,y)\sqrt{\mathrm{d}x^2 + \mathrm{d}y^2}$, or more generally $f(P)\,|\mathrm{d}P|$. It's traditional to write $\mathrm{d}s$ for $|\mathrm{d}P|$ (or

$|\mathrm{d}\mathbf{r}|$, which is the same), but it's important that there is no quantity s defined everywhere on the coordinate plane that $\mathrm{d}s$ is the differential of. To emphasize this, you can write $đs$; 'đ' is a symbol that some people use when something is traditionally written with 'd' but is not really a differential.

As long as the differentials $\mathrm{d}x$ etc appear only in $đs$, then the result of the integral is independent of orientation, because replacing $\mathrm{d}x$ with $-\mathrm{d}x$ (as would happen upon reversing the orientation) doesn't change $đs$. For this reason, you can integrate a function on an *unoriented* curve. When parametrizing, everything will come out using $|\mathrm{d}t|$ instead of $\mathrm{d}t$, but as long as the integral involving t has its bounds set up so that t is increasing, then $\mathrm{d}t$ is positive and so $|\mathrm{d}t| = \mathrm{d}t$, after which you can integrate normally.

For example, to integrate $f(x, y) = 6x^2y$ on the same semicircle as in the previous examples, you get

$$đs = \sqrt{\mathrm{d}x^2 + \mathrm{d}y^2} = \sqrt{(-2\sin t\,\mathrm{d}t)^2 + (2\cos t\,\mathrm{d}t)^2} = \sqrt{(4\sin^2 t + 4\cos^2 t)\,\mathrm{d}t^2} = \sqrt{4}\sqrt{\mathrm{d}t^2} = 2\,|\mathrm{d}t|.$$

Thus, the integral is

$$\int_{x^2+y^2=4,\ y\geq 0} 6x^2y\,đs = \int_{t=0}^{\pi} 6(\cos t)^2(\sin t)(2\,|\mathrm{d}t|) = \int_{t=0}^{\pi} 12\sin t\cos^2 t\,\mathrm{d}t = 8.$$

Since x^2y is positive everywhere on this curve, you should have expected a positive result.

If for some reason you set the integral up backward, then $\mathrm{d}t$ would be negative and so $|\mathrm{d}t|$ would be $-\mathrm{d}t$, and the result would be the same in the end:

$$\int_C đs = \int_{t=\pi}^{0} 12\sin t\cos^2 t\,|\mathrm{d}t| = \int_{t=\pi}^{0} 12\sin t\cos^2 t(-\mathrm{d}t) = -\int_{t=\pi}^{0} 12\sin t\cos^2 t\,\mathrm{d}t = -(-8) = 8.$$

(But it's simpler to always set things up so that the parameter is increasing.)

Pseudooriented curves

In 2 dimensions, you'll sometimes be asked to integrate a vector field *across* a curve rather than *along* it as usual. Although there is no standard notation for this, you can write it as as $\mathbf{F} \times \mathrm{d}\mathbf{r}$ in analogy with the usual $\mathbf{F} \cdot \mathrm{d}\mathbf{r}$. The book sometimes writes $\mathbf{F} \cdot \mathbf{n}\,\mathrm{d}s$, where $\mathbf{n} = \times\mathbf{T}$ and $\mathrm{d}\mathbf{r} = \mathbf{T}\,\mathrm{d}s$, but this just results in $\mathbf{F} \cdot \times\mathrm{d}\mathbf{r} = \mathbf{F} \times \mathrm{d}\mathbf{r}$.

This is the 2-dimensional cross product, so the result is still a scalar. Technically, however, it is actually a **pseudoscalar**, because its sign depends on how you orient the plane (counterclockwise as is the convention, or clockwise instead). Similarly, specifying a direction across a curve really gives the curve a **pseudoorientation**, because it only defines a direction along the curve (an orientation) by picking a convention about how these directions correspond. In practice, we orient the plane counterclockwise, meaning that counterclockwise cross products are positive, the rotation $\times\mathbf{v}$ of a vector $\mathbf{v}$ is obtained by rotating it clockwise, a direction across a curve turns into a direction along it by rotation counterclockwise, and a direction along a curve turns into a direction across it by rotating clockwise. But if you consistently did all of these the other way, then the results of all integrals would be the same.

For example, to integrate $\langle 2x, 3xy\rangle$ across our semicircle, now pseudooriented upwards, integrate

$$\langle 2x, 3xy\rangle \times \langle \mathrm{d}x, \mathrm{d}y\rangle = 2x\,\mathrm{d}y - 3xy\,\mathrm{d}x,$$

and use the orientation counterclockwise from upwards, which is leftwards (the same as in first example):

$$\begin{aligned}\int_{\substack{x^2+y^2=4,\ y\geq 0\\ \mathrm{d}y\geq 0}} \langle 2x, 3xy\rangle \times \mathrm{d}\mathbf{r} &= \int_{\substack{x^2+y^2=4,\ y\geq 0\\ \mathrm{d}x\leq 0}} (2x\,\mathrm{d}y - 3xy\,\mathrm{d}x)\\ &= \int_{t=0}^{\pi} \Big((2(2\cos t)(2\cos t\,\mathrm{d}t)) - 3(2\cos t)(2\sin t)(-2\sin t\,\mathrm{d}t)\Big)\\ &= \int_{t=0}^{\pi} (8\cos^2 t + 24\sin^2 t\cos t)\,\mathrm{d}t = 4\pi.\end{aligned}$$

Since the vector $\langle 2x, 3xy\rangle$ points to the right where we cross the curve to the right (on the right side) and points to the left where we cross to the left, this suggests that the horizontal component should give a positive result. However, since this vector points upwards on the right side and downwards on the left side, while we cross the curve consistently upwards, this suggests that the vertical component should cancel. So you should again expect a positive result before doing the calculation.

The Fundamental Theorem of Calculus

In one-variable Calculus, the second Fundamental Theorem states that

$$\int_{x=a}^{b} f'(x)\,\mathrm{d}x = f(b) - f(a).$$

If we write u for the quantity $f(x)$, then its differential $\mathrm{d}u$ is precisely the integrand $f'(x)\,\mathrm{d}x$, so the Fundamental Theorem can also be written as

$$\int_{a}^{b} \mathrm{d}u = u\big|_a^b.$$

This works just as well when there are several independent variables as when there is just one. Now if $u = f(P)$, then $\mathrm{d}u$ is $\nabla f(P) \cdot \mathrm{d}\mathbf{r}$, so

$$\int_{P=a}^{b} \nabla f(P) \cdot \mathrm{d}\mathbf{r} = f(b) - f(a).$$

Although this is now a theorem about integrating a gradient along a curve, in essence it is still just the FTC, a theorem about integrating differentials. This has a massive generalization to higher-rank differential forms, called the *Stokes Theorem*, which we'll get to later.

A differential form is called **exact** if there exists a quantity u such that $\alpha = \mathrm{d}u$. Similarly, a vector field $\mathbf{F}$ is called **conservative** if there is a scalar field f such that $\mathbf{F} = \nabla f$. The connection between these is that $\mathbf{F}$ is conservative if and only if $\mathbf{F}(P) \cdot \mathrm{d}\mathbf{r}$ is exact. (After all, if $\mathbf{F} = \nabla f$, then $\mathbf{F}(P) \cdot \mathrm{d}\mathbf{r} = \mathrm{d}\big(f(P)\big)$.) An oriented curve is called **closed** if its beginning and ending points are the same; one sometimes emphasizes that an integral is along a closed curve by writing $\oint$ in place of $\int$. Then the integral of an exact differential form or a conservative vector field along a closed curve is zero, because

$$\oint_C \alpha = \int_a^a \mathrm{d}u = u\big|_a^a = u\big|_a - u\big|_a = 0.$$

Similarly, the integral of a conservative vector field along a closed curve is zero. In this case, we can use notation more like that of a definite integral in one variable:

$$\int_{P=P_1}^{P_2} \alpha$$

means the integral of α along *any* curve from P_1 to P_2. It doesn't matter which curve you use; if C_1 and C_2 are both curves like this, then these combine into a closed curve $C_1 - C_2$, in which you start at P_1, follow C_1 to P_2, then follow C_2 backwards (hence the minus sign) back to P_1. Then

$$\int_{C_1} \alpha - \int_{C_2} \alpha = \oint_{C_1 - C_2} \alpha = 0,$$

so $\int_{C_1} \alpha = \int_{C_2} \alpha$. (This is still undefined if there is *no* curve from P_1 to P_2 through the domain of α. This is analogous to the case in one dimension of an integral $\int_{x=a}^{b} f(x)\,\mathrm{d}x$ where f is undefined somewhere between a and b.)

Conversely, if the integral of a differential form or of a vector field is zero along *every* closed curve, then that differential form must be exact or that vector field must be conservative. The reason is that in this case (and only in this case) we can pick a point P_0 to start from and define a semidefinite integral

$$u = \int_{P=P_0} \alpha = \int_{P_0}^{P} \alpha.$$

Because α is exact, you get the same result no matter which path you use from P_0 to P. (Ideally, the domain of α should be *path-connected*, meaning that there exists a curve between any two points. If not, then you must split the domain into various path-connected components and pick a point in each.) That $\mathrm{d}u = \alpha$ in this case is essentially the multivariable version of the *first* Fundamental Theorem of Calculus.

Given a differential form α, finding such an expression u is a form of *indefinite* integration. It's not practical to check every possible curve, of course, so we need other methods to decide if α is exact, and

this can also help us to find u. There are actually several methods; one is given in the textbook, essentially reversing the process of partial differentiation with a kind of partial integration. (If you try this method when α is not exact, then it will fail.)

If the domain of α is reasonably simple, then it's possible to pick a point P_0 and write down a general formula for a parametrized curve from P_0 to any point P. (For example, you could always use a straight line segment, as long as these line segments always lie entirely within the domain.) If you try this method when α is not exact, then you may get a result; but when you check it, then you'll find that it's wrong when α is not exact.

It's often possible to tell ahead of time whether α is exact. To really explain what's going on here, I'll need to talk about the *exterior differential*, which is a topic that we'll get to in a couple of weeks. For now, I'll describe it in terms of partial derivatives. So, if $\alpha = \mathrm{d}u$, then

$$\alpha = \frac{\partial u}{\partial x}\,\mathrm{d}x + \frac{\partial u}{\partial y}\,\mathrm{d}y + \cdots.$$

(The dots are meant to indicate that more terms may appear if there are more than two variables.) Assuming that u is twice differentiable, then mixed second partial derivatives are equal:

$$\frac{\partial^2 u}{\partial x\,\partial y} = \frac{\partial^2 u}{\partial y\,\partial x}.$$

So if you start with an arbitrary linear differential 1-form

$$\alpha = \alpha_x\,\mathrm{d}x + \alpha_y\,\mathrm{d}y + \cdots,$$

then it could only be exact if it is **closed**, meaning that

$$\frac{\partial \alpha_x}{\partial y} = \frac{\partial \alpha_y}{\partial x}$$

(and similarly for other mixtures of derivatives if there are more than two variables), assuming that it's differentiable in the first place. Similarly, a vector field

$$\mathbf{F}(x, y, \ldots) = \mathbf{F}_1(x, y, \ldots)\mathbf{i} + \mathbf{F}_2(x, y, \ldots)\mathbf{j} + \cdots$$

can only be conservative if it is **irrotational**, meaning that

$$D_2\mathbf{F}_1 = D_1\mathbf{F}_2$$

(and similarly for other mixtures of derivatives if there are more than two variables), assuming that it's differentiable in the first place.

Conversely, a closed differential form or an irrotational vector field must be exact or conservative (respectively) if its domain is **precisely-simply connected**, which means that any simple closed curve (one that doesn't intersect itself except where its two endpoints are equal) in the domain of the differential form or the vector field is the boundary of a region that lies entirely within that domain. (The domain is *simply connected* if it is both path-connected and precisely-simply connected. Conversely, it is precisely-simply connected if each of its path-connected components is simply connected. If you take a class in Topology such as MATH 471 at UNL, then you'll learn a hundred specific terms like these.) But a full discussion of the reasons for this must wait until we've covered higher-order differential forms.

Fubini theorems

I want to record here some theorems about double (and higher) Riemann integrals, culminating in using the Fubini theorems to turn them into iterated integrals.

First, I want to note some notation that is a little more precise than the notation in the textbook. The notation in the textbook is very common, and it's usually quite clear what it means, but it's not good enough if you want to be completely unambiguous about what variables you're using and where. So rather than write, for example,

$$\iint_D f(x,y)\,\mathrm{d}A,$$

where D is a region in 2 dimensions (formally a relation between 2 variables) and f is a function of 2 variables, I'll write

$$\int_{(x,y)\in D} f(x,y)\,|\mathrm{d}x\wedge\mathrm{d}y|.$$

So to begin with, since the integrand (all of the stuff after the integral symbols) makes it clear that there are 2 variables of integration, it's not necessary to repeat the integral symbol. But just as we write $f(x,y)$ (rather than just f) after that symbol to indicate the value of the function f at particular values of the variables x and y (rather than its value somewhere else), so I write $(x,y)\in D$ (rather than just D) beneath that symbol to indicate that the point whose coordinates are those values (rather than some other point) belongs to the region D.

At the end, since $\mathrm{d}A$ (and $\mathrm{d}V$ in 3 dimensions) don't indicate which variables are being used, I use the notation $|\mathrm{d}x\wedge\mathrm{d}y|$ (or $|\mathrm{d}x\wedge\mathrm{d}y\wedge\mathrm{d}z|$ in 3 dimensions). This notation is more complicated than necessary just to indicate the variables, but there is a reason for it; just as the notation $\mathrm{d}y/\mathrm{d}x$ for a derivative is not merely an arbitrary symbol but can be literally understood as the result of dividing expressions (called differentials) obtained by applying an operator d, so the notation $|\mathrm{d}x\wedge\mathrm{d}y|$ for an area element (or $|\mathrm{d}x\wedge\mathrm{d}y\wedge\mathrm{d}z|$ for a volume element) is not merely an arbitrary symbol but can be literally understood as the absolute value of an expression (called an exterior differential form) involving an operator $\wedge$. However, don't worry about that for now; just treat it as a notation used to indicate precisely which variables are used in the area (or volume) element.

Using that notation, here are the important theorems:

1 The integral of a continuous function on a compact (that is closed and bounded) region always exists: $\int_{(x,y)\in D} f(x,y)\,|\mathrm{d}x\wedge\mathrm{d}y|$ exists if f is continuous and D is compact (and similarly in more variables).

2 If two regions D_1 and D_2 are completely disjoint (no overlap at all), or if their overlap is contained within a single point/line/plane/etc of fewer dimensions than the overall number of variables, and if a function f has integrals on both of these regions, then the integral of f on their union (the combined region $D_1\cup D_2$) also exists and is the sum of the separate integrals:

$$\int_{(x,y)\in D_1\cup D_2} f(x,y)\,|\mathrm{d}x\wedge\mathrm{d}y| = \int_{(x,y)\in D_1} f(x,y)\,|\mathrm{d}x\wedge\mathrm{d}y| + \int_{(x,y)\in D_2} f(x,y)\,|\mathrm{d}x\wedge\mathrm{d}y|$$

(and similarly in more variables) if the integrals on the right exist and the overlap is small.

3 In any double (or higher) integral, if two of the variables are swapped in both the function being integrated and in the region over which it is integrated (or equivalently, by renaming the variables, by swapping the variables only with the area/volume/etc element), then the result is the same (so that if either integral exists, then so does the other, and then they are equal):

$$\int_{(x,y)\in D} f(x,y)\,|\mathrm{d}x\wedge\mathrm{d}y| = \int_{(x,y)\in D} f(x,y)\,|\mathrm{d}y\wedge\mathrm{d}x|$$

(and similarly in more variables).

4 For a region D in 2 dimensions, if there are constants a and b with $a \leq b$ and continuous functions g and h (each of 1 variable) such that $(x, y) \in D$ if and only if $a \leq x \leq b$ and $g(x) \leq y \leq h(x)$, and if $g(x) \leq h(x)$ whenever $a \leq x \leq b$, then the integral of any continuous function f on D is the same as the corresponding iterated integral:

$$\int_{(x,y)\in D} f(x,y)\,|\mathrm{d}x \wedge \mathrm{d}y| = \int_{x=a}^{b} \left(\int_{y=g(x)}^{h(x)} \mathrm{d}y \right) \mathrm{d}x.$$

5 For a region D in 3 (or more) variables, if there are a compact region R in 2 variables (or in general a compact region of one fewer dimension) and continuous functions g and h of 2 variables each (or in general with the same number of variables as R has dimensions) such that $(x, y, z) \in D$ if and only if $(x, y) \in R$ and $g(x, y) \leq z \leq h(x, y)$, and if $g(x, y) \leq h(x, y)$ whenever $(x, y) \in R$, then the integral of any continuous function f on D is the same as the corresponding iterated integral:

$$\int_{(x,y,z)\in D} f(x,y,z)\,|\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z| = \int_{(x,y)\in R} \left(\int_{z=g(x,y)}^{h(x,y)} f(x,y,z)\,\mathrm{d}z \right) |\mathrm{d}x \wedge \mathrm{d}y|$$

(and similarly in more variables).

The last two of these are the Fubini Theorem (for Riemann integrals of continuous functions).

By itself, the Fubini Theorem only works for regions of particular shapes, but the other theorems combine to make it more useful. First of all, Theorem 3 allows us to put the variables in whatever order we like. Even so, the regions still require particular shapes; we can just orient those however we wish. Theorem 2, in principle, allows us to divide a region up into smaller regions appropriate for the Fubini Theorem; the only question is whether the integrals exist. Theorem 1 guarantees this existence for continuous functions.

So using these in order, if you want to integrate over a crazy region, then divide the region into pieces of suitable shape. If the function is continuous and these smaller regions are all compact, then you know that their integrals exist; and if the regions overlap only slightly, then you can recover the answer to the original problem by adding them up. Finally, to get the integrals on these small regions, think of the variables as coming in whichever order works best, and use the Fubini Theorem (possibly more than once) to replace double and triple integrals with iterated integrals. Hopefully, these will be integrals that you can do!

Change of variables in multiple integrals

I often say that the differentials in expressions such as $3\,\mathrm{d}x + x^2\,\mathrm{d}y + \mathrm{e}^y\,\mathrm{d}z$, $\int_{x=0}^{1} 3x^2\,\mathrm{d}x$, and $\mathrm{d}y/\mathrm{d}x$ can and should be treated literally, not merely as mnemonics for appreciable changes in a limit or an approximation. For this to work in multiple (double, triple, etc) integrals, this requires a little care.

Change of variables in single integrals

One example of how it's useful to take differentials literally is that one can do a change of variables in a single-variable integral by calculating with differentials; for example, to integrate $\sqrt{1-x^2}\,\mathrm{d}x$ (say from $x=0$ to $x=1$), let $u = \operatorname{asin} x$, so that $x = \sin u$ and $\mathrm{d}x = \cos u\,\mathrm{d}u$, and calculate:

$$\int_{x=0}^{1} \sqrt{1-x^2}\,\mathrm{d}x = \int_{u=\operatorname{asin} 0}^{\operatorname{asin} 1} \sqrt{1-(\sin u)^2}(\cos u\,\mathrm{d}u) = \int_{u=0}^{\pi/2} \cos^2 u\,\mathrm{d}u = \left(\frac{1}{2}u + \frac{1}{4}\sin(2u)\right)\Big|_{u=0}^{\pi/2} = \frac{\pi}{4}.$$

(Incidentally, to integrate $\cos^2 u\,\mathrm{d}u$, I used the trigonometric identity that $\cos^2\theta = 1/2 + 1/2\cos(2\theta)$. This, along with $\sin^2\theta = 1/2 - 1/2\cos(2\theta)$, will come up a lot in the rest of this course.) You can even develop a general formula for this change of variables:

$$\int_{x=a}^{b} f(x)\,\mathrm{d}x = \int_{u=\operatorname{asin} a}^{\operatorname{asin} b} f(\sin u)\cos u\,\mathrm{d}u.$$

If you use this formula with $a=0$, $b=1$, and $f(x) = \sqrt{1-x^2}$ for all x, then you recover the previous calculation. (This is really the same idea that I used for integrating along curves in the handout from February 6.)

There is one big difference between single-variable integrals as they are usually done in Calculus and multiple integrals: single-variable integrals are oriented ($\int_{x=a}^{b}$ is the integral as x runs from a to b, whereas $\int_{x=b}^{a}$ is the integral as x runs from b to a, regardless of whether $a \le b$ or $b \le a$), while multiple integrals are unoriented ($\int_{(x,y)\in R}$ is the integral on the region R in the (x,y)-plane, without specifying any particular direction in that region). In other words, single-variable integrals are like the integrals from Section 15.2 of the textbook, integrals along oriented curves; while multiple integrals are like the integrals from Section 15.1, integrals on unoriented curves. So, to make the single-variable example above more like a multiple integral, I'll write it as

$$\int_{0\le x\le 1} \sqrt{1-x^2}\,|\mathrm{d}x|.$$

You can interpret this directly as an integral on a curve, where the curve in question (actually a straight line segment) is the interval $[0,1]$ on the real number line. Like the integrals in Section 15.1, this needs $|\mathrm{d}x| = \sqrt{\mathrm{d}x^2}$ so that the orientation (from 0 to 1 or from 1 to 0) makes no difference:

$$\int_{0\le x\le 1} \sqrt{1-x^2}\,|\mathrm{d}x| = \int_{x=0}^{1} \sqrt{1-x^2}\,\mathrm{d}x = \int_{x=1}^{0} \sqrt{1-x^2}\,(-\mathrm{d}x),$$

where first $|\mathrm{d}x| = \mathrm{d}x$ because x is increasing from 0 to 1 and next $|\mathrm{d}x| = -\mathrm{d}x$ because x is decreasing from 1 to 0; both integrals evaluate to $\pi/4$.

Then to redo the subsitution $u = \operatorname{asin} x$, instead of simply $\mathrm{d}x = \cos u\,\mathrm{d}u$, what really matters is that

$$|\mathrm{d}x| = |\cos u\,\mathrm{d}u| = |\cos u|\,|\mathrm{d}u| = \cos u\,|\mathrm{d}u|.$$

(I can simplify $|\cos u|$ to $\cos u$ because $u = \operatorname{asin} x$ means that $-\pi/2 \le u \le \pi/2$, so that $\cos u \ge 0$. Actually, I already used this fact, when I simplified $\sqrt{1-\sin^2 u}$ to $\cos u$ instead of to $|\cos u|$.) Now the general formula for the substitution is

$$\int_{a\le x\le b} f(x)\,|\mathrm{d}x| = \int_{\operatorname{asin} a\le u\le \operatorname{asin} b} f(\sin u)\cos u\,|\mathrm{d}u|,$$

and the specific example is

$$\int_{0\le x\le 1}\sqrt{1-x^2}\,|\mathrm{d}x| = \int_{0\le u\le \pi/2}\sqrt{1-\sin^2 u}\cos u\,|\mathrm{d}u| = \int_{u\in 0}^{\pi/2}\cos^2\,\mathrm{d}u = \frac{\pi}{4}.$$

To actually evaluate this integral, I had to switch from $\int_{0\le u\le\pi/2}$ to $\int_{u=0}^{\pi/2}$ and turn $|\mathrm{d}u|$ into $\mathrm{d}u$ (because u is increasing from 0 to $\pi/2$); you should think of this as the one-dimensional analogue of turning a multiple integral into an iterated integral (where again the normal way of doing this sets up the bounds on the integrals so that the variables are increasing).

Although I gave a general formula for the substitution $u = \operatorname{asin} x$, I can give an even more general formula, for an arbitrary substitution, where u is an arbitrary function of x (well, as long as that function is differentiable and one-to-one with a differentiable inverse). To make it look more like the formulas for multiple integrals, I'll write this as $x = g(u)$ instead of $u = g(x)$; since g is one-to-one, however, you can also write $u = g^{-1}(x)$. Then $\mathrm{d}x = g'(u)\,\mathrm{d}u$, so

$$\int_{a\le x\le b} f(x)\,|\mathrm{d}x| = \int_{g^{-1}(a)\le u\le g^{-1}(b)} f\big(g(u)\big)\,|g'(u)|\,|\mathrm{d}u|$$

if g (and hence g^{-1}) is increasing, or

$$\int_{a\le x\le b} f(x)\,|\mathrm{d}x| = \int_{g^{-1}(b)\le u\le g^{-1}(a)} f\big(g(u)\big)\,|g'(u)|\,|\mathrm{d}u|$$

if g (and hence g^{-1}) is decreasing. (A one-to-one function defined on an interval must be either increasing or decreasing to be continuous; otherwise, it would violate the Intermediate Value Theorem.)

To avoid the ambiguity of whether g is increasing or decreasing (and to make things look even more like the multi-variable case), I'll write $x \in R$ instead of $a \le x \le b$, so that R is the interval $[a, b]$, and I'll write $u \in G$ instead of either $g^{-1}(a) \le u \le g^{-1}(b)$ or $g^{-1}(b) \le u \le g^{-1}(a)$, so that G is $[g^{-1}(a), g^{-1}(b)]$ or $[g^{-1}(b), g^{-1}(a)]$, whichever makes sense. (You could also specify G as $\{u \mid a \le g(u) \le b\}$, or even as simply $\{u \mid g(u) \in R\}$. This G is called the *preimage* of R under g.) Then the formula is

$$\int_{x\in R} f(x)\,|\mathrm{d}x| = \int_{u\in G} f\big(g(u)\big)\,|g'(u)|\,|\mathrm{d}u|.$$

This is the complete analogue of the change-of-variables formula for double integrals that appears at the top of page 49 of these notes.

The wedge product

There is another complication that only appears with more than one variable. In 2 variables, for example, the textbook writes

$$\iint_R f(x,y)\,\mathrm{d}x\,\mathrm{d}y,$$

and in class I usually write

$$\int_{(x,y)\in R} f(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

(because I am more pedantic about putting variables in the correct place). But on page 43, I wrote even more precisely

$$\int_{(x,y)\in R} f(x,y)\,|\mathrm{d}x\wedge\mathrm{d}y|.$$

You can already see where the absolute value is coming from; as with $|\mathrm{d}x|$ in the one-variable case, it's because we're integrating over an unoriented region R. But now I want to explain the wedge ($\wedge$).

The **wedge product** of differential forms is kind of like the cross product of vectors; however, instead of trying to interpret it as another vector (or a scalar), it is simply another differential form, but one of higher 'rank' than the original forms. (Just as the operation that produces the cross product may be called outer multiplication of vectors, so the operation that produces the wedge product may be called

exterior multiplication of differential forms, but the term 'wedge product' is much more common.) You've used differential forms earlier in this course; those have rank 1, and they can be evaluted at a point and a vector. To evaluate a differential form of rank 2, you need a point and 2 vectors; to evaluate a differential form of rank 3, you need a point and 3 vectors; and so on.

The wedge product also involves subtracting one thing from another (again like the cross product); if α and β are 1-forms (differential forms of rank 1, as we've been using so far), P_0 is a point, and $\mathbf{v}_1$ and $\mathbf{v}_2$ are vectors, then

$$(\alpha \wedge \beta)\big|_{\substack{P=P_0,\\ \mathrm{d}P=\mathbf{v}_1,\mathbf{v}_2}} = \frac{1}{2}\alpha\big|_{\substack{P=P_0,\\ \mathrm{d}P=\mathbf{v}_1}} \beta\big|_{\substack{P=P_0,\\ \mathrm{d}P=\mathbf{v}_2}} - \frac{1}{2}\alpha\big|_{\substack{P=P_0,\\ \mathrm{d}P=\mathbf{v}_2}} \beta\big|_{\substack{P=P_0,\\ \mathrm{d}P=\mathbf{v}_1}}.$$

That is, to evaluate the wedge product $\alpha \wedge \beta$ at a point P_0 and two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$, first evaluate α at P_0 and $\mathbf{v}_1$ and evaluate β at P_0 and $\mathbf{v}_2$, multiply the results, then swap which vector goes with which differential form, evaluate and multiply again, then subtract the two products, and divide by 2. For example, if $\alpha = x^2\,\mathrm{d}x + xy\,\mathrm{d}y$, $\beta = y^2\,\mathrm{d}x - xy\,\mathrm{d}y$, $P_0 = (2,3)$, $\mathbf{v}_1 = \langle 0.01, 0.04\rangle$, and $\mathbf{v}_2 = \langle -0.01, 0\rangle$, then

$$\begin{aligned}
&\Big((x^2\,\mathrm{d}x + xy\,\mathrm{d}y) \wedge (y^2\,\mathrm{d}x - xy\,\mathrm{d}x)\Big)\Big|_{\substack{(x,y)=(2,3),\\ \mathrm{d}(x,y)=\langle 0.01,0.04\rangle,\langle -0.01,0\rangle}} \\
&\quad = \frac{1}{2}(x^2\,\mathrm{d}x + xy\,\mathrm{d}y)\big|_{\substack{(x,y)=(2,3),\\ \langle \mathrm{d}x,\mathrm{d}y\rangle=\langle 0.01,0.04\rangle}} (y^2\,\mathrm{d}x - xy\,\mathrm{d}y)\big|_{\substack{(x,y)=(2,3),\\ \langle \mathrm{d}x,\mathrm{d}y\rangle=\langle -0.01,0\rangle}} \\
&\qquad - \frac{1}{2}(x^2\,\mathrm{d}x + xy\,\mathrm{d}y)\big|_{\substack{(x,y)=(2,3),\\ \langle \mathrm{d}x,\mathrm{d}y\rangle=\langle -0.01,0\rangle}} (y^2\,\mathrm{d}x - xy\,\mathrm{d}y)\big|_{\substack{(x,y)=(2,3),\\ \langle \mathrm{d}x,\mathrm{d}y\rangle=\langle 0.01,0.04\rangle}} \\
&\qquad\quad = \frac{1}{2}\Big((2)^2(0.01) + (2)(3)(0.04)\Big)\Big((3)^2(-0.01) - (2)(3)(0)\Big) \\
&\qquad\qquad - \frac{1}{2}\Big((2)^2(-0.01) + (2)(3)(0)\Big)\Big((3)^2(0.01) - (2)(3)(0.04)\Big) \\
&\qquad\qquad\quad = (0.28)(-0.09) - (-0.04)(-0.15) = -0.0156.
\end{aligned}$$

A few basic properties of the wedge product follow immediately:

$$\begin{gathered}
\alpha \wedge (u\beta) = (u\alpha) \wedge \beta = u(\alpha \wedge \beta); \\
(\alpha + \beta) \wedge \gamma = \alpha \wedge \gamma + \beta \wedge \gamma; \\
\alpha \wedge (\beta + \gamma) = \alpha \wedge \beta + \alpha \wedge \gamma; \\
\alpha \wedge \beta = -\beta \wedge \alpha; \\
\alpha \wedge \alpha = 0,
\end{gathered}$$

where α, β, and γ are 1-forms and u is a 0-form, that is an ordinary non-differential quantity. (What these equations technically mean is that if you evaluate each side at the same point and vectors, then you'll get the same result on both sides, assuming that the operations appearing in the expressions are defined.) So if you treat the wedge product as a kind of multiplication, then you can use the ordinary rules of algebra, so long as you keep track of the order of multiplication in the wedge product and throw in a minus sign whenever you reverse the order of multiplication of two 1-forms (similarly to the cross product of vectors).

To see how this works, revisit the example above where $\alpha = x^2\,\mathrm{d}x + xy\,\mathrm{d}y$ and $\beta = y^2\,\mathrm{d}x - xy\,\mathrm{d}y$. The wedge product $\alpha \wedge \beta$ can be simplified as follows:

$$\begin{aligned}
\alpha \wedge \beta &= (x^2\,\mathrm{d}x + xy\,\mathrm{d}y) \wedge (y^2\,\mathrm{d}x - xy\,\mathrm{d}y) && * \\
&= (x^2\,\mathrm{d}x) \wedge (y^2\,\mathrm{d}x) + (x^2\,\mathrm{d}x) \wedge (-xy\,\mathrm{d}y) + (xy\,\mathrm{d}y) \wedge (y^2\,\mathrm{d}x) + (xy\,\mathrm{d}y) \wedge (-xy\,\mathrm{d}y) \\
&= (x^2)(y^2)(\mathrm{d}x \wedge \mathrm{d}x) + (x^2)(-xy)(\mathrm{d}x \wedge \mathrm{d}y) + (xy)(y^2)(\mathrm{d}y \wedge \mathrm{d}x) + (xy)(-xy)(\mathrm{d}y \wedge \mathrm{d}y) \\
&= x^2y^2(0) - x^3y\,\mathrm{d}x \wedge \mathrm{d}y + xy^3(-\mathrm{d}x \wedge \mathrm{d}y) - x^2y^2(0) && * \\
&= (-x^3y - xy^3)\,\mathrm{d}x \wedge \mathrm{d}y = -xy(x^2 + y^2)\,\mathrm{d}x \wedge \mathrm{d}y. && *
\end{aligned}$$

I've written this out in detail so that each step uses only one of the basic algebraic properties of the wedge product; but with a little practice, you should only need to write down the lines with asterisks after them. When you multiply the expressions (think FOIL), make sure to keep track of the order in which you multiply the differentials; if you multiply a differential by itself (such as $\mathrm{d}x \wedge \mathrm{d}x$), then you get zero, and if you multiply differentials in an order different from the order that you prefer (such as $\mathrm{d}y \wedge \mathrm{d}x$ instead of $\mathrm{d}x \wedge \mathrm{d}y$ if you prefer alphabetical order), then you can rearrange the order if you throw in a minus sign whenever two differentials switch places. In this way, you can go from the first line in the calculation above to the next line with an asterisk, skipping over the lines in between. (With a little more practice, you can even skip that line and go straight from the first line to the last line.)

To check that this simplification of $\alpha \wedge \beta$ is correct, I'll evaluate it again at $P_0 = (2, 3)$, $\mathbf{v}_1 = \langle 0.01, 0.04 \rangle$, and $\mathbf{v}_2 = \langle -0.01, 0 \rangle$. I get

$$\begin{aligned}(-xy(x^2+y^2)\,\mathrm{d}x \wedge \mathrm{d}y)\big|_{\substack{(x,y)=(2,3),\\ \mathrm{d}(x,y)=\langle 0.01,0.04\rangle,\langle -0.01,0\rangle}} &= \Big(-xy(x^2+y^2)\Big)\Big|_{(x,y)=(2,3)} (\mathrm{d}x \wedge \mathrm{d}y)|_{\langle \mathrm{d}x,\mathrm{d}y\rangle=\langle 0.01,0.04\rangle,\langle -0.01,0\rangle} \\ &= -(2)(3)\Big((2)^2+(3)^2\Big)\left(\frac{1}{2}(0.01)(0) - \frac{1}{2}(0.04)(-0.01)\right) = -0.0156,\end{aligned}$$

the same result as before. (Technically, what makes the original and simplified versions of $\alpha \wedge \beta$ equal to each other as differential forms is precisely that you will get the same results when evaluating them as long as you use the same point and vectors, no matter which point and vectors those are.)

To define a wedge product between forms of higher rank, you have to add and subtract all possible permutations of the possible orders in which to write the vectors at which the result is evaluated. Keeping track of all of this in a general formula is complicated, but the important point for our calculations is that the rules above continue to apply, and additionally we have an associative law for wedge products:

$$(\alpha \wedge \beta) \wedge \gamma = \alpha \wedge (\beta \wedge \gamma).$$

(This associative law is *not* true for cross products of vectors, so the wedge product is easier to work with.) We will not actually need to evaluate these higher-rank forms in this course; what's necessary is to work with them algebraically. In other words, the only new calculation in this chapter so far that you really need to know how to do is the one in the bottom part of the previous page.

Change of variables in multiple integrals

I'm now ready to explain change of variables in multiple integrals. If $x = g(u, v)$ and $y = h(u, v)$, where g and h are fixed differentiable binary functions, then

$$\begin{aligned}\mathrm{d}x \wedge \mathrm{d}y &= (D_1 g(u,v)\,\mathrm{d}u + D_2 g(u,v)\,\mathrm{d}v) \wedge (D_1 h(u,v)\,\mathrm{d}u + D_2 h(u,v)\,\mathrm{d}v) \\ &= D_1 g(u,v) D_1 h(u,v)\,\mathrm{d}u \wedge \mathrm{d}u + D_1 g(u,v) D_2 h(u,v)\,\mathrm{d}u \wedge \mathrm{d}v \\ &\quad + D_2 g(u,v) D_1 h(u,v)\,\mathrm{d}v \wedge \mathrm{d}u + D_2 g(u,v) D_2 h(u,v)\,\mathrm{d}v \wedge \mathrm{d}v \\ &= 0 + D_1 g(u,v) D_2 h(u,v)\,\mathrm{d}u \wedge \mathrm{d}v - D_2 g(u,v) D_1 h(u,v)\,\mathrm{d}u \wedge \mathrm{d}v + 0 \\ &= \Big(D_1 g(u,v) D_2 h(u,v) - D_2 g(u,v) D_1 h(u,v)\Big)\,\mathrm{d}u \wedge \mathrm{d}v.\end{aligned}$$

In other words,

$$\mathrm{d}x \wedge \mathrm{d}y = \left(\left(\frac{\partial x}{\partial u}\right)_v \left(\frac{\partial y}{\partial v}\right)_u - \left(\frac{\partial x}{\partial v}\right)_u \left(\frac{\partial y}{\partial u}\right)_v\right) \mathrm{d}u \wedge \mathrm{d}v.$$

You can also write this as

$$\mathrm{d}x \wedge \mathrm{d}y = \frac{\partial(x,y)}{\partial(u,v)}\,\mathrm{d}u \wedge \mathrm{d}v,$$

where

$$\frac{\partial(x,y)}{\partial(u,v)} = \left(\frac{\partial x}{\partial u}\right)_v \left(\frac{\partial y}{\partial v}\right)_u - \left(\frac{\partial y}{\partial u}\right)_v \left(\frac{\partial x}{\partial v}\right)_u = \begin{vmatrix} (\partial x/\partial u)_v & (\partial y/\partial u)_v \\ (\partial x/\partial v)_u & (\partial y/\partial v)_u \end{vmatrix}$$

is the **Jacobian determinant** of (x, y) with respect to (u, v).

The general formula for change of variables now simply requires absolute values:

$$\int_{(x,y)\in R} f(x,y)\,|\mathrm{d}x \wedge \mathrm{d}y| = \int_{(u,v)\in G} f\Big(g(u,v), h(u,v)\Big) \left|\frac{\partial(x,y)}{\partial(u,v)}\right| |\mathrm{d}u \wedge \mathrm{d}v|,$$

as long as $(u, v) \in G$ if and only if $\Big(g(u, v), h(u, v)\Big) \in R$ and (g, h) is jointly one-to-one (meaning that $(u_1, v_1) = (u_2, v_2)$ whenever $\Big(g(u_1, v_1), h(u_1, v_1)\Big) = \Big(g(u_2, v_2), h(u_2, v_2)\Big)$). It actually still works even if this one-to-one condition is violated, so long as the exceptions form a space of smaller dimension. (I'll explain this by way of example at the top of the next page, as part of my discussion of polar coordinates.)

The general formula in 3 dimensions is similar, but more complicated:

$$\int_{(x,y,z)\in R} f(x,y,z)\,|\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z| = \int_{(u,v,w)\in G} f\Big(g(u,v,w), h(u,v,w), k(u,v,w)\Big) \left|\frac{\partial(x,y,z)}{\partial(u,v,w)}\right| |\mathrm{d}u \wedge \mathrm{d}v \wedge \mathrm{d}w|,$$

where

$$\begin{aligned}\frac{\partial(x,y,z)}{\partial(u,v,w)} &= \left(\frac{\partial x}{\partial u}\right)_{v,w}\left(\frac{\partial y}{\partial v}\right)_{u,w}\left(\frac{\partial z}{\partial w}\right)_{u,v} - \left(\frac{\partial x}{\partial u}\right)_{v,w}\left(\frac{\partial z}{\partial v}\right)_{u,w}\left(\frac{\partial y}{\partial w}\right)_{u,v} - \left(\frac{\partial y}{\partial u}\right)_{v,w}\left(\frac{\partial x}{\partial v}\right)_{u,w}\left(\frac{\partial z}{\partial w}\right)_{u,v} \\ &\quad + \left(\frac{\partial y}{\partial u}\right)_{v,w}\left(\frac{\partial z}{\partial v}\right)_{u,w}\left(\frac{\partial x}{\partial w}\right)_{u,v} + \left(\frac{\partial z}{\partial u}\right)_{v,w}\left(\frac{\partial x}{\partial v}\right)_{u,w}\left(\frac{\partial y}{\partial w}\right)_{u,v} - \left(\frac{\partial z}{\partial u}\right)_{v,w}\left(\frac{\partial y}{\partial v}\right)_{u,w}\left(\frac{\partial x}{\partial w}\right)_{u,v} \\ &= \begin{vmatrix} (\partial x/\partial u)_{v,w} & (\partial y/\partial u)_{v,w} & (\partial z/\partial u)_{v,w} \\ (\partial x/\partial v)_{u,w} & (\partial y/\partial v)_{u,w} & (\partial z/\partial v)_{u,w} \\ (\partial x/\partial w)_{u,v} & (\partial y/\partial w)_{u,v} & (\partial z/\partial w)_{u,v} \end{vmatrix},\end{aligned}$$

as long as (f, g, h) is jointly one-to-one (or close) and $(u, v, w) \in G$ if and only if $\Big(g(u, v, w), h(u, v, w), k(u, v, w)\Big) \in R$.

Polar coordinates

Polar coordinates are a widely used example. In 2 dimensions, using $x = r\cos\theta$ and $y = r\sin\theta$,

$$\begin{aligned}\mathrm{d}x \wedge \mathrm{d}y &= (\cos\theta\,\mathrm{d}r - r\sin\theta\,\mathrm{d}\theta) \wedge (\sin\theta\,\mathrm{d}r + r\cos\theta\,\mathrm{d}\theta) \\ &= \cos\theta\sin\theta\,(0) + r\cos^2\theta\,(\mathrm{d}r \wedge \mathrm{d}\theta) - r\sin^2\theta\,(-\mathrm{d}r \wedge \mathrm{d}\theta) - r^2\sin\theta\cos\theta\,(0) \\ &= (r\cos^2\theta + r\sin^2\theta)\,\mathrm{d}r \wedge \mathrm{d}\theta = r\,\mathrm{d}r \wedge \mathrm{d}\theta,\end{aligned}$$

so

$$|\mathrm{d}x \wedge \mathrm{d}y| = |r|\,|\mathrm{d}r \wedge \mathrm{d}\theta| = r\,|\mathrm{d}r \wedge \mathrm{d}\theta|$$

as long as we only use coordinates where $r \geq 0$ (which is always possible). In 3 dimensions, throwing in z gives cylindrical coordinates:

$$|\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z| = |r|\,|\mathrm{d}z \wedge \mathrm{d}r \wedge \mathrm{d}\theta| = r\,|\mathrm{d}z \wedge \mathrm{d}r \wedge \mathrm{d}\theta|.$$

Then switching from (z, r) to (ρ, ϕ) in exactly the same way that polar coordinates switch from (x, y) to (r, θ) (so $z = \rho\cos\phi$ and $r = \rho\sin\phi$) gives spherical coordinates:

$$|\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z| = |r|\,|\rho|\,|\mathrm{d}\rho \wedge \mathrm{d}\phi \wedge \mathrm{d}\theta| = \rho^2\,|\sin\phi|\,|\mathrm{d}\rho \wedge \mathrm{d}\phi \wedge \mathrm{d}\theta| = \rho^2\sin\phi\,|\mathrm{d}\rho \wedge \mathrm{d}\phi \wedge \mathrm{d}\theta|$$

if $r \geq 0$ and $\rho \geq 0$. (Since $r = \rho\sin\phi$, if $r \geq 0$ and $\rho \geq 0$, then $\sin\phi \geq 0$ too.)

These must all be used with restrictions on the allowed values of the polar coordinates, in order for the change of variables to be one-to-one (mostly). The usual choices are $r \geq 0$, $0 \leq \theta \leq 2\pi$, $\rho \geq 0$, and $0 \leq \phi \leq \pi$. (These are consistent, since $r = \rho \sin\phi \geq 0$ when $\rho \geq 0$ and $0 \leq \phi \leq \pi$.) If you don't use $r \geq 0$ and $\rho \geq 0$, then you need more absolute values in the formulas, and $0 \leq \phi \leq \pi$ is the only good choice for ϕ (since it produces $\phi = \operatorname{asin}(r/\rho)$ when $\rho \neq 0$), but people sometimes use $-\pi \leq \theta \leq \pi$ or $-\pi/2 \leq \theta \leq 3\pi/2$ instead of $0 \leq \theta \leq 2\pi$, especially when integrating over a region that doesn't go all of the way around. Any choice $a \leq \theta \leq b$ is valid as long as $b - a = 2\pi$.

Whatever you use for θ, there is overlap where θ comes back to where it started, since $\sin a = \sin b$ and $\cos a = \cos b$ when $b - a = 2\pi$ (and θ only appears in those forms). However, this is contained within a single line in 2 dimensions and contained within a single plane in 3 dimensions, which doesn't affect the value of any integral. (There is no corresponding overlap with ϕ, since there we have $b - a = \pi$ rather than 2π.) Besides this, all values of θ produce the same result when $r = 0$, but again, this is contained within a single line in 2 dimensions and contained within a single plane in 3 dimensions. (It looks even lower in dimension in rectangular coordinates, a point in the (x, y)-plane and a line in (x, y, z)-space, but the dimensions that matter are in the (r, θ)-plane and in (z, r, θ)-space.) Similarly, all values of ϕ produce the same result when $\rho = 0$, but this is contained within a single plane (in (ρ, ϕ, θ)-space).

Therefore,

$$\int_{(x,y)\in R} f(x,y)\,|\mathrm{d}x \wedge \mathrm{d}y| = \int_{(r,\theta)\in G} f(r\cos\theta, r\sin\theta) r\,|\mathrm{d}r \wedge \mathrm{d}\theta|,$$

as long as

$$G = \{r, \theta \mid (r\cos\theta, r\sin\theta) \in R,\ r \geq 0,\ 0 \leq \theta \leq 2\pi\};$$

also,

$$\int_{(x,y,z)\in R} f(x,y,z)\,|\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z| = \int_{(z,r,\theta)\in G} f(r\cos\theta, r\sin\theta, z) r\,|\mathrm{d}z \wedge \mathrm{d}r \wedge \mathrm{d}\theta|,$$

as long as

$$G = \{z, r, \theta \mid (r\cos\theta, r\sin\theta, z) \in R,\ r \geq 0,\ 0 \leq \theta \leq 2\pi\};$$

finally,

$$\int_{(x,y,z)\in R} f(x,y,z)\,|\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z| = \int_{(\rho,\phi,\theta)\in G} f(\rho\sin\phi\cos\theta, \rho\cos\phi\sin\theta, \rho\cos\phi) \rho^2 \sin\phi\,|\mathrm{d}\rho \wedge \mathrm{d}\phi \wedge \mathrm{d}\theta|,$$

as long as

$$G = \{\rho, \phi, \theta \mid (\rho\sin\phi\cos\theta, \rho\sin\phi\sin\theta, \rho\cos\phi) \in R,\ \rho \geq 0,\ 0 \leq \phi \leq \pi,\ 0 \leq \theta \leq 2\pi\}.$$

While I'm on the subject, I should also warn you that different disciplines and different countries use different standard symbols for the polar coordinates. It's very common for ϕ and θ to be swapped completely (including using only ϕ in 2 dimensions), and r and ρ are also usually swapped, at least in 3 dimensions. It's pretty much only mathematicians in the USA who use the symbols as they are used in our textbook and as I have used them here. So watch out for this if you go to Europe or take a physics class!

Integration on surfaces

Just as you can integrate a differential 1-form (the ordinary kind without the wedge product) on an oriented curve, so you can integrate a **differential 2-form** (two 1-forms multiplied together by the wedge product or an expression built out of such products) on an oriented surface. (Similarly, you can integrate a differential 3-form on an oriented region of space, and so on for higher rank forms in spaces of higher dimension, but we're not doing any of that except for the volume integrals that we've already covered.)

Similarly, just as you can integrate a vector field along an oriented curve by taking a dot product with d$\mathbf{r}$ to get a differential 1-form and you can also integrate a vector field across a pseudooriented curve by taking a cross product with d$\mathbf{r}$ to get a differential pseudo-1-form (and then reintepreting this as an honest differential 1-form on an oriented curve), so you can integrate a vector field across a pseudooriented surface by taking a dot product with d$\mathbf{S}$ to get a differential pseudo-2-form (and then reintepreting this as an honest differential 2-form on an oriented surface).

So now I need to explain what all of this means.

Parametrizing surfaces

Just as you use 1 parameter (often called t) to parametrize a curve, so you use 2 variables (often called u and v) to parametrize a surface. For example, on the surface of the unit sphere (the sphere of radius 1 centred at $(x, y, z) = (0, 0, 0)$), we can use spherical coordinates with $\rho = 1$, so that

$$\begin{aligned} x &= r\cos\theta = \rho\sin\phi\cos\theta = \sin\phi\cos\theta, \\ y &= r\sin\theta = \rho\sin\phi\sin\theta = \sin\phi\sin\theta, \text{ and} \\ z &= \rho\cos\theta = \cos\theta. \end{aligned}$$

That is, ϕ and θ are the parameters. (You can call them u and v instead, but it's convenient to call them by more familiar names when possible.) Strictly speaking, the parametrization should also indicate the range of values taken by the parameters; in this case,

$$0 \le \phi \le \pi,\ 0 \le \theta \le 2\pi.$$

Now I have made this sphere into a **parametrized surface** (in 3-dimensional space).

In general, you can use ϕ and θ as parameters whenever the surface can be described by giving ρ as a function of ϕ and θ. (In the example above, that function was the constant function with value 1.) Besides using spherical coordinates, cylindrical coordinates are also often useful for parametrization. Most often, you'll use r and θ as the parameters, but sometimes you'll use z and θ; in any case, you'll need a way to express the other variable as a function of the two that you're using as parameters. Then using $x = r\cos\theta$ and $y = r\sin\theta$, you have x, y, and z all given as functions of the parameters. Finally, if you can express z as a function of x and y, then you can use x and y themselves as the parameters. (You could also use x and z or y and z, as long as the missing variable is given as a function of the two that you use.)

While most examples will use familiar coordinates as the parameters, in general, so long as you have $P = (x, y, z)$ given as a point-valued function of two variables u and v, then the range of this function is a **parametrized surface**. For purposes of integrals, this function should ideally be one-to-one, but as long as the overlap is contained within a few lines in the (u, v)-plane, then it won't affect the value of any integrals. (This is the same condition as for change of variables in a double integral.) In the case of cylindrical coordinates, the overlap is when θ is 0 or 2π, or (if r is being used as a parameter) when $r = 0$; but these are contained within lines. In the case of spherical coordinates, the overlap is when θ is 0 or 2π again, when $\phi = 0$, or when $\phi = \pi$; again, these are contained within lines. So cylindrical and spherical coordinates are always acceptable for integrals. (With rectangular coordinates, there is no overlap, so they are definitely acceptable.)

Orienting surfaces

In the case of a curve, there are two ways to go along the curve, giving two orientations. In the case of a surface, there are many ways to go along it, but if you start going in some direction, then you can *turn* from that direction in one way or the other; these give the two **orientations** of the surface. (Actually, not every surface can be oriented; a Möbius strip is a famous example of a surface that cannot be orinted continuously everywhere. However, any parametrized surface can be broken into pieces on which it can be oriented, so it is possible to do some integrals on unorientable surfaces, as long as they are integrals whose values don't depend on the orientation. Surface area and other integrals of scalar fields, discussed below, are examples of these.)

A differential form such as $\mathrm{d}u \wedge \mathrm{d}v$ *matches* the orientation of a surface if moving in the direction in which u increases and then turning in the direction in which v increases matches the surface's orientation. For example, the (x, y)-plane can be oriented clockwise or counterclockwise; $\mathrm{d}x \wedge \mathrm{d}y$ matches the counterclockwise orientation (if (x, y) is a counterclockwise coordinate system as usual), while $\mathrm{d}y \wedge \mathrm{d}x$ matches the clockwise orientation.

It's often easier to think of a **pseudoorientation** of a surface, which (in a 3-dimensional space) is a direction *across* the surface. The textbook never refers directly to orientations of surfaces, but only to pseudoorientations, which it (confusingly) calls 'orientations'. However, you can switch between orientations and pseudoorientations using the right-hand rule: if you curl the fingers of your right hand in the direction of turning indicated by an orientation, then your thumb will point in the direction of crossing indicated by the corresponding pseudoorientation. So the textbook applies this right-hand rule whenever it needs an orientation but really has a pseudoorientation.

Defining surface integrals

As with other definitions of integrals, people never use this directly if they can help it, and you'll never need to use it to solve any of the problems. But for the record, here it is.

So, suppose that you have a differential 2-form α written using the variables $P = (x, y, z)$ and their differentials, and an oriented surface in (x, y, z)-space, given by some parametrization function S (so that $P = (x, y, z) = S(u, v)$ on the surface) whose domain is a compact region R. Then we can try to integrate α along the surface, by defining the integral

$$\int_{P=S(u,v)} \alpha.$$

To form a Riemann sum to approximate this integral, dividing the region R into n triangles, pick one vertex of each triangle, and let $\mathbf{v}_k$ and $\mathbf{w}_k$ (where $k = 1, 2, \ldots n$ counts the triangles) the vectors (in the ambient (x, y, z)-space) from that vertex to the other two vertices; select which is $\mathbf{v}_k$ and which is $\mathbf{w}_k$ so that, when you turn from $\mathbf{v}_k$ to $\mathbf{w}_k$, this matches the orientation of the surface. Finally, tag this partition with a point c_k within each triangle. The **Riemann sum** is

$$\sum_{k=1}^{n} \alpha\big|_{\substack{P=S(c_k),\\ \mathrm{d}P=\mathbf{u}_k,\mathbf{v}_k}} .$$

If you require that the areas of the triangles to all be less than δ and take the limit of the Riemann sums as $\delta \to 0^+$, then the value of the integral is defined to be this limit, if it exists.

There is a theorem that this limit does exist, at least if α is piecewise continuous and S is piecewise continuously differentiable (and sometimes otherwise); I don't know a nice proof of this directly, but you can prove that it exists because the practical calculation method in the next section works. Similarly, there is now a theorem that the value of this integral does not depend on the parametrization of the surface, only the orientation.

Calculating integrals

The practical method of evaluating an integral along a surface is to pick any convenient parametrization (preferably one that is continuously differentiable) and put everything in terms of those parameters.

For example, I'll integrate $z\,\mathrm{d}x \wedge \mathrm{d}y$ on the top half of the unit sphere, oriented to turn clockwise when viewed from above the sphere. I'll use the parametrization given earlier using spherical coordinates:

$$\begin{aligned} x &= \sin\phi\cos\theta, \\ y &= \sin\phi\sin\theta, \\ z &= \cos\phi. \end{aligned}$$

Since I only want the top half of the sphere, I use

$$0 \le \phi \le \frac{\pi}{2},\ 0 \le \theta \le 2\pi.$$

Now I differentiate the parametrization:

$$\begin{aligned} \mathrm{d}x &= \cos\phi\cos\theta\,\mathrm{d}\phi - \sin\phi\sin\theta\,\mathrm{d}\theta, \\ \mathrm{d}y &= \cos\phi\sin\theta\,\mathrm{d}\phi + \sin\phi\cos\theta\,\mathrm{d}\theta, \\ \mathrm{d}z &= -\sin\phi\,\mathrm{d}\phi. \end{aligned}$$

Then

$$\mathrm{d}x \wedge \mathrm{d}y = \cos\phi\sin\phi\cos^2\theta\,\mathrm{d}\phi\wedge\mathrm{d}\theta - \sin\phi\cos\phi\sin^2\theta\,\mathrm{d}\theta\wedge\mathrm{d}\phi = \sin\phi\cos\phi\,\mathrm{d}\phi\wedge\mathrm{d}\theta.$$

(Remember that $\mathrm{d}\phi\wedge\mathrm{d}\phi$ and $\mathrm{d}\theta\wedge\mathrm{d}\theta$ are 0, so that half of the terms immediately vanish, and that $\mathrm{d}\theta\wedge\mathrm{d}\phi = -\mathrm{d}\phi\wedge\mathrm{d}\theta$, so that the other two terms can be combined into one.) Finally,

$$z\,\mathrm{d}x\wedge\mathrm{d}y = \sin\phi\cos^2\phi\,\mathrm{d}\phi\wedge\mathrm{d}\theta.$$

So, I am basically looking at

$$\int_{\substack{0\le\phi\le\pi/2,\\ 0\le\theta\le2\pi}} \sin\phi\cos^2\phi\,\mathrm{d}\phi\,\mathrm{d}\theta,$$

but I still need to think about the orientation. I really have $\mathrm{d}\phi\wedge\mathrm{d}\theta$ rather than $\mathrm{d}\phi\,\mathrm{d}\theta$, and this matches an orientation in which I turn from a direction in which ϕ increases to a direction in which θ increases. But this appears counterclockwise from above, while the orientation of the surface is clockwise from above. To fix this, I could rewrite the form to use $\mathrm{d}\theta\wedge\mathrm{d}\phi$, or equivalently put in a minus sign wherever $\mathrm{d}\phi\wedge\mathrm{d}\theta$ appears. So my real integral is

$$\int_{\theta=0}^{2\pi}\int_{\phi=0}^{\pi/2}(-\sin\phi\cos^2\phi)\,\mathrm{d}\phi\,\mathrm{d}\theta = \int_{\theta=0}^{2\pi}\left(-\frac{1}{3}\right)\mathrm{d}\theta = -\frac{2}{3}\pi.$$

You should be able to visualize this example geometrically well enough to see that the answer would have to be negative. Since $\mathrm{d}x\wedge\mathrm{d}y$ matches an orientation in which you turn from a direction in which x increases to a direction in which y increases, which appears counterclockwise from above, while the orientation is supposed to be clockwise from above, the factor $\mathrm{d}x\wedge\mathrm{d}y$ will always contribute something negative. The factor z, on the other hand, will always contribute something positive, since z is always positive on the top half of the sphere. So, the product $z\,\mathrm{d}x\wedge\mathrm{d}y$ will always be negative, so the overall integral must also be negative.

In this way, you can integrate any continuous differential 2-form on any surface with a continuously differentiable parametrization, because this process will always leave you with a continuous double integral to do.

Integrating vector fields

In the textbook, you'll never be directly given differential forms to integrate (other that 1-forms to integrate along curves). In some of Section 15.6 and much of Sections 15.7 and 15.8, you integrate a vector field across a surface; to integrate the vector field $\mathbf{F}$, you integrate the differential form $\mathbf{F}(x,y,z)\cdot đ\mathbf{S}$, where $đ\mathbf{S}$ is the **oriented surface element**

$$đ\mathbf{S} = \frac{1}{2}\,\mathrm{d}P \hat{\times} \mathrm{d}P = \langle \mathrm{d}y\wedge\mathrm{d}z, \mathrm{d}z\wedge\mathrm{d}x, \mathrm{d}x\wedge\mathrm{d}y\rangle = \frac{\partial P}{\partial u}\times\frac{\partial P}{\partial v}\,\mathrm{d}u\wedge\mathrm{d}v.$$

(People often write $đ\mathbf{S}$ as simply $\mathrm{d}\mathbf{S}$, although there is no quantity $\mathbf{S}$ that it is the differential of.) Here, $P=(x,y,z)$ as usual; the book prefers $\mathbf{r}=\langle x,y,z\rangle$, but since $\mathrm{d}P=\mathrm{d}\mathbf{r}$, partial derivatives of P and of $\mathbf{r}$ are the same, so we can equally well write

$$đ\mathbf{S} = \frac{1}{2}\,\mathrm{d}\mathbf{r} \hat{\times} \mathrm{d}\mathbf{r} = \langle \mathrm{d}y\wedge\mathrm{d}z, \mathrm{d}z\wedge\mathrm{d}x, \mathrm{d}x\wedge\mathrm{d}y\rangle = \frac{\partial \mathbf{r}}{\partial u}\times\frac{\partial \mathbf{r}}{\partial v}\,\mathrm{d}u\wedge\mathrm{d}v.$$

(When I write $\hat{\times}$ between vector-valued differential forms, I mean to multiply them as vectors using the cross product and as differential forms using the wedge product. Note that you get two minus signs when switching the order of multiplication, so the result of multiplying $\mathrm{d}P=\mathrm{d}\mathbf{r}$ by itself is not zero but rather twice something, and that something is what we mean by $đ\mathbf{S}$.)

The middle formula for $đ\mathbf{S}$ (the one without P or $\mathbf{r}$ in it) requires the use of the right-hand rule for the cross product. This is because $đ\mathbf{S}$ is really a **pseudoform**, meaning that it changes sign if you switch between right-hand and lett-hand rules. (Recall that multiplying vectors with the cross product similarly results in a pseudovector, also called an axial vector.) In this way, it makes sense to integrate a vector field through a pseudooriented surface; if you consistently use the left-hand rule instead of the right-hand rule, then the final result will be the same.

(The textbook never writes $đ\mathbf{S}$ or even $\mathrm{d}\mathbf{S}$; instead, it writes $\mathbf{n}\,đ\sigma$, or rather $\mathbf{n}\,\mathrm{d}\sigma$. But $đ\sigma$ is just $\|đ\mathbf{S}\|$, the magnitude of $đ\mathbf{S}$; and $\mathbf{n}$ is just $\widehat{đ\mathbf{S}}$, a unit vector in the direction of $đ\mathbf{S}$, that is a unit vector perpendicular to the surface pointing in the direction given by its pseudoorientation. So $\mathbf{n}\,\mathrm{d}\sigma$ is really just a complicated way of saying $đ\mathbf{S}$. To actually calculate $\mathbf{n}$ and $đ\sigma$ is a waste of time if $đ\mathbf{S}$ is all that you really want.)

So for example, integrating the constant vector field $\mathbf{F}(x,y,z)=\langle 0,0,z\rangle = z\mathbf{k}$ through the top half of the unit sphere pseudooriented downwards is the same as integrating the rank-2 differential form

$$\mathbf{F}(x,y,z)\cdot đ\mathbf{S} = \langle 0,0,z\rangle\cdot\langle \mathrm{d}y\wedge\mathrm{d}z, \mathrm{d}z\wedge\mathrm{d}x, \mathrm{d}x\wedge\mathrm{d}y\rangle = 0+0+z\,\mathrm{d}x\wedge\mathrm{d}y = z\,\mathrm{d}x\wedge\mathrm{d}y$$

on that hemisphere oriented clockwise when viewed from above, because turning the fingers of your right hand clockwise results in your thumb pointing downwards. Above, I calculated this integral to be $-2/3\,\pi$), and that is exactly how I would finish this problem.

Since the vector field that we integrated points upwards while the surface through which we integrated is pseudooriented downwards, you should expect the final result to be negative; guessing the sign of the integral ahead of time like this can help you to avoid mistakes with orientation. (If you used the left-hand rule instead, then you'd turn the fingers of your left hand counterclockwise to make your left thumb point downwards, but you'd also use $\langle \mathrm{d}z\wedge\mathrm{d}y, \mathrm{d}x\wedge\mathrm{d}z, \mathrm{d}y\wedge\mathrm{d}x\rangle$ for $đ\mathbf{S}$, and the final result would be the same.)

Integrating scalar fields

In Section 15.5 and some of Section 15.6, you integrate a scalar field (that is a functions of 3 variables) on a surface; to integrate the scalar field f, you integrate the differential form $f(x,y,z)\,đ\sigma$, where

$$đ\sigma = |đ\mathbf{S}| = \sqrt{(\mathrm{d}y\wedge\mathrm{d}z)^2+(\mathrm{d}z\wedge\mathrm{d}x)^2+(\mathrm{d}x\wedge\mathrm{d}y)^2} = \left|\frac{\partial\mathbf{r}}{\partial u}\times\frac{\partial\mathbf{r}}{\partial v}\right|\,|\mathrm{d}u\wedge\mathrm{d}v|.$$

Because the differentials only appear inside a vector magnitude, square, or absolute value (depending on which version you look at), orientation is irrelevant; instead, simply make sure that all parameters are increasing in the iterated integral.

So for example, integrating the scalar field $f(x,y,z) = z$ on the top half of the unit sphere is the same as integrating the rank-2 differential form

$$f(x,y,z)\,đ\sigma = z\sqrt{(\mathrm{d}y\wedge\mathrm{d}z)^2 + (\mathrm{d}z\wedge\mathrm{d}x)^2 + (\mathrm{d}x\wedge\mathrm{d}y)^2}$$

on that hemisphere with either orientation. To work out that expression using the parameters ϕ and θ, I can use $\mathrm{d}x\wedge\mathrm{d}y = \sin\phi\cos\phi\,\mathrm{d}\phi\wedge\mathrm{d}\theta$ from earlier, but I also need to find $\mathrm{d}y\wedge\mathrm{d}z$ and $\mathrm{d}z\wedge\mathrm{d}x$. I already have the individual differentials from page 3, so

$$\mathrm{d}y\wedge\mathrm{d}z = (\cos\phi\sin\theta\,\mathrm{d}\phi + \sin\phi\cos\theta\,\mathrm{d}\theta)\wedge(-\sin\phi\,\mathrm{d}\phi) = \sin^2\phi\cos\theta\,\mathrm{d}\phi\wedge\mathrm{d}\theta$$

and

$$\mathrm{d}z\wedge\mathrm{d}x = (-\sin\phi\,\mathrm{d}\phi)\wedge(\cos\phi\cos\theta\,\mathrm{d}\phi - \sin\phi\sin\theta\,\mathrm{d}\theta) = \sin^2\phi\sin\theta\,\mathrm{d}\phi\wedge\mathrm{d}\theta.$$

Therefore, I am integrating

$$\begin{aligned}&\cos\phi\sqrt{\sin^4\phi\cos^2\theta\,(\mathrm{d}\phi\wedge\mathrm{d}\theta)^2 + \sin^4\phi\sin^2\theta\,(\mathrm{d}\theta\wedge\mathrm{d}\phi)^2 + \sin^2\phi\cos^2\phi\,(\mathrm{d}\phi\wedge\mathrm{d}\theta)^2}\\ &= \cos\phi\sqrt{\sin^4\phi\,(\mathrm{d}\phi\wedge\mathrm{d}\theta)^2 + \sin^2\phi\cos^2\phi\,(\mathrm{d}\phi\wedge\mathrm{d}\theta)^2} = \cos\phi\sqrt{\sin^2\phi\,(\mathrm{d}\phi\wedge\mathrm{d}\theta)^2} = \sin\phi\cos\phi\,|\mathrm{d}\phi\wedge\mathrm{d}\theta|.\end{aligned}$$

(Here, I simplified $\sqrt{\sin^2\phi}$ to $\sin\phi$ rather than to $|\sin\phi|$, since $0\le\phi\le\pi$, so that $\sin\phi\ge 0$.)

The value of the integral is now

$$\int_{\theta=0}^{2\pi}\int_{\phi=0}^{\pi/2}\sin\phi\cos\phi\,\mathrm{d}\phi\,\mathrm{d}\theta = \int_{\theta=0}^{2\pi}\frac{1}{2}\,\mathrm{d}\theta = \pi.$$

(You should expect the integral to be positive, since z is always positive on the top hemisphere.) We don't have to think about orientation when setting up this iterated integral, since the integrand involves $|\mathrm{d}\phi\wedge\mathrm{d}\theta|$ rather than $\mathrm{d}\phi\wedge\mathrm{d}\theta$ itself; just make sure that the bounds are set up on the integrals so that each variable is increasing.

If instead I simply want the area of this surface, then I can simply integrate $đ\sigma$ itself, which gives

$$\int_{\theta=0}^{2\pi}\int_{\phi=0}^{\pi/2}\sin\phi\,\mathrm{d}\phi\,\mathrm{d}\theta = \int_{\theta=0}^{2\pi}\mathrm{d}\theta = 2\pi.$$

(And that is indeed the area of a hemisphere of radius 1.)

The Stokes theorems

The Stokes theorems generalize the (second) Fundamental Theorem of Calculus. The basic idea is that the integral of some differential form α on some manifold M (that is a curve, surface, etc) is equal to the integral of an antidifferential of α on the boundary of M.

In the case of one-variable Calculus, the theorem is

$$\int_{x=a}^{b} f(x)\,\mathrm{d}x = F(b) - F(a)$$

whenever $f = F'$. Here, $f(x)\,\mathrm{d}x$ is the differential form α; one of its antidifferentials is $F(x)$ (because the differential of $F(x)$ is $\mathrm{d}\big(F(x)\big) = F'(x)\,\mathrm{d}x = f(x)\,\mathrm{d}x = \alpha$). Also, the manifold M is the portion of the number line where x lies between a and b, thought of as a curve oriented from $x = a$ to $x = b$; its boundary consists of the points $x = a$ and $x = b$, with $x = a$ counted negatively and $x = b$ positively. In place of integrating the antidifferential $F(x)$ on these points, we evaluate it at those points and add the results (which really involves a subtraction since one of them is counted negatively).

In higher dimensions, the situation is in some ways easier to understand, because the boundary will be something more than just a few points, something that we're more used to talking about integrating on. Specifically, the boundary of a compact surface (whether a region in the 2-dimensional plane or a curved surface in 3-dimensional space) is a curve or a few curves, and the boundary of a compact region in 3-dimensional space is a surface or a few surfaces. Still, if you can think of evaluating a quantity as integrating it at a point, then all versions of the theorem can be expressed in the same way.

This is how the general **Stokes Theorem** looks:

$$\int_{\partial M} \alpha = \int_{M} \mathrm{d} \wedge \alpha.$$

Here, α is a differential form of a certain kind, called an *exterior* differential form, of some rank p, while M is an oriented manifold of dimension $p + 1$ (so a 1-dimensional curve if $p = 0$, a 2-dimensional surface if $p = 1$, etc). Also, ∂M is the **boundary** of M, which is an oriented manifold of dimension p (or perhaps a **chain** of several such manifolds); you know the symbol '∂' as used for partial derivatives, but it is also used to mean a boundary. Finally, $\mathrm{d} \wedge \alpha$ is a kind of differential of α, called the *exterior* differential, which I will explain next.

Exterior forms and exterior differentials

You may recall from the previous handout that a differential form of rank p (or p-form for short) may be constructed by taking p ordinary differential forms (those of rank 1) and multiplying them together using exterior multiplication to form their wedge product; more generally, if you apply operations such as addition to such expressions, then this still leaves you with a p-form. This can then be evaluated at a point along p vectors. So for example, $2y\,\mathrm{d}x \wedge \mathrm{d}y - 2x\,\mathrm{d}y \wedge \mathrm{d}z$ is formed by taking the wedge product of $2y\,\mathrm{d}x$ and $\mathrm{d}y$, giving the 2-form $2y\,\mathrm{d}x \wedge \mathrm{d}y$, taking the wedge product of $2x\,\mathrm{d}y$ and $\mathrm{d}z$, giving the 2-form $2x\,\mathrm{d}y \wedge \mathrm{d}z$, and subtracting these, giving the 2-form $2y\,\mathrm{d}x \wedge \mathrm{d}y - 2x\,\mathrm{d}y \wedge \mathrm{d}z$. This can now be evaluated at a point $(x, y, z) = P_0$ along two different vectors $\langle \mathrm{d}x, \mathrm{d}y, \mathrm{d}z \rangle = \mathbf{v}_1$ and $\langle \mathrm{d}x, \mathrm{d}y, \mathrm{d}z \rangle = \mathbf{v}_2$, by the method described on page 3 in the handout from February 21 (although you never to actually perform such an evaluation in this course).

You may also recall from the handout from January 18 that the differential of an ordinary quantity u (which we can now think of as a differential form of rank 0) is a 1-form, which you can evaluate at a point P_0 along a vector $\mathbf{v}_0$ by considering a parametrized curve through P_0 with $\mathbf{v}_0$ as its velocity vector there and seeing how fast u changes along that curve. Specifically, if you evaluate u at various points on the curve, then the result is a function of the parameter of the curve, and the derivative of this function (at the value of the parameter that gives the point P_0 and the tangent vector $\mathbf{v}_0$) is the result of evaluating $\mathrm{d}u$ at P_0 along $\mathbf{v}_0$.

We can combine these ideas to define the exterior differential of a p-form α. This will be a $(p + 1)$-form, so it can be evaluated at a point P_0 along $p + 1$ vectors $\mathbf{v}_0$, $\mathbf{v}_1$, ..., and $\mathbf{v}_p$. To evaluate this, consider a parametrized curve through P_0 with $\mathbf{v}_0$ as its velocity vector there. You can evaluate α at any

point on that curve along the same p vectors $\mathbf{v}_1$, $\mathbf{v}_2$, ..., and $\mathbf{v}_p$ to get a number that is a function of the parameter of the curve. The derivative of this function (at the value of the parameter that gives the point P_0 and the tangent vector $\mathbf{v}_0$) is *sort of* the result of evaluating $\mathrm{d}\wedge\alpha$ at P_0 along $\mathbf{v}_0$, $\mathbf{v}_1$, ..., and $\mathbf{v}_p$. I say 'sort of', because you now you need to consider a curve through P_0 whose velocity vector is $\mathbf{v}_1$ rather than $\mathbf{v}_0$ and do the same thing, evaluating α at a point on the curve along $\mathbf{v}_0$, $\mathbf{v}_2$, $\mathbf{v}_3$, ..., and $\mathbf{v}_p$. Continue in this way until you've gone along a curve whose tangent vector is $\mathbf{v}_p$. Now add up all of those results where you used a curve whose tangent vector was $\mathbf{v}_i$ for an even value of i, then subtract from this all of the results where you used a curve whose tangent vector was $\mathbf{v}_i$ for an odd value of i, and finally divide all of this by $p+1$. Now you *really* have the result of evaluating $\mathrm{d}\wedge\alpha$ at P_0 along $\mathbf{v}_0$, $\mathbf{v}_1$, ..., and $\mathbf{v}_p$.

Again, you're not actually going to need to do any such evaluation in this course. However, this definition leads to some fairly simple rules for calculating exterior differentials, and this will be useful. Here are some of the most important rules:

- $\mathrm{d}\wedge u = \mathrm{d}u$ when u is a differentiable 0-form;
- $\mathrm{d}\wedge\mathrm{d}u = 0$ when u is a twice-differentiable 0-form;
- $\mathrm{d}\wedge(\alpha+\beta) = \mathrm{d}\wedge\alpha + \mathrm{d}\wedge\beta$ when α and β are differentiable p-forms (for any whole number p);
- $\mathrm{d}\wedge(u\alpha) = \mathrm{d}u\wedge\alpha + u\,\mathrm{d}\wedge\alpha$ when u is a differentiable 0-form and α is a differentiable p-form (for any whole number p);
- $\mathrm{d}\wedge(\alpha\wedge\beta) = (\mathrm{d}\wedge\alpha)\wedge\beta - \alpha\wedge(\mathrm{d}\wedge\beta)$ when α is a differentiable 1-form and β is a differentiable p-form (for any whole number p).

In words: the exterior differential of an ordinary quantity is its ordinary differential that we've been using all along (because the definition is the same in this case); the exterior differential of one of these differentials is zero (ultimately because of the equality of mixed partial derivatives); the exterior differential obeys the Sum Rule (because every process in the definition obeys a Sum Rule); the exterior differential obeys a kind of Product Rule when multiplying by a 0-form, in which the differentials are multiplied by the wedge product; and the exterior differential of a wedge product with a 1-form obeys a kind of Product Rule with a minus sign in the term where the differential operator and the 1-form switch order. (These are all special cases of more complicated rules that apply to differential forms of any rank, except that the first rule really does only apply to 0-forms and the Sum Rule doesn't get any more complicated.)

Besides the Sum Rule (which should be very easy to use), the main rule that you'll really use is a combination of all of the others:

- $\mathrm{d}\wedge(u\,\mathrm{d}v\wedge\mathrm{d}w\wedge\cdots) = \mathrm{d}u\wedge\mathrm{d}v\wedge\mathrm{d}w\wedge\cdots$.

This will allow you to easily take the exterior differential of any differential form which is a sum of expressions like this. A differential form that is such a sum is called an **exterior differential form**, and it is these that we are most interested in. For example, $2y\,\mathrm{d}x\wedge\mathrm{d}y - 2x\,\mathrm{d}y\wedge\mathrm{d}z$ is an exterior differential 2-form, and its exterior differential is

$$\begin{aligned}\mathrm{d}\wedge(2y\,\mathrm{d}x\wedge\mathrm{d}y - 2x\,\mathrm{d}y\wedge\mathrm{d}z) &= \mathrm{d}(2y)\wedge\mathrm{d}x\wedge\mathrm{d}y - \mathrm{d}(2x)\wedge\mathrm{d}y\wedge\mathrm{d}z \\ &= 2\,\mathrm{d}y\wedge\mathrm{d}x\wedge\mathrm{d}y - 2\,\mathrm{d}x\wedge\mathrm{d}y\wedge\mathrm{d}z = -2\,\mathrm{d}x\wedge\mathrm{d}y\wedge\mathrm{d}z\end{aligned}$$

(because the first term, which multiplies $\mathrm{d}y$ by itself, is zero). I've written out this whole section for completeness, but the only thing that you *really* need to know about the exterior differential is how to perform calculations like the one above.

A note on notation: Besides simple 1-forms, the exterior differential forms are the most widely studied differential forms; and while the exterior differential is not the only way to differentiate these, it is by far the most widely studied way. For this reason, it's common to leave out the symbol '$\wedge$' in the wedge product and extremely common to leave out that symbol in the exterior differential. So the previous example may be written

$$\mathrm{d}(2y\,\mathrm{d}x\,\mathrm{d}y - 2x\,\mathrm{d}y\,\mathrm{d}z) = -2\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z.$$

The main reason why I'm *not* using this simplified notation myself is that we do occasionally multiply differential forms using ordinary multiplication (rather than exterior multplication) in expressions such as $đs = \sqrt{\mathrm{d}x^2 + \mathrm{d}y^2}$; here, $\mathrm{d}x^2$ really means the 1-form $\mathrm{d}x\,\mathrm{d}x$, *not* the 2-form $\mathrm{d}x\wedge\mathrm{d}x$, which would be zero.

(The formula for $đ\sigma$ in the previous handout even mixes wedge products with squares. There are also similar expressions, used in the geometry of surfaces and in general relativity for example, that have a $\mathrm{d}x\,\mathrm{d}y$ term under the square root, so it's not enough just to treat exponents differently.) But if you read other material on exterior differential forms, then it's very likely that some or all of the wedges will be left out.

Cohomology

One very important property of the exterior differential is that

$$\mathrm{d} \wedge (\mathrm{d} \wedge \alpha) = 0$$

whenever α is a twice-differentiable exterior form of any rank; that is, an exterior differential of an exterior differential is zero. The reason for this is the equality of mixed partial derivatives; if you write out the left-hand side explicitly in terms of partial derivatives of expressions appearing in α (which is very complicated but can be done), then you will see that everything cancels.

This fits in very nicely with the Stokes Theorem; if you apply it twice, then you get

$$\int_{\partial\partial M} \alpha = \int_{\partial M} \mathrm{d} \wedge \alpha = \int_M \mathrm{d} \wedge \mathrm{d} \wedge \alpha;$$

the right-hand side is zero because of the previous fact, while the left-hand side is zero because everything in the boundary of a boundary cancels. (For example, a curve bounding a surface must end where it began; similarly, a surface bounding a three-dimensional region must close in on itself and have no bounding curves.)

A manifold (or chain of manifolds) is called **closed** if it has no boundary (but this is different from being a closed set of points), and an exterior form is called **exact** if it's the exterior differential of some other exterior form. So the integral of an exact form on a closed manifold must be zero. This is where the terms 'closed' and 'exact' come from (in this context), but people also turn them around and use them this way: a chain of manifolds is **exact** if it's the boundary of some other manifold, and an exterior form is **closed** if its exterior differential is zero. Then the integral of a closed form on an exact manifold is also zero.

Because $\partial\partial M = 0$ and $\mathrm{d} \wedge \mathrm{d} \wedge \alpha = 0$, anything exact must also be closed. Conversely, a closed manifold in $\mathbf{R}^n$ must be exact, because you can simply fill it in to get something of one higher dimension that it bounds. Similarly, a closed exterior form in n variables is exact *if* it is defined for all possible values of those variables. However, if it is sometimes undefined, then it might not be, in particular if there are any gaps or holes in its domain.

It's therefore possible to study the topology of a manifold (roughly, those features of its shape that cannot be changed by continuously stretching or otherwise distorting it) by studying the exterior forms defined on it. The existence of closed but non-exact forms shows the existence of holes or gaps in the manifold, and the rank of the form in question even shows what kind of hole or gap. (The hole in a doughnut is different from both the gap between a pair of separated blobs and the hollow inside a sphere; these come from closed but non-exact forms of ranks 1, 0, and 2, respectively. This study of the topology of a shape by identifying and classifying holes in it is called *cohomology* (or specifically *de Rham cohomology* if you use closed but non-exact exterior differential forms to find the holes).

We won't really be doing any cohomology in this course; all that you really need to know are these facts:

- An exact differential form, if it is differntiable, is also closed;
- A closed differential form, if it is defined everywhere, is also exact.

The special case of this for 0-forms has already come up, on pages 4 and 5 in the handout from February 6.

Green's Theorem

Suppose that R is a compact region in the 2-dimensional plane, and suppose that the boundary of R is a curve C. (It's not true that every compact region has a curve for its boundary; Green's Theorem applies only to regions whose boundaries may be parametrized as a curve or a chain of curves.) Pseudoorient the curve C in the direction from within R to outside of R, and turn this into an orientation following the right-hand rule (so that you turn counterclockwise from the pseudoorientation to the orientation). In other words, C is oriented counterclockwise overall, with the region R on the left as you travel along its boundary C.

If f and g are functions of 2 variables (scalar fields) that are continuously differentiable at least on all of R, then

$$\int_{(x,y)\in C}(f(x,y)\,\mathrm{d}x + g(x,y)\,\mathrm{d}y) = \iint_{(x,y)\in R}\Big(D_1 g(x,y) - D_2 f(x,y)\Big)\,\text{đ}A.$$

Equivalently, if $u = f(x,y)$ and $v = g(x,y)$ are variable quantities that are continuously differentiable at least whenever $(x,y) \in R$, then

$$\int_{(x,y)\in C}(u\,\mathrm{d}x + v\,\mathrm{d}y) = \iint_{(x,y)\in R}\left(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}\right)\text{đ}A.$$

This result is Green's Theorem.

To see how this is a special case of the general Stokes Theorem, look at the exterior differential of $u\,\mathrm{d}x + v\,\mathrm{d}y$:

$$\begin{aligned}\mathrm{d}\wedge(u\,\mathrm{d}x + v\,\mathrm{d}y) = \mathrm{d}u\wedge\mathrm{d}x + \mathrm{d}v\wedge\mathrm{d}y &= \left(\frac{\partial u}{\partial x}\,\mathrm{d}x + \frac{\partial u}{\partial y}\,\mathrm{d}y\right)\wedge\mathrm{d}x + \left(\frac{\partial v}{\partial x}\,\mathrm{d}x + \frac{\partial v}{\partial y}\,\mathrm{d}y\right)\wedge\mathrm{d}y \\ &= 0 + \frac{\partial u}{\partial y}(-\mathrm{d}x\wedge\mathrm{d}y) + \frac{\partial v}{\partial x}\,\mathrm{d}x\wedge\mathrm{d}y + 0 = \left(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}\right)\mathrm{d}x\wedge\mathrm{d}y.\end{aligned}$$

Since $\mathrm{d}x\wedge\mathrm{d}y$ corresponds to the counterclockwise orientation of the region R, the boundary curve C must also be oriented counterclockwise around R.

If $\mathbf{F}$ is a continuously differentiable vector field in 2 dimensions, then we can integrate both $\mathbf{F}(x,y)\cdot\mathrm{d}\mathbf{r}$ and $\mathbf{F}(x,y)\times\mathrm{d}\mathbf{r}$ on a curve such as C. These will correspond to two different ways of differentiating $\mathbf{F}$. Writing $\mathbf{F} = \langle M, N\rangle$,

$$\mathbf{F}(x,y)\cdot\mathrm{d}\mathbf{r} = \langle M(x,y), N(x,y)\rangle\cdot\langle\mathrm{d}x,\mathrm{d}y\rangle = M(x,y)\,\mathrm{d}x + N(x,y)\,\mathrm{d}y,$$

so (using $f = M$ and $g = N$) Green's Theorem says that

$$\int_C \mathbf{F}(x,y)\cdot\mathrm{d}\mathbf{r} = \iint_R(\nabla\times\mathbf{F})(x,y)\,\mathrm{d}A,$$

where $\nabla\times\mathbf{F}$ is a scalar field called the **curl** of $\mathbf{F}$: $\nabla\times\mathbf{F} = \langle D_1, D_2\rangle\times\langle M, N\rangle = D_1 N - D_2 M$; or more explicitly,

$$(\nabla\times\mathbf{F})(x,y) = \frac{\partial\Big(N(x,y)\Big)}{\partial x} - \frac{\partial\Big(M(x,y)\Big)}{\partial y}.$$

On the other hand,

$$\mathbf{F}(x,y)\times\mathrm{d}\mathbf{r} = \langle M(x,y), N(x,y)\rangle\times\langle\mathrm{d}x,\mathrm{d}y\rangle = M(x,y)\,\mathrm{d}y - N(x,y)\,\mathrm{d}x,$$

so (now using $f = -N$ and $g = M$) Green's Theorem also says that

$$\int_C \mathbf{F}(x,y)\times\mathrm{d}\mathbf{r} = \iint_R \nabla\cdot\mathbf{F}(x,y)\,\mathrm{d}A,$$

where $\nabla \cdot \mathbf{F}$ is a scalar field called the **divergence** of $\mathbf{F}$: $\nabla \cdot \mathbf{F} = \langle D_1, D_2 \rangle \cdot \langle M, N \rangle = D_1 M + D_2 N$; or more explicitly,

$$(\nabla \cdot \mathbf{F})(x,y) = \frac{\partial\big(M(x,y)\big)}{\partial x} + \frac{\partial\big(N(x,y)\big)}{\partial y}.$$

These are both forms of Green's Theorem; thought of as theorems about vector fields, the first of these generalizes to Stokes's Theorem on page 6 below, while the last of these generalizes to Gauss's Theorem on page 7. The fact that $\mathrm{d} \wedge \mathrm{d}u = 0$, where $u = f(x,y)$, can be intepreted to say that $\nabla \times \nabla f = 0$.

If the boundary of R consists of several curves, then we can still write Green's Theorem as

$$\int_{\partial R} (u\,\mathrm{d}x + v\,\mathrm{d}y) = \iint_R \left(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}\right) \mathrm{d}A$$

(or similarly for any of the versions involving scalar or vector fields), where you integrate along the boundary ∂R by integrating along each curve in that boundary and adding the integrals. But now only the outermost curve of the boundary is oriented clockwise overall; the inner curves are oriented clockwise instead. This still matches the orientation of R, because an inner curve surrounds a hole in R rather than R itself. You can also write the integral on the left-hand side as

$$\int_{C_0} (u\,\mathrm{d}x + v\,\mathrm{d}y) - \int_{C_1} (u\,\mathrm{d}x + v\,\mathrm{d}y) - \int_{C_2} (u\,\mathrm{d}x + v\,\mathrm{d}y) - \cdots,$$

where C_0 is the outermost curve surrounding R and C_1, C_2, etc are the inner curves surrounding the holes, if you orient them all counterclockwise this time. (This still assumes that R consists of a single piece; if R consists of more than one disconnected piece, then there will be more than one outer curve being added.)

The proof of Green's Theorem essentially relies on showing that it is true on a small rectangular region, say

$$R = \{x, y \mid x_1 \le x \le x_2,\ y_1 \le y \le y_2\}.$$

The boundary of this region is a rectangle, running from (x_1, y_1) to (x_2, y_1) to (x_2, y_2) to (x_1, y_2) back to (x_1, y_1), which I will divide into four line segments, each parametrized by x or y itself. Then the integral along the boundary is

$$\begin{aligned}
&\int_{x=x_1}^{x_2} \big(f(x,y_1)\,\mathrm{d}x + g(x,y_1)\,\mathrm{d}(y_1)\big) + \int_{y=y_1}^{y_2} (f(x_2,y)\,\mathrm{d}(x_2) + g(x_2,y)\,\mathrm{d}y) \\
&\quad + \int_{x=x_2}^{x_1} \big(f(x,y_2)\,\mathrm{d}x + g(x,y_2)\,\mathrm{d}(y_2)\big) + \int_{y=y_2}^{y_1} (f(x_1,y)\,\mathrm{d}(x_1) + g(x_1,y)\,\mathrm{d}y) \\
&= \int_{x=x_1}^{x_2} f(x,y_1)\,\mathrm{d}x + \int_{y=y_1}^{y_2} g(x_2,y)\,\mathrm{d}y - \int_{x=x_1}^{x_2} f(x,y_2)\,\mathrm{d}x - \int_{y=y_1}^{y_2} g(x_1,y)\,\mathrm{d}y \\
&= \int_{x=x_1}^{x_2} \big(f(x,y_1) - f(x,y_2)\big) + \int_{y=y_1}^{y_2} \big(g(x_2,y) - g(x_1,y)\big).
\end{aligned}$$

Meanwhile, the integral on the rectangular region itself is

$$\begin{aligned}
&\int_{\substack{x_1 \le x \le x_2, \\ y_1 \le y \le y_2}} \left(\frac{\partial\big(g(x,y)\big)}{\partial x} - \frac{\partial\big(f(x,y)\big)}{\partial y}\right) \mathrm{d}A \\
&= \int_{\substack{x_1 \le x \le x_2 \\ y_1 \le y \le y_2}} \frac{\partial\big(g(x,y)\big)}{\partial x}\,\mathrm{d}A - \int_{\substack{x_1 \le x \le x_2 \\ y_1 \le y \le y_2}} \frac{\partial\big(f(x,y)\big)}{\partial y}\,\mathrm{d}A \\
&= \int_{y=y_1}^{y_2} \left(\int_{x=x_1}^{x_2} \frac{\partial\big(g(x,y)\big)}{\partial x}\,\mathrm{d}x\right) \mathrm{d}y - \int_{x=x_1}^{x_2} \left(\int_{y=y_1}^{y_2} \frac{\partial\big(f(x,y)\big)}{\partial y}\,\mathrm{d}y\right) \mathrm{d}x \\
&= \int_{y=y_1}^{y_2} \big(g(x_2,y) - g(x_1,y)\big)\,\mathrm{d}y - \int_{x=x_1}^{x_2} \big(f(x,y_2) - f(x,y_1)\big)\,\mathrm{d}x \\
&= \int_{x=x_1}^{x_2} \big(f(x,y_1) - f(x,y_2)\big) + \int_{y=y_1}^{y_2} \big(g(x_2,y) - g(x_1,y)\big).
\end{aligned}$$

So Green's Theorem is definitely true for a rectangular region. Now, the integral on a region R is a limit of sums of integrals on rectangular regions, while the corresponding sums of integrals on the regions' boundaries mostly cancel, as the same line segment is integrated along in first one direction and then the other; only the integrals along the line segments near the boundary of R survive. The limit of this is the integral along the boundary itself, because the differential 1-form being integrated is an exterior form; the integral along a diagonal line segment from one point on the boundary curve to another nearby point is approximately the same as an integral along a horizontal line segment followed by an integral along a vertical line segment: $\int(u\,\mathrm{d}x + v\,\mathrm{d}y) = \int u\,\mathrm{d}x + \int v\,\mathrm{d}y$. This is ultimately why all of these theorems apply only to exterior differential forms; none of this would work for something like $|u\,\mathrm{d}x + v\,\mathrm{d}y|$. (You can similarly prove the general Stokes Theorem —in any rank and any dimension— all at once, if you take care to keep careful track of everything, but this is a lot of detail when written out in full.)

Stokes's Theorem

While Green's Theorem is about a region in the plane, Stokes's Theorem is about a curved surface in space. (Stokes's Theorem is also called the Kelvin–Stokes Theorem or the Curl Theorem; the general Stokes Theorem is named after this particular case.)

If $\mathbf{F} = \langle M, N, P\rangle$ is a continuously differentiable vector field in 3 dimensions and S is a compact surface in 3 dimensions with boundary curve C, then Stokes's Theorem says that

$$\int_C \mathbf{F}(x,y,z)\cdot\mathrm{d}\mathbf{r} = \iint_S (\nabla\times\mathbf{F})(x,y,z)\cdot\mathrm{d}\mathbf{S},$$

where $\nabla\times\mathbf{F}$ is a vector field called the **curl** of $\mathbf{F}$: $\nabla\times\mathbf{F} = \langle D_1, D_2, D_3\rangle \times \langle M, N, P\rangle = \langle D_2P - D_3N, D_3M - D_1P, D_1N - D_2M\rangle$; or more explicitly,

$$(\nabla\times\mathbf{F})(x,y,z) = \left\langle \frac{\partial\big(P(x,y,z)\big)}{\partial y} - \frac{\partial\big(N(x,y,z)\big)}{\partial z}, \frac{\partial\big(M(x,y,z)\big)}{\partial z} - \frac{\partial\big(P(x,y,z)\big)}{\partial x}, \frac{\partial\big(N(x,y,z)\big)}{\partial x} - \frac{\partial\big(M(x,y,z)\big)}{\partial y} \right\rangle.$$

Both the surface S and its boundary C are oriented here, and these orientations must match; that is, the direction in which you travel along C, as given by its orientation, must agree with the direction in which you turn along S, as given by its orientation, near the boundary. We usually think of a surface as being pseudooriented, that is given with a direction across it instead of directions along it, relating this to the orientation of its boundary with the right-hand rule; but the formula for the curl also relies on the right-hand rule, so you could just as easily use the left-hand rule for both.

As with Green's Theorem, this generalizes to a surface whose boundary consists of several curves, but it won't apply to a surface whose boundary cannot be parametrized as curves at all. On the other hand, a parametrized surface could cross itself (just as a curve can), and Stokes's Theorem can continue to apply in that case. What ultimately matters is the region in the (u,v)-plane that the parametrization transforms into the actual surface; if the boundary of that region can be parametrized as curves, then Stokes's Theorem applies to the resulting surface. This also shows one way to prove Stokes's Theorem: by reducing it to Green's Theorem in the (u,v)-plane.

This classical Stokes's Theorem is a special case of the general Stokes Theorem: writing u for $M(x,y,z)$, v for $N(x,y,z)$, and w for $P(x,y,z)$,

$$\begin{aligned}
&\mathrm{d}\wedge(\mathbf{F}(x,y,z)\cdot\mathrm{d}\mathbf{r}) = \mathrm{d}\wedge(u\,\mathrm{d}x + v\,\mathrm{d}y + w\,\mathrm{d}z) = \mathrm{d}u\wedge\mathrm{d}x + \mathrm{d}v\wedge\mathrm{d}y + \mathrm{d}w\wedge\mathrm{d}z\\
&= \left(\frac{\partial u}{\partial x}\,\mathrm{d}x + \frac{\partial u}{\partial y}\,\mathrm{d}y + \frac{\partial u}{\partial z}\,\mathrm{d}z\right)\wedge\mathrm{d}x + \left(\frac{\partial v}{\partial x}\,\mathrm{d}x + \frac{\partial v}{\partial y}\,\mathrm{d}y + \frac{\partial v}{\partial z}\,\mathrm{d}z\right)\wedge\mathrm{d}y + \left(\frac{\partial w}{\partial x}\,\mathrm{d}x + \frac{\partial w}{\partial y}\,\mathrm{d}y + \frac{\partial w}{\partial z}\,\mathrm{d}z\right)\wedge\mathrm{d}z\\
&= 0 + \frac{\partial u}{\partial y}(-\mathrm{d}x\wedge\mathrm{d}y) + \frac{\partial u}{\partial z}\,\mathrm{d}z\wedge\mathrm{d}x + \frac{\partial v}{\partial x}\,\mathrm{d}x\wedge\mathrm{d}y + 0 + \frac{\partial v}{\partial z}(-\mathrm{d}y\wedge\mathrm{d}z) + \frac{\partial w}{\partial x}(-\mathrm{d}z\wedge\mathrm{d}x) + \frac{\partial w}{\partial y}\,\mathrm{d}y\wedge\mathrm{d}z + 0\\
&= \left(\frac{\partial w}{\partial y} - \frac{\partial v}{\partial z}\right)\mathrm{d}y\wedge\mathrm{d}z + \left(\frac{\partial u}{\partial z} - \frac{\partial w}{\partial x}\right)\mathrm{d}z\wedge\mathrm{d}x + \left(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}\right)\mathrm{d}x\wedge\mathrm{d}y = (\nabla\times\mathbf{F})(x,y,z)\cdot\mathrm{d}\mathbf{S}.
\end{aligned}$$

The fact that $\mathrm{d} \wedge \mathrm{d}u = 0$, when $u = f(x, y, z)$, becomes the result that

$$\nabla \times \nabla f = 0.$$

Conversely, if $\nabla \times \mathbf{F} = 0$, then $\mathbf{F}$ must be the gradient of some scalar field f, unless there must be a hole in the domain of $\mathbf{F}$ similar to the hole through a doughnut.

Gauss's Theorem

While Green's Theorem is about a region in the plane, Gauss's Theorem is about a region in space. (Gauss's Theorem is also called the Ostrogradsky–Gauss Theorem or the Divergence Theorem.)

If $\mathbf{G} = \langle M, N, P \rangle$ is a continuously differentiable vector field in 3 dimensions and D is a compact region in 3 dimensions with boundary surface S, then Gauss's Theorem says that

$$\int_S \mathbf{G}(x, y, z) \cdot \mathrm{d}\mathbf{S} = \iiint_D (\nabla \cdot \mathbf{G})(x, y, z)\, \mathrm{d}V,$$

where $\nabla \cdot \mathbf{G}$ is a scalar field called the **divergence** of $\mathbf{G}$: $\nabla \cdot \mathbf{G} = \langle D_1, D_2, D_3 \rangle \cdot \langle M, N, P \rangle = D_1 M + D_2 N + D_3 P$; or more explicitly,

$$(\nabla \cdot \mathbf{G})(x, y, z) = \frac{\partial\big(M(x, y, z)\big)}{\partial x} + \frac{\partial\big(N(x, y, z)\big)}{\partial y} + \frac{\partial\big(P(x, y, z)\big)}{\partial z}.$$

If you give the region D a right-handed orientation, then the orientation on the boundary surface S is counterclockwise as viewed from outside the region, which in turn corresponds under the right-hand rule to a pseudoorientation from inside to outside. You can just as easily use the left hand for everything, but the pseudoorientation on S will always be outwards.

As with Green's Theorem, this generalizes to a region whose boundary consists of several surfaces, but it won't apply to a region whose boundary cannot be parametrized as surfaces at all. Unlike Stokes's Theorem, Gauss's Theorem cannot be reduced to Green's Theorem using a parametrization; however, the proof of Gauss's Theorem is quite similar to the proof of Green's Theorem, only with an extra dimension: the integral along the boundary of a box splits into six pieces, grouped into three pairs, while the integral on the box splits into three terms whose integrals lead to the same three expressions.

Gauss's Theorem is also a special case of the Stokes Theorem: writing u for $M(x, y, z)$, v for $N(x, y, z)$, and w for $P(x, y, z)$,

$$\begin{aligned}
\mathrm{d} \wedge (\mathbf{G}(x, y, z) \cdot \mathrm{d}\mathbf{S}) &= \mathrm{d} \wedge (u\, \mathrm{d}y \wedge \mathrm{d}z + v\, \mathrm{d}z \wedge \mathrm{d}x + w\, \mathrm{d}x \wedge \mathrm{d}y) \\
&= \mathrm{d}u \wedge \mathrm{d}y \wedge \mathrm{d}z + \mathrm{d}v \wedge \mathrm{d}z \wedge \mathrm{d}x + \mathrm{d}w \wedge \mathrm{d}x \wedge \mathrm{d}y \\
&= \left(\frac{\partial u}{\partial x}\, \mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z + 0 + 0\right) + \left(0 + \frac{\partial v}{\partial y}\, \mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z + 0\right) + \left(0 + 0 + \frac{\partial w}{\partial z}\, \mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z\right) \\
&= \left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z}\right) \mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z = (\nabla \cdot \mathbf{G})(x, y, z)\, \mathrm{d}V.
\end{aligned}$$

The fact that $\mathrm{d} \wedge (\mathrm{d} \wedge \alpha) = 0$, when $\alpha = \mathbf{F}(x, y, z) \cdot \mathrm{d}\mathbf{r}$, becomes the result that

$$\nabla \cdot \nabla \times \mathbf{F} = 0.$$

Conversely, if $\nabla \cdot \mathbf{G} = 0$, then $\mathbf{G}$ must be the curl of some vector field $\mathbf{F}$, unless there is a hollow in the domain of $\mathbf{G}$ similar to the hollow inside a sphere.

of the integrals in vector calculus can be thought of as integrals of *differential forms* of one sort or another. Since integration of differential forms generalizes in ways that integration of vector fields cannot (some of which are important in applications, especially to physics), it's useful to be able to think about differential forms. Furthermore, you then need fewer formulas for the various derivatives of vector fields and for the theorems that relate derivatives to integrals.

General principles

Here I spell out the general principles of integrating differential forms, but it's really the examples that follow that will make the ideas clear.

There are three sorts of differential forms that we'll need: *exterior* forms, *pseudoexterior* forms, and *absolute* forms. The exterior forms are the most straightforward kind and the simplest to calculate with. The pseudoexterior forms are essentially the same as exterior forms, except that their sign is determined by using the right-hand rule; if you used the left-hand rule instead, then the pseudoexterior forms would have opposite sign but the results of all integrals would stay the same. (In general, you can put 'pseudo' before the name of a concept to get the name of a related concept where the sign depends on the right-hand rule. It is sometimes handy to keep track of whether something is pseudo or not; for example, if you ever add something pseudo to something nonpseudo, then you know that you're making a mistake, much as you would be if you added quantities measured in different units. However, you can ignore the difference in calculations as long as you always use the right-hand rule.) The absolute forms are least used in applications; they typically arise by taking the absolute value of another form (and then possibly multiplying by a scalar quantity). However, they are still important, since lengths, areas, and volumes may be found by integrating absolute forms. (If you read other material on differential forms, the exterior ones are the most commonly studied, and people will often leave out the word 'exterior'. Then the pseudoexterior forms are just called 'pseudoforms', and there is no common name for the absolute forms at all; 'absolute' is a term for them that I made up. On the other hand, there are yet other kinds of differential forms besides all of these.)

You integrate these forms along various regions in space, called *manifolds*. These manifolds can correspondingly be oriented, pseudooriented, or unoriented. Now it's the unoriented manifolds that are the simplest; they are just shapes of consistent dimension. With an oriented manifold, you also make a choice of which direction to go along the manifold; with a pseudooriented manifold, you instead make a choice of which direction to go around or across the manifold. You integrate exterior forms on oriented manifolds, pseudoexterior forms on pseudooriented manifolds, and absolute forms on unoriented manifolds. (If you read other material, the pseudooriented manifolds are sometimes also called 'transversely oriented'.) People also talk about integrating on *chains*: a chain is just a list of manifolds, each with a real number (its *weight*); to integrate a differential form on a chain, you multiply the integral on each manifold by that manifold's weight and then add these products. You'll see some simple examples of chains when we get to the Stokes Theorem below.

To calculate integrals, you want to parametrize your manifolds; you'll have one or more variables $t, u, v, \ldots$ (the *parameters*), running over some domain of values, and a point-valued function (the *parametrization*) of those variables specifying which point in space corresponds to which values of the parameters. Running this function over the entire domain of parameters carves out the manifold. (You'll want your parametrization functions to be continuously differentiable, in order to avoid worrying about whether the integrals are defined. For the same reason, the forms themselves should be continuous, and the domains of the paremetrizations should be compact, that is closed and bounded. The integrals may be defined in other cases, but they are guaranteed to exist if these conditions are met.)

The number of parameters used is the *dimension* of the manifold. This must match the *rank* of the differential form, which is the number of differentials in each term of the form. These differentials are combined using the *wedge product*, $\wedge$. A key property of the wedge product is that it is *anticommutative* between differentials; that is,

$$\mathrm{d}x \wedge \mathrm{d}y = -\mathrm{d}y \wedge \mathrm{d}x$$

(much like the cross product of vectors). This also means that $\mathrm{d}x \wedge \mathrm{d}x = 0$. However, for absolute forms, you take the absolute value of the wedge product; then $|\mathrm{d}x \wedge \mathrm{d}y| = |-\mathrm{d}y \wedge \mathrm{d}x| = |\mathrm{d}y \wedge \mathrm{d}x|$, while $|\mathrm{d}x \wedge \mathrm{d}x| = |0| = 0$ still.

To calculate the integral, you use the parametrization to express the coordinates $x, y, \ldots$ in terms of the parameters $t, u, v, \ldots$, then differentiate this to get $\mathrm{d}x, \mathrm{d}y, \ldots$ in terms of $\mathrm{d}t, \mathrm{d}u, \mathrm{d}v, \ldots$, so that the integral is entirely in terms of the parameters. You then express this as an iterated integral, checking the orientation or pseudoorientation and putting a minus sign out front if it goes the wrong way.

Summary of the integrals

This section repeats what we've already done, but shows explicitly how every integral that you deal with in this course is either the integral of an exterior form on an oriented manifold, the integral of a pseudoexterior form on a pseudooriented manifold, or the integral of an absolute form on an unoriented manifold.

Curves

A **curve** C is a manifold of dimension 1. So it may be parametrized by a function (which we'll assume is continuously differentiable) that takes one variable t to a point $P = (x, y, \ldots)$. Note that the differential $\mathrm{d}P = \langle \mathrm{d}x, \mathrm{d}y, \ldots \rangle$ is a vector; if you write $\mathbf{r}$ for the vector $P - (0, 0, \ldots)$, then $\mathrm{d}P = \mathrm{d}\mathbf{r}$, and $\mathrm{d}\mathbf{r}$ is the more usual notation (even though P is the more fundamental concept). When you orient a curve, you specify which direction to travel along the curve; when you pseudoorient a curve in 2 dimensions, you specify which direction to travel across the curve. (You won't need to pseudoorient a curve in more dimensions in this class, although it can be done by specifying directions around the curve.)

To integrate a vector quantity $\mathbf{F} = \langle M, N, \ldots \rangle$ along an oriented curve C, you integrate the rank-1 exterior form $\mathbf{F} \cdot \mathrm{d}\mathbf{r}$:

$$\int_C \mathbf{F} \cdot \mathrm{d}\mathbf{r} = \int_C \langle M, N, \ldots \rangle \cdot \langle \mathrm{d}x, \mathrm{d}y, \ldots \rangle = \int_C (M\,\mathrm{d}x + N\,\mathrm{d}y + \cdots) = \int_C \left(M\frac{\mathrm{d}x}{\mathrm{d}t} + N\frac{\mathrm{d}y}{\mathrm{d}t} + \cdots \right) \mathrm{d}t$$

or

$$\int_C \mathbf{F} \cdot \mathrm{d}\mathbf{r} = \int_C \mathbf{F} \cdot \frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t}\,\mathrm{d}t = \int_C \langle M, N, \ldots \rangle \cdot \left\langle \frac{\mathrm{d}x}{\mathrm{d}t}, \frac{\mathrm{d}y}{\mathrm{d}t}, \ldots \right\rangle \mathrm{d}t = \int_C \left(M\frac{\mathrm{d}x}{\mathrm{d}t} + N\frac{\mathrm{d}y}{\mathrm{d}t} + \ldots \right) \mathrm{d}t.$$

(There's no need to learn all of these formulas; just put everything in terms of t and push through.) To match orientations, make sure that the direction along the curve as t increases is the same direction as the curve's orientation; or if not, then put a minus sign out front.

To integrate a vector quantity $\mathbf{F} = \langle M, N \rangle$ across a pseudooriented curve C in 2 dimensions, you integrate the rank-1 pseudoexterior form $\mathbf{F} \times \mathrm{d}\mathbf{r}$ (where the cross product in 2 dimensions produces a scalar, or rather a pseudoscalar since the sign depends on the right-hand rule):

$$\int_C \mathbf{F} \times \mathrm{d}\mathbf{r} = \int_C \langle M, N \rangle \times \langle \mathrm{d}x, \mathrm{d}y \rangle = \int_C (M\,\mathrm{d}y - N\,\mathrm{d}x) = \int_C \left(M\frac{\mathrm{d}y}{\mathrm{d}t} - N\frac{\mathrm{d}x}{\mathrm{d}t} \right) \mathrm{d}t$$

or

$$\int_C \mathbf{F} \times \mathrm{d}\mathbf{r} = \int_C \mathbf{F} \times \frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t}\,\mathrm{d}t = \int_C \langle M, N \rangle \times \left\langle \frac{\mathrm{d}x}{\mathrm{d}t}, \frac{\mathrm{d}y}{\mathrm{d}t} \right\rangle \mathrm{d}t = \int_C \left(M\frac{\mathrm{d}y}{\mathrm{d}t} - N\frac{\mathrm{d}x}{\mathrm{d}t} \right) \mathrm{d}t.$$

To match pseudoorientations using the right-hand rule, make sure that the direction along the curve as t changes is counterclockwise from the direction of the curve's pseudoorientation; or if not, then put a minus sign out front.

To integrate a scalar quantity f on an unoriented curve C, you integrate the rank-1 absolute form $f\,đs$, where s has no meaning by itself but instead $đs$ is the absolute form $\|\mathrm{d}\mathbf{r}\|$:

$$\int_C f\,đs = \int_C f\,\|\mathrm{d}\mathbf{r}\| = \int_C f\,\|\langle \mathrm{d}x, \mathrm{d}y, \ldots \rangle\| = \int_C f\sqrt{(\mathrm{d}x)^2 + (\mathrm{d}y)^2 + \cdots} = \int_C f\sqrt{\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2 + \left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2 + \cdots}\,|\mathrm{d}t|$$

or

$$\int_C f\,đs = \int_C f\,\|\mathrm{d}\mathbf{r}\| = \int_C f\left\|\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t}\right\|\,|\mathrm{d}t| = \int_C f\left\|\left\langle \frac{\mathrm{d}x}{\mathrm{d}t}, \frac{\mathrm{d}y}{\mathrm{d}t}, \ldots \right\rangle\right\|\,|\mathrm{d}t| = \int_C f\sqrt{\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2 + \left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2 + \cdots}\,|\mathrm{d}t|.$$

Now there is no orientation to match; instead, make sure that t is increasing, so that $|\mathrm{d}t| = \mathrm{d}t$ in the integral; or if not, then put a minus sign out front.

Surfaces

A **surface** S is a manifold of dimension 2, given by a function (which we'll assume is continuously differentiable) that takes two variables u, v to a point $P = (x, y, z, \ldots)$ When you pseudoorient a surface in 3 dimensions, you specify which direction to travel across the surface. (You won't need to pseudoorient a surface in more dimensions, nor will you need to orient any at all, although again these can be done.)

To integrate a vector quantity $\mathbf{F} = \langle M, N, O \rangle$ across a pseudooriented surface S in 3 dimensions, you integrate the rank-2 pseudoexterior form $\mathbf{F} \cdot đ\mathbf{S}$, where $\mathbf{S}$ has no meaning by itself, but instead $đ\mathbf{S}$ is the pseudovector-valued form $1/2\, d\mathbf{r} \mathbin{\hat{\times}} d\mathbf{r}$ (which as a vector is multiplied by the cross product and as a differential form is multiplied by the wedge product). This works out to $\langle dy \wedge dz, dz \wedge dx, dx \wedge dy \rangle$ (using the right-hand rule) or $\partial\mathbf{r}/\partial u \times \partial\mathbf{r}/\partial v \, du \wedge dv$:

$$\begin{aligned}\int_S \mathbf{F} \cdot đ\mathbf{S} &= \int_S \langle M, N, O \rangle \cdot \langle dy \wedge dz, dz \wedge dx, dx \wedge dy \rangle = \int_S (M\, dy \wedge dz + N\, dz \wedge dx + O\, dx \wedge dy) \\ &= \int_S \left(M\left(\frac{\partial y}{\partial u}\frac{\partial z}{\partial v} - \frac{\partial y}{\partial v}\frac{\partial z}{\partial u}\right) + N\left(\frac{\partial z}{\partial u}\frac{\partial x}{\partial v} - \frac{\partial z}{\partial v}\frac{\partial x}{\partial u}\right) + O\left(\frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u}\right)\right) du \wedge dv\end{aligned}$$

or

$$\begin{aligned}\int_S \mathbf{F} \cdot đ\mathbf{S} &= \int_S \langle M, N, O \rangle \cdot \frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v}\, du \wedge dv = \int_S \langle M, N, O \rangle \cdot \left\langle \frac{\partial x}{\partial u}, \frac{\partial y}{\partial u}, \frac{\partial z}{\partial u} \right\rangle \times \left\langle \frac{\partial x}{\partial v}, \frac{\partial y}{\partial v}, \frac{\partial z}{\partial v} \right\rangle du \wedge dv \\ &= \int_S \left(M\left(\frac{\partial y}{\partial u}\frac{\partial z}{\partial v} - \frac{\partial y}{\partial v}\frac{\partial z}{\partial u}\right) + N\left(\frac{\partial z}{\partial u}\frac{\partial x}{\partial v} - \frac{\partial z}{\partial v}\frac{\partial x}{\partial u}\right) + O\left(\frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u}\right)\right) du \wedge dv.\end{aligned}$$

To match pseudoorientations using the right-hand rule, make sure that, as you turn the fingers of your right hand from the direction in which u changes towards the direction in which v changes, your thumb points in the direction of the surface's pseudoorientation; or if not, then put a minus sign out front.

To integrate a scalar quantity f on an unoriented surface S, you integrate the rank-2 absolute form $f\, đ\sigma$, where σ has no meaning by itself but instead $đ\sigma$ is the absolute form $\|đ\mathbf{S}\|$:

$$\begin{aligned}\int_S f\, đ\sigma &= \int_S f\, \|đ\mathbf{S}\| = \int_S f\, \|\langle dy \wedge dz, dz \wedge dx, dx \wedge dy \rangle\| = \int_S f \sqrt{(dy \wedge dz)^2 + (dz \wedge dx)^2 + (dx \wedge dy)^2} \\ &= \int_S f \sqrt{\left(\frac{\partial y}{\partial u}\frac{\partial z}{\partial v} - \frac{\partial y}{\partial v}\frac{\partial z}{\partial u}\right)^2 + \left(\frac{\partial z}{\partial u}\frac{\partial x}{\partial v} - \frac{\partial z}{\partial v}\frac{\partial x}{\partial u}\right)^2 + \left(\frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u}\right)^2}\, |du \wedge dv|\end{aligned}$$

or

$$\begin{aligned}\int_S f\, đ\sigma &= \int_S f\, \|đ\mathbf{S}\| = \int_S f \left\| \frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v} \right\| |du \wedge dv| = \int_S f \left\| \left\langle \frac{\partial x}{\partial u}, \frac{\partial y}{\partial u}, \frac{\partial z}{\partial u} \right\rangle \times \left\langle \frac{\partial x}{\partial v}, \frac{\partial y}{\partial v}, \frac{\partial z}{\partial v} \right\rangle \right\| |du \wedge dv| \\ &= \int_S f \sqrt{\left(\frac{\partial y}{\partial u}\frac{\partial z}{\partial v} - \frac{\partial y}{\partial v}\frac{\partial z}{\partial u}\right)^2 + \left(\frac{\partial z}{\partial u}\frac{\partial x}{\partial v} - \frac{\partial z}{\partial v}\frac{\partial x}{\partial u}\right)^2 + \left(\frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u}\right)^2}\, |du \wedge dv|.\end{aligned}$$

Again there is no orientation to match; instead, make sure that u and v are both increasing, so that $|du \wedge dv| = du\, dv$ in the integral; or if not, then put a minus sign out front for each one that doesn't.

Area integrals

The coordinate plane is both an ambient space of dimension 2 and a manifold of dimension 2 within itself. You can parametrize it simply by the coordinates x and y, although there are other ways to parametrize it (such as by polar coordinates).

Instead of $đ\mathbf{S}$, we can look at the pseudoexterior form $1/2\, d\mathbf{r} \mathbin{\hat{\times}} d\mathbf{r}$, which works out to $dx \wedge dy$ (using the right-hand rule). Alternatively, instead of $đ\sigma$, we can look at the absolute form $|dx \wedge dy|$. These are actually two equivalent ways to think of the area form $đA$, because there is nothing to do to pseudorient a manifold within itself; it's not possible to go around or across the plane while staying within the plane. In the past, we've thought of $đA$ as an absolute form, which means that you didn't have to worry about orientation or the right-hand rule. But when applying the Stokes Theorem later on, you'll have to think of $đA$ as a pseudoexterior form, because the Stokes Theorem doesn't apply to absolute forms in general.

In any case, to integrate a scalar quantity f on a region in the plane, you integrate the rank-2 form $f\, đA$; make sure that x and y are both increasing, so that $|dx \wedge dy| = dx\, dy$ in the integral; or if not, then put a minus sign out front for each one that doesn't.

Volume integrals

Similarly, ordinary three-dimensional space is both an ambient space of dimension 3 and a manifold of dimension 3 within itself. You can parametrize it by the coordinates x, y, and z, although again there are other ways to parametrize it (such as by cylindrical or spherical coordinates).

Instead of $đA$ we can look at the pseudoexterior form $1/6\,\mathrm{d}\mathbf{r} \mathbin{\hat{\cdot}} \mathrm{d}\mathbf{r} \mathbin{\hat{\times}} \mathrm{d}\mathbf{r}$, which works out to $\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z$ (using the right-hand rule), or the absolute form $|\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z|$. Again, these are two equivalent ways to think of the volume form $đV$. In the past, we've thought of $đV$ as an absolute form; but when applying the Stokes Theorem later on, you'll have to think of $đV$ as a pseudoexterior form.

In any case, to integrate a scalar quantity f in a region in space, you integrate the rank-3 form $f\,đV$; make sure that x, y, and z are all increasing, so that $|\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z| = \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$ in the integral; or if not, then put a minus sign out front for each one that doesn't.

The Stokes Theorem

The (second) Fundamental Theorem of Calculus states that

$$\int_a^b \mathrm{d}u = u|_a^b.$$

This works just as well when there are several independent variables as when there is just one. In this case, you can also write $\mathrm{d}\big(f(P)\big)$ as $\nabla f(P) \cdot \mathrm{d}\mathbf{r}$ to get the theorem

$$\int_{P=a}^b \nabla f(P) \cdot \mathrm{d}\mathbf{r} = f(b) - f(a).$$

Although this is now a theorem about integrating a gradient along a curve, in essence it is still just the FTC, a theorem about integrating differentials.

To keep the notation simple, I'll continue to refer to scalar- and vector-valued quantities rather than to scalar and vector fields (which are kinds of functions). The only real imprecision here is that the symbol written '∇' should properly be $\mathrm{d}/\mathrm{d}P$ (or $\mathrm{d}/\mathrm{d}\mathbf{r}$) to indicate the variables with respect to which you're differentiating; however, '∇' is much more common. So for example, I'll write the preceding statement about gradients as

$$\int_a^b \nabla f \cdot \mathrm{d}\mathbf{r} = f|_a^b,$$

where the f here is really the same as what was u before.

This theorem generalizes to differential forms of higher rank, where it is called the **Stokes Theorem**:

$$\int_M \mathrm{d} \wedge \alpha = \int_{\partial M} \alpha.$$

Here, α is any exterior or pseudoexterior differential form and M is any oriented or pseudooriented manifold, so long as they have the same kind of orientation and the dimension of M is 1 more than the rank of α (so that the dimensions and ranks in each integral match up). To do this properly, you need to know two things: how to take the differential of a differential form, which is the $\mathrm{d} \wedge \alpha$ in the Stokes Theorem; and how to take the endpoints of a manifold other than a curve, which is the ∂M in the Stokes Theorem (which traditionally, but unfortunately, uses the same symbol as for partial derivatives).

With endpoints, you're really dealing with the *boundary* of a manifold. The boundary of a curve oriented from a to b consists of both the point $\{a\}$ and the point $\{b\}$, the former negatively and the latter positively. (Technically, this is a chain: the point $\{a\}$ has weight -1, while the point $\{b\}$ has weight 1.) If you think of a point $\{a\}$ as a manifold of dimension 0 and think of a scalar quantity f as a differential form of rank 0, then you integrate f on $\{a\}$ by simply taking the value of f at a: $\int_{\{a\}} f = f|_a$, so $\int_{-1\{a\}+1\{b\}} f = -1f|_a + 1f|_b = f|_a^b$. Then the FTC can be written as

$$\int_C \mathrm{d}f = \int_{\partial C} f.$$

The boundary of a surface is a curve (or a chain made up of several curves), and the boundary of a region of space is a surface (or a chain made up of several surfaces).

When you take the differential of an exterior differential form α, you get another exterior differential form if you use the *exterior* differential $\mathrm{d}\wedge\alpha$ (which is usually written just '$\mathrm{d}\alpha$' by people who study only exterior and pseudoexterior forms, even though there is also an ordinary nonexterior differential that could be used instead). You can also apply this to a pseudoexterior form to get another pseudoexterior form. When you add forms, the exterior differential obeys the Sum Rule as usual; when you multiply them, you have a kind of Product Rule too. This is the same as the usual Product Rule, except that you must keep track of the order of multiplication. However, this caveat really doesn't matter due to the next rule: the exterior differential of a differential is zero. For example,

$$\mathrm{d}\wedge(x\,\mathrm{d}y\wedge\mathrm{d}z)=\mathrm{d}x\wedge\mathrm{d}y\wedge\mathrm{d}z+x\,\mathrm{d}\wedge\mathrm{d}y\wedge\mathrm{d}z-x\,\mathrm{d}y\wedge\mathrm{d}\wedge\mathrm{d}z=\mathrm{d}x\wedge\mathrm{d}y\wedge\mathrm{d}z+0-0=\mathrm{d}x\wedge\mathrm{d}y\wedge\mathrm{d}z.$$

So in the end, you just take the differential of the non-differential factor of each term, then stick this with a wedge in front of the previous differential factors.

When you relate differential forms to vector fields, you can also use various ways of taking derivatives of vector fields. These can be expressed using ∇ and one of the ways of multiplying vectors: the **divergence** $\nabla\cdot\mathbf{F}$ is a scalar field, and the **curl** $\nabla\times\mathbf{F}$ is a pseudovector field in 3 dimensions or a pseudoscalar field in 2 dimensions. Specifically, if $\mathbf{F}(x,y,\ldots)=\langle M,N,\ldots\rangle$, then

$$\nabla\cdot\mathbf{F}(x,y,\ldots)=\langle\partial/\partial x,\partial/\partial y,\ldots\rangle\cdot\langle M,N,\ldots\rangle=\frac{\partial M}{\partial x}+\frac{\partial N}{\partial y}+\cdots;$$

and

$$\nabla\times\mathbf{F}(x,y,z)=\langle\partial/\partial x,\partial/\partial y,\partial/\partial z\rangle\times\langle M,N,O\rangle=\left\langle\frac{\partial O}{\partial y}-\frac{\partial N}{\partial z},\frac{\partial M}{\partial z}-\frac{\partial O}{\partial x},\frac{\partial N}{\partial x}-\frac{\partial M}{\partial y}\right\rangle$$

in 3 dimensions, while

$$\nabla\times\mathbf{F}(x,y)=\langle\partial/\partial x,\partial/\partial y\rangle\times\langle M,N\rangle=\frac{\partial N}{\partial x}-\frac{\partial M}{\partial y}$$

in 2 dimensions.

The connection between these and differentials is as follows (where now I'll conflate the functions f and $\mathbf{F}$ with their values $f(x,y,\ldots)$ and $\mathbf{F}(x,y,\ldots)$ to keep the notation short):

- $\mathrm{d}f=\nabla f\cdot\mathrm{d}\mathbf{r}$ in any number of dimensions;
- $\mathrm{d}\wedge(\mathbf{F}\cdot\mathrm{d}\mathbf{r})=\nabla\times\mathbf{F}\,đA$ in 2 dimensions;
- $\mathrm{d}\wedge(\mathbf{F}\cdot\mathrm{d}\mathbf{r})=\nabla\times\mathbf{F}\cdot đ\mathbf{S}$ in 3 dimensions;
- $\mathrm{d}\wedge(\mathbf{F}\times\mathrm{d}\mathbf{r})=\nabla\cdot\mathbf{F}\,đA$ in 2 dimensions; and
- $\mathrm{d}\wedge(\mathbf{F}\cdot đ\mathbf{S})=\nabla\cdot\mathbf{F}\,đV$ in 3 dimensions.

(These are not new principles, but rather facts that you can verify by writing everything in terms of the components of $\mathbf{F}$, partial derivatives, and differentials.) Here, $đA$ is the area form $|\mathrm{d}x\wedge\mathrm{d}y|$, which you should now think of as a pseudoexterior form that you can identify with $\mathrm{d}x\wedge\mathrm{d}y$ using the right-hand rule, and $đV$ is the volume form $|\mathrm{d}x\wedge\mathrm{d}y\wedge\mathrm{d}z|$, which you should now think of as a pseudoexterior form that you can identify with $\mathrm{d}x\wedge\mathrm{d}y\wedge\mathrm{d}z$ using the right-hand rule.

Now suppose that a surface S is bounded by a curve ∂S. The Stokes Theorem tells you that

$$\int_S\mathrm{d}\wedge\alpha=\int_{\partial S}\alpha,$$

where α is any (exterior or pseudoexterior) differential form of rank 1. If you integrate a vector quantity $\mathbf{F}$ along ∂S, then you're really integrating the differential form $\mathbf{F}\cdot\mathrm{d}\mathbf{r}$, so

$$\int_{\partial S}\mathbf{F}\cdot\mathrm{d}\mathbf{r}=\int_S\mathrm{d}\wedge(\mathbf{F}\cdot\mathrm{d}\mathbf{r})=\int_S\nabla\times\mathbf{F}\cdot đ\mathbf{S}$$

in 3 dimensions, or

$$\int_{\partial S}\mathbf{F}\cdot\mathrm{d}\mathbf{r}=\int_S\mathrm{d}\wedge(\mathbf{F}\cdot\mathrm{d}\mathbf{r})=\int_S\nabla\times\mathbf{F}\,đA$$

in 2 dimensions (where S is now a region in the plane). These are the theorems traditionally called *Stokes's Theorem* and *Green's Theorem*, respectively. If, in 2 dimensions, you integrate $\mathbf{F}$ across ∂S, then

$$\int_{\partial S} \mathbf{F} \times \mathrm{d}\mathbf{r} = \int_S \mathrm{d} \wedge (\mathbf{F} \times \mathrm{d}\mathbf{r}) = \int_S \nabla \cdot \mathbf{F} \, đA,$$

which is another form of Green's Theorem; in terms of differentials, it's just like the previous version, except that the form being integrated is pseudoexterior instead of exterior. (These theorems are not new principles either, but follow from the general Stokes Theorem and the exterior differentials listed above.)

Next, suppose that a region Q in space is bounded by a surface ∂Q. Now the Stokes Theorem tells you that

$$\int_Q \mathrm{d} \wedge \alpha = \int_{\partial Q} \alpha,$$

where now α is any (exterior or pseudoexterior) differential form of rank 2. If you integrate a vector field $\mathbf{F}$ across ∂Q, then you're really integrating $\mathbf{F} \cdot đ\mathbf{S}$, so

$$\int_{\partial Q} \mathbf{F} \cdot đ\mathbf{S} = \int_Q \mathrm{d} \wedge (\mathbf{F} \cdot đ\mathbf{S}) = \int_Q \nabla \cdot \mathbf{F} \, đV.$$

This is the theorem traditionally called *Gauss's Theorem*, although many textbooks simply call it the *Divergence Theorem*. (Once more, you can verify these by explicit calculation.)

Since the boundary ∂M for any manifold is closed in on itself, the boundary of the boundary, $\partial\partial M$, is always empty. This means that

$$\int_M \mathrm{d} \wedge \mathrm{d} \wedge \alpha = \int_{\partial M} \mathrm{d} \wedge \alpha = \int_{\partial\partial M} \alpha = 0;$$

since this is true no matter how small M may be, you can conclude that

$$\mathrm{d} \wedge \mathrm{d} \wedge \alpha = 0$$

for any (exterior or pseudoexterior) differential form α. In terms of vector fields, this has two consequences:

$$\nabla \times \nabla f = 0$$

in 2 or 3 dimensions, and

$$\nabla \cdot \nabla \times \mathbf{F} = 0$$

in 3 dimensions. If you write these facts out using partial derivatives, then you'll see that they simply state the equality of mixed partial derivatives. (As a technicality, that equality is not always guaranteed, but it is guaranteed when the mixed partial derivatives are continuous; we derived these facts by considering integrals that likewise are only guaranteed to exist when the forms being integrated are continuous. Conversely, the Stokes Theorem can be proved in the first place by using the equality of mixed partial derivatives and the ordinary FTC applied to iterated integrals, by carefully keeping track of everything.)

Optional material

This material doesn't come up in the course, but it's used a lot and fills in some gaps in the concepts.

Hodge duals

You may notice that a vector quantity $\mathbf{F}$ can be turned into a differential form in two different ways: in 2 dimensions, $\mathbf{F} \cdot \mathrm{d}\mathbf{r}$ is an exterior form of rank 1, while $\mathbf{F} \times \mathrm{d}\mathbf{r}$ is a pseudoexterior form of rank 1; in 3 dimensions, $\mathbf{F} \cdot \mathrm{d}\mathbf{r}$ is again an exterior form of rank 1, while now $\mathbf{F} \cdot đ\mathbf{S}$ is a pseudoexterior form of rank 2. Either way, the two differential forms related to a single vector field are called *Hodge duals* of each other. If you work directly with differential forms instead of vectors, then you can use the Hodge duals to bring in geometric ideas of length and angle. In this way, you can work as much as possible with the objects that you integrate to get measurable quantities.

The Hodge dual of a differential form α is denoted $*\alpha$. In rectangular coordinates, it's easy to calculate Hodge duals; you replace the differential factors of each term with whatever is missing in the area or volume form (written in the order given by the right-hand rule), paying attention to the sign. This gives you

$$*\mathrm{d}x = \mathrm{d}y,\ *\mathrm{d}y = -\mathrm{d}x$$

in 2 dimensions; and

$$*\mathrm{d}x = \mathrm{d}y \wedge \mathrm{d}z,\ *\mathrm{d}y = -\mathrm{d}x \wedge \mathrm{d}z = \mathrm{d}z \wedge \mathrm{d}x,\ *\mathrm{d}z = \mathrm{d}x \wedge \mathrm{d}y$$

and

$$*(\mathrm{d}y \wedge \mathrm{d}z) = \mathrm{d}x,\ *(\mathrm{d}z \wedge \mathrm{d}x) = \mathrm{d}y,\ *(\mathrm{d}x \wedge \mathrm{d}y) = \mathrm{d}z$$

in 3 dimensions. (The Hodge dual of an exterior form is a pseudoexterior form and vice versa, and these rules are written using the right-hand rule.) Now you can check that

$$*(\mathbf{F} \cdot \mathrm{d}\mathbf{r}) = \mathbf{F} \times \mathrm{d}\mathbf{r},\ *(\mathbf{F} \times \mathrm{d}\mathbf{r}) = -\mathbf{F} \cdot \mathrm{d}\mathbf{r},$$

in 2 dimensions; and

$$*(\mathbf{F} \cdot \mathrm{d}\mathbf{r}) = \mathbf{F} \cdot đ\mathbf{S},\ *(\mathbf{F} \cdot đ\mathbf{S}) = \mathbf{F} \cdot \mathrm{d}\mathbf{r}$$

in 3 dimensions. You can even extend this to forms of top rank and to scalar quantities (which are differential forms of rank 0):

$$*(đA) = *(\mathrm{d}x \wedge \mathrm{d}y) = 1,\ *1 = \mathrm{d}x \wedge \mathrm{d}y = đA$$

in 2 dimensions; and

$$*(đV) = *(\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z) = 1,\ *1 = \mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z = đV$$

in 3 dimensions.

Laplacians

The **Laplacian** of a form α is

$$\Delta\alpha = *\Big(\mathrm{d} \wedge *(\mathrm{d} \wedge \alpha)\Big) \pm \mathrm{d} \wedge *(\mathrm{d} \wedge *\alpha),$$

where you use $+$ or $-$ on the second term depending on whether the ambient space has even or odd dimension, and you must throw in another overall minus sign if both the space's dimension and the form's rank are odd. In other words, take the exterior differential, then the Hodge dual, then repeat; and also do this in reverse order; then add or subtract these according to the parity of the dimension, and possibly take the opposite of the entire result. (I know, that's kind of a complicated rule; it's been chosen just so to make everything below work out.) Notice that $\Delta\alpha$ has both the same rank and the same orientation as α, so it is a nice notion of second derivative.

If you think of a scalar field f as an exterior form of rank 0, then $\mathrm{d} \wedge f = \mathrm{d}f$, while $*f$ has top rank, so $\mathrm{d} \wedge *f = 0$. Then

$$\Delta f = *(\mathrm{d} \wedge *\mathrm{d}f) = *\Big(\mathrm{d} \wedge *(\nabla f \cdot \mathrm{d}\mathbf{r})\Big) = *\Big(\mathrm{d} \wedge (\nabla f \times \mathrm{d}\mathbf{r})\Big) = *(\nabla \cdot \nabla f\, đA) = \nabla \cdot \nabla f$$

in 2 dimensions; and

$$\Delta f = *(\mathrm{d} \wedge *\mathrm{d}f) = *\Big(\mathrm{d} \wedge *(\nabla f \cdot \mathrm{d}\mathbf{r})\Big) = *\Big(\mathrm{d} \wedge (\nabla f \cdot đ\mathbf{S})\Big) = *(\nabla \cdot \nabla f \, đV) = \nabla \cdot \nabla f$$

in 3 dimensions. In fact, the rule that $\Delta f = \nabla \cdot \nabla f$ is correct in *any* number of dimensions (and the weird rules about minus signs are designed to make that work out); for this reason, the Laplacian operator Δ is often written as '$\|\nabla\|^2$' or just '∇^2' (think of $\|\mathbf{v}\|^2 = \mathbf{v} \cdot \mathbf{v}$).

Other Laplacians are

$$\Delta(\mathbf{F} \cdot \mathrm{d}\mathbf{r}) = \nabla(\nabla \cdot \mathbf{F}) \cdot \mathrm{d}\mathbf{r} + \nabla(\nabla \times \mathbf{F}) \times \mathrm{d}\mathbf{r}, \; \Delta(\mathbf{F} \times \mathrm{d}\mathbf{r}) = \nabla(\nabla \cdot \mathbf{F}) \times \mathrm{d}\mathbf{r} - \nabla(\nabla \times \mathbf{F}) \cdot \mathrm{d}\mathbf{r}$$

in 2 dimensions; and

$$\Delta(\mathbf{F} \cdot \mathrm{d}\mathbf{r}) = \nabla(\nabla \cdot \mathbf{F}) \cdot \mathrm{d}\mathbf{r} - \nabla \times (\nabla \times \mathbf{F}) \cdot \mathrm{d}\mathbf{r}, \; \Delta(\mathbf{F} \cdot đ\mathbf{S}) = \nabla(\nabla \cdot \mathbf{F}) \cdot đ\mathbf{S} - \nabla \times (\nabla \times \mathbf{F}) \cdot đ\mathbf{S}$$

in 3 dimensions. If you define $\Delta\mathbf{F}$ so that $\Delta\mathbf{F} \cdot \mathrm{d}\mathbf{r} = \Delta(\mathbf{F} \cdot \mathrm{d}\mathbf{r})$, you can see (by working out their components) that $\Delta\mathbf{F} \times \mathrm{d}\mathbf{r} = \Delta(\mathbf{F} \times \mathrm{d}\mathbf{r})$ in 2 dimensions and that $\Delta\mathbf{F} \cdot đ\mathbf{S} = \Delta(\mathbf{F} \cdot đ\mathbf{S})$ in 3 dimensions; furthermore, each component of $\Delta\mathbf{F}$ is the Laplacian of the corresponding component of $\mathbf{F}$. So Laplacians work very nicely indeed.

Maxwell's equations

One of the basic applications of vector calculus —arguably the original application— is the classical theory of electromagnetic fields that was fully worked almost 150 years ago by James Clerk Maxwell. Maxwell's equations of electromagnetism have been expressed in many formalisms over the years: explicitly using partial derivatives of component functions (the way Maxwell presented them), using quaternions (like complex numbers with three imaginary dimensions, which is how Maxwell really thought of them), using the vector calculus of Oliver Heaviside and Willard Gibbs (the simplification of quaternionic calculus that is taught in the course textbook), using differential forms in three-dimensional space (which is how I usually think of them), and using differential forms in four-dimensional space-time. Each is simpler and more elegant than the last.

Nearly all of the differential forms appearing in these notes will be exterior or pseudoexterior differential forms. To keep the notation simple, *I will leave out the symbol '$\wedge$' in the wedge product and the exterior differential.* So unless I explicitly state otherwise, if you see two differentials (or differential forms) multiplied together, then they're being multiplied by the wedge product (aka the exterior product); and if you see the differential of a differential form, then it's the exterior differential. (People who work with exterior differential forms usually do this anyway, especially for the differential. Note that the exterior differential of a nondifferential expression is the same as its ordinary differential, so there is no confusion there.) Also, unlike in the specific problems that you've done in this course, I'll use variables that refer directly to differential forms; typically, these variables with be in a fancy calligraphic font ('$\mathcal{A}$', '$\mathcal{B}$', '$\mathcal{C}$', ...).

The quantities in the equations

To be very definite, I will give operational definitions of the physical quantities that appear in Maxwell's equations, describing how you would (in principle) measure them.

I will take as a basic notion the idea of **electric charge**. Electric charge may be positive or negative, and the difference between these is perfectly arbitrary (which is in some ways similar to the right-hand rule); what's important is that there is a difference, and positive and negative charges cancel each other out. In any given region of space, there is a certain total charge in that region, which we'll assume is given by integrating a continuous rank-3 pseudoexterior differential form, the **charge form** $\mathcal{Q}$. (The existence of this differential form is actually a theorem, under certain assumptions about additivity and continuity of charge.) We may write

$$\mathcal{Q} = \rho \, đV = \rho \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z,$$

where the scalar field ρ is the **charge density**. The SI unit of charge is the coulomb (named after Charles-Augustin de Coulomb, who discovered the inverse-square law of static electricity); charge density is measured in coulombs per cubic metre.

Together with electric charge, we have **electric current**, which is the flow of electric charge. We measure current through a pseudooriented surface; the total rate (with respect to time) at which charge moves through the surface in the given direction is the current through that surface. (Negative charge moving through the surface in the negative direction counts positively, like positive charge moving in the positive direction; negative charge moving in the positive direction and positive charge moving in the negative direction count negatively.) The current through a pseudooriented surface is given by integrating a continuous rank-2 pseudoexterior form, the **current form** $\mathcal{J}$. We may write

$$\mathcal{J} = \mathbf{J} \cdot đ\mathbf{S} = J_1 \, \mathrm{d}y \, \mathrm{d}z + J_2 \, \mathrm{d}z \, \mathrm{d}x + J_3 \, \mathrm{d}x \, \mathrm{d}y,$$

where the vector field $\mathbf{J}$ is the **current density**. The SI unit of current is the coulomb per second, or ampere (named after André-Marie Ampère, who discovered Ampere's Law, discussed below); current density is measured in amperes per square metre.

Based on these, we can now define some other quantities. When the work (transfer of energy) done on a charged object is proportional to its charge, we consider that the work is done by an *electric field*. If a charged object travels through an electric field along an oriented curve, then the work done on the particle is the product of the particle's charge and the **electric potential** along the curve. Since the charge on any actual object is spread out over space and charge (as you'll see below) also affects the electric field,

we really need to consider the limiting case of an object with both infinitesimal volume and infinitesimal charge density. The electric potential along an oriented curve is given by integrating a continuous rank-1 exterior form, the **electric potential form** $\mathcal{E}$. We may write

$$\mathcal{E} = \mathbf{E} \cdot \mathrm{d}\mathbf{r} = E_1 \,\mathrm{d}x + E_2 \,\mathrm{d}y + E_3 \,\mathrm{d}z,$$

where the vector field $\mathbf{E}$ is the **electric field strength**. The SI unit of electric potential is the joule per coulomb, or volt; electric field strength is measured in volts per metre.

The electric field not only affects charges but also is created by them. As charges move in response to the work done on them by the electric field, this tends to cancel out the original field. (This is a general theme in electromagnetism, that any phenomenon has effects that counteract the original cause.) In particular, if a sheet of material that conducts electric current (a *Faraday shield*) is placed in an electric field, then the free charged particles in the shield will move to opposite sides, blocking out the electric field in the interior of the sheet. The **electric flux** through a pseudooriented surface is the total charge induced by the electric field on the outside of a continuous Faraday shield along that surface (or opposite the charge induced on the inside of the shield). Again, we must really consider a limiting case, that of a sheet with infinitesimal thickness and infinite conductance. The electric flux through a pseudooriented surface is given by integrating a continuous rank-2 pseudoexterior form, the **electric flux form** $\mathcal{D}$. We may write

$$\mathcal{D} = \mathbf{D} \cdot \text{đ}\mathbf{S} = D_1 \,\mathrm{d}y \,\mathrm{d}z + D_2 \,\mathrm{d}z \,\mathrm{d}x + D_3 \,\mathrm{d}x \,\mathrm{d}y,$$

where the vector field $\mathbf{D}$ is the **electric displacement**. The SI unit of electric flux is the coulomb again; electric displacement is measured in coulombs per square metre.

Besides the electric field, there is also a *magnetic field*. Although this may be thought of as dealing with magnetic poles (instead of electric charges), magnetic poles are not individual objects but always come in pairs. We now understand (and Maxwell already understood) that magnetism deals with electric currents, with a north pole and a south pole appearing on either side of a rotating current. If a wire with current flowing through it travels through a magnetic field, then it traces out a surface, which we orient (not pseudoorient!) as the direction of travel followed by the direction of the current. Then the work done on the wire is the product of the wire's current and the **magnetic flux** on the surface. Since any actual conducting wire has some thickness and current (as you'll see below) also affects the magnetic field, we really need to consider the limiting case of a wire with both infinitesimal thickness and infinitesimal current density. The magnetic flux on an oriented surface is given by integrating a continuous rank-2 exterior form, the **magnetic flux form** $\mathcal{B}$. We may write

$$\mathcal{B} = \mathbf{B} \cdot \text{đ}\mathbf{S} = B_1 \,\mathrm{d}y \,\mathrm{d}z + B_2 \,\mathrm{d}z \,\mathrm{d}x + B_3 \,\mathrm{d}x \,\mathrm{d}y,$$

where the pseudovector field $\mathbf{B}$ is the **magnetic flux density**. The SI unit of magnetic flux is the joule per ampere, or weber; magnetic flux density is measured in webers per square metre, or teslas.

Just as the electric field causes charges to move to counteract it, so the magnetic field creates currents that counteract it. In particular, if a tube of conductive material (a *solenoid*) is placed in a magnetic field, then the field will induce a current on the inside of the solenoid, blocking the magnetic field within the solenoid. The **magnetic potential** around a pseudooriented curve (not oriented!) is the total current induced by the magnetic field in a continuous solenoid surrounding the curve in the direction opposite the curve's pseudoorientation. Once more, we must really consider a limiting case, that of a tube with infinitesimal radius and infinite conductance. The magnetic potential around a pseudooriented curve is given by integrating a continuous rank-1 pseudoexterior form, the **magnetic potential form** $\mathcal{H}$. We may write

$$\mathcal{H} = \mathbf{H} \cdot \mathrm{d}\mathbf{r} = H_1 \,\mathrm{d}x + H_2 \,\mathrm{d}y + H_3 \,\mathrm{d}z,$$

where the pseudovector field $\mathbf{H}$ is the **magnetizing field strength**. The SI unit of magnetic potential is the ampere again; magnetizing field strength is measured in amperes per metre.

The constitutive relations

Before I get to the four equations generally called Maxwell's, I need to clear something up. We have two ways to measure an electric field, the electric potential along a curve (the integral of $\mathcal{E}$) and the electric flux through a surface (the integral of $\mathcal{D}$); similarly, we have two ways to measure a magnetic field, the magnetic flux on a surface (the integral of $\mathcal{B}$) and the magnetic potential around a curve (the integral of $\mathcal{H}$). Since $\mathcal{E}$ and $\mathcal{D}$ measure the same physical field, there should be a relationship between them, and the same for $\mathcal{B}$ and $\mathcal{H}$. The simplest relationship would be that each of these quantities is the Hodge dual of its partner; after all, the Hodge dual of an exterior 1-form is a pseudoexterior 2-form, etc. (Then we would also have $\mathbf{E} = \mathbf{D}$ and $\mathbf{B} = \mathbf{H}$.) However, there are a few complications with that.

First, if we measure $\mathcal{D}$ and $\mathcal{H}$ with actual conducting materials, then (even in the limit of infinite conductance!) there will always be charges that are bound in the material, unable to be moved by the fields, and there will also be bound currents sometimes (as in a magnet). Thus, $\mathcal{D}$ and $\mathcal{H}$ effectively measure only the free charge and current. When people express Maxwell's equations using only $\mathcal{E}$ and $\mathcal{B}$ instead, they speak of Maxwell's equations *in a vacuum*.

Secondly, even in vacuum, $\mathcal{E}$ and $\mathcal{D}$ are measured in different units (and similarly for $\mathcal{B}$ and $\mathcal{H}$). Up to a point, this is expected; since volume has units of cubic metres, we expect the Hodge dual to affect units. However, this only affects units of length, and we need more than that (in particular, the units of charge are reversed). In vaccum, the unit conversion is done by fundamental physical constants, the *electric constant* ϵ_0 and the *magnetic constant* μ_0; then we have

$$*\mathcal{E} = \frac{\mathcal{D}}{\epsilon_0}$$

(so $*\mathcal{D} = \epsilon_0 \mathcal{E}$) and

$$*\mathcal{B} = \mu_0 \mathcal{H}$$

(so $*\mathcal{H} = \mathcal{B}/\mu_0$). Ultimately, the SI units are defined so that ϵ_0 is exactly

$$\frac{2^3 5^7}{7^2 73^2 293339^2 \pi} \approx 8.85 \times 10^{-12}$$

farads per metre and μ_0 is exactly

$$\frac{\pi}{2^5 5^7} \approx 1.26 \times 10^{-6}$$

henries per metre. (A farad is a square coulomb per joule, named after Michael Faraday, who discoved Faraday's Law, below; a henry is a joule per square ampere. By the way, there are only two more SI units related specifically to electromagnetism: the siemens is a farad per second, and the ohm is a henry per second. But we will not need these here.)

In a medium, we typically have $*\mathcal{E} = \mathcal{D}/\epsilon$ and $*\mathcal{B} = \mu\mathcal{H}$ (or $\mathbf{D} = \epsilon\mathbf{E}$ and $\mathbf{H} = \mathbf{B}/\mu$ in terms of vector fields) for some constants ϵ and μ, the *permittivity* and *permeability* of the medium. (Then ϵ_0 and μ_0 are respectively the permittivity and permeability of the vacuum.) Sometimes things are not so simple (for example, the permittivity or permeability may depend on the direction); but we always have some relationship between these quantities, called the *constitutive relations* of the material. When we use differential forms instead of vector fields, the constitutive relations are the *only* equations in which the Hodge dual operator appears, hence the only place where geometric ideas (such as length, angle, and volume) play a role; using vector fields obscures this fact.

Static systems

Maxwell found four equations, which I will state first for *static* systems, that is those in which the distribution of charges, currents, and fields does not change with time. In a static system, the total current through the boundary of any region of space must be zero, because otherwise the total charge inside that region would be changing; this is the *continuity equation*

$$\int_{\partial Q} \mathcal{J} = 0,$$

which is not counted as one of Maxwell's four. Assuming that $\mathcal{J}$ is continuously differentiable, then the Stokes Theorem turns this into $\int_Q \mathrm{d}\mathcal{J} = 0$; since this holds for any region Q, we conclude that

$$\mathrm{d}\mathcal{J} = 0,$$

which is $\nabla \cdot \mathbf{J} = 0$ in terms of the current density. Like the continuity equation, each of Maxwell's equations will have an integral and differential form.

The simplest of Maxwell's equations is

$$\int_{\partial Q} \mathcal{B} = 0,$$

stating that the magnetic flux through the boundary of any region in space is zero. In other words, magnetic flux, like current in a static system, flows continuously with no sink or source. The differential form is

$$\mathrm{d}\mathcal{B} = 0,$$

or $\nabla \cdot \mathbf{B} = 0$ in vector calculus.

Similarly, *Faraday's Law* for static systems states that the electric potential along the boundary of any oriented surface is zero:

$$\int_{\partial R} \mathcal{E} = 0.$$

In differential form, this becomes

$$\mathrm{d}\mathcal{E} = 0,$$

which is $\nabla \times \mathbf{E} = 0$ in vector calulus. Thus, $\mathcal{E}$ is an exact differential, and $\mathbf{E}$ is a conservative vector field.

Next, Gauss's Law (after Carl Gauß) states that the total electric flux outward through the boundary of any region in space equals the total electric charge contained in that region:

$$\int_{\partial Q} \mathcal{D} = \int_Q \mathcal{Q}.$$

In differential form,

$$\mathrm{d}\mathcal{D} = \mathcal{Q};$$

in vector calculus, $\nabla \cdot \mathbf{D} = \rho$. Thus, unlike magnetic flux, electric flux has sources and sinks, which are electric charges.

Finally, Ampere's Law for static systems states that the magnetic potential around the boundary of a pseudooriented surface equals the total current through that surface:

$$\int_{\partial R} \mathcal{H} = \int_R \mathcal{J}.$$

In differential form,

$$\mathrm{d}\mathcal{H} = \mathcal{J};$$

in vector calculus, $\nabla \times \mathbf{H} = \mathbf{J}$. Thus, currents are sources for the magnetic field.

The reason that the continuity equation is not counted as one of Maxwell's equations is that it actually follows from Ampere's Law. Specifically (in a static system), we have

$$\int_{\partial Q} \mathcal{J} = \int_{\partial \partial Q} \mathcal{H} = 0,$$

since the boundary of a boundary is empty.

Electrodynamics

Some of the equations above only apply when the charges, currents, and fields don't change with time. Maxwell's equations also come in a more general form that drops this assumption. It is easy enough to state the integral forms of these equations, but the differential forms require taking seriously the four-dimensional nature of our universe in space and time. In vector calculus, this is done by treating space and time separately, but differential forms make sense in any number of dimensions; this ultimately simplifies Maxwell's equations. Finally, the constitutive relations in 4 dimensions clarify the nature of the geometry of spacetime in our universe, which leads naturally to Albert Einstein's special theory of relativity.

Here are Maxwell's equations in integral form:

$$\begin{aligned}
\int_{\partial Q} \mathcal{B} &= 0, \\
\int_{\partial R} \mathcal{E} &= -\frac{\mathrm{d}}{\mathrm{d}t} \int_R \mathcal{B}, \\
\int_{\partial Q} \mathcal{D} &= \int_Q \mathcal{Q}, \\
\int_{\partial R} \mathcal{H} &= \int_R \mathcal{J} + \frac{\mathrm{d}}{\mathrm{d}t} \int_R \mathcal{D}.
\end{aligned}$$

In words, the magnetic flux on the boundary of an oriented region of space is still zero, but the electric potential along the boundary of an oriented surface is now the opposite of the rate of change with time of the magnetic flux on that surface. Similarly, the electric flux out of the boundary of a region of space is still the total electric charge in that region, but the magnetic potential around the boundary of a pseudooriented surface is now the sum of the electric current through that surface and the rate of change with time of the electric flux through that surface. The continuity equation (which now relies on both Ampere's Law and Gauss's Law) becomes

$$\int_{\partial Q} \mathcal{J} = \int_{\partial \partial Q} \mathcal{H} - \frac{\mathrm{d}}{\mathrm{d}t} \int_{\partial Q} \mathcal{D} = -\frac{\mathrm{d}}{\mathrm{d}t} \int_Q \mathcal{Q};$$

in words, if current flows out of the boundary of a region of space, then the total charge in that region goes down accordingly. (The reason that we credit these equations to Maxwell, when all of them are laws discovered earlier by other people, is that Ampère didn't know about the contribution of $\mathcal{D}$ to his law; Maxwell realized that it had to be there to get the correct continuity equation, and this is what made the system complete.)

If we separate space from time, writing ∂ for the exterior differential on space (holding time t constant, so giving a merely *partial* exterior differential) and using a dot to indicate differentiation with respect to time, then here are the equations in differential form:

$$\begin{aligned}
\partial \mathcal{B} &= 0, \\
\partial \mathcal{E} &= -\dot{\mathcal{B}}, \\
\partial \mathcal{D} &= \mathcal{Q}, \\
\partial \mathcal{H} &= \mathcal{J} + \dot{\mathcal{D}}.
\end{aligned}$$

The continuity equation in differential form is

$$\partial \mathcal{J} = -\dot{\mathcal{Q}}.$$

Rewriting in vector calculus (which is how you usually find Maxwell's equations on T-shirts):

$$\begin{aligned}
\nabla \cdot \mathbf{B} &= 0, \\
\nabla \times \mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t}, \\
\nabla \cdot \mathbf{D} &= \rho, \\
\nabla \times \mathbf{H} &= \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t};
\end{aligned}$$

the continuity equation is

$$\nabla \cdot \mathbf{J} = -\frac{\partial \rho}{\partial t}.$$

This is a little unsatisfying, because differential forms are supposed to take care of *all* variation of a quantity, which in this context is variation in both space and time. In general, we have $\mathrm{d}\omega = \partial\omega + \dot{\omega}\,\mathrm{d}t$, for ω any differential form defined on spacetime. Then $\mathrm{d}(\omega\,\mathrm{d}t) = \mathrm{d}\omega\,\mathrm{d}t = (\partial\omega + \dot{\omega}\,\mathrm{d}t)\,\mathrm{d}t = \partial\omega\,\mathrm{d}t + 0 = \partial\omega\,\mathrm{d}t$ (since $\mathrm{d}t\,\mathrm{d}t = 0$ with the wedge product). This works for $\mathcal{E}$, $\mathcal{H}$, and $\mathcal{J}$, since $\dot{\mathcal{E}}$, $\dot{\mathcal{H}}$, and $\dot{\mathcal{J}}$ never appear. In fact, it works out very nicely to multiply Faraday's Law and Ampere's Law by $\mathrm{d}t$. If we then add or subtract these equations from the ones that precede them, then we can make $\mathrm{d}\mathcal{B}$ and $\mathrm{d}\mathcal{D}$ appear as well. That is, the first pair adds as follows:

$$\begin{aligned}
\partial\mathcal{B} + \partial\mathcal{E}\,\mathrm{d}t &= 0 - \dot{\mathcal{B}}\,\mathrm{d}t, \\
\partial\mathcal{B} + \dot{\mathcal{B}}\,\mathrm{d}t + \partial\mathcal{E}\,\mathrm{d}t &= 0, \\
\mathrm{d}\mathcal{B} + \mathrm{d}(\mathcal{E}\,\mathrm{d}t) &= 0, \\
\mathrm{d}F &= 0.
\end{aligned}$$

In the last step, I've introduced

$$F = \mathcal{B} + \mathcal{E}\,\mathrm{d}t,$$

sometimes called the **Faraday form** (although the letter originally simply stood for 'field'). Similarly, the second pair subtracts as follows:

$$\begin{aligned}
\partial\mathcal{D} - \partial\mathcal{H}\,\mathrm{d}t &= \mathcal{Q} - \mathcal{J}\,\mathrm{d}t - \dot{\mathcal{D}}\,\mathrm{d}t, \\
\partial\mathcal{D} + \dot{\mathcal{D}}\,\mathrm{d}t - \partial\mathcal{H}\,\mathrm{d}t &= \mathcal{Q} - \mathcal{J}\,\mathrm{d}t, \\
\mathrm{d}\mathcal{D} - \mathrm{d}(\mathcal{H}\,\mathrm{d}t) &= \mathcal{Q} - \mathcal{J}\,\mathrm{d}t, \\
\mathrm{d}M &= j.
\end{aligned}$$

Now in the last step, I've introduced both the **Maxwell form**

$$M = \mathcal{D} - \mathcal{H}\,\mathrm{d}t$$

and the **four-current form**

$$j = \mathcal{Q} - \mathcal{J}\,\mathrm{d}t.$$

Let's take stock of where we are. We have a continuous rank-2 exterior differential form F, measured in webers (which are the same as volt-seconds), a continuous rank-2 pseudoexterior differential form M, measured in coulombs (which are the same as ampere-seconds), and a continuous rank-3 pseudoexterior differential form j, also measured in coulombs. There are now only two Maxwell's equations:

$$\begin{aligned}
\mathrm{d}F &= 0, \\
\mathrm{d}M &= j;
\end{aligned}$$

the continuity equation is simply

$$\mathrm{d}j = 0.$$

We can also write these equations in integral form:

$$\begin{aligned}
\int_{\partial R} F &= 0, \\
\int_{\partial R} M &= j; \\
\int_{\partial Q} j &= 0.
\end{aligned}$$

Here, R is a 2-dimensional surface embedded in four-dimensional spacetime, which could be a surface as we normally think of it, for an instant, but is typically what we would think of as a curve, persisting through time (and perhaps moving, growing, or shrinking). Similarly, Q is a 3-dimensional hypersurface in spacetime, which could be a region of space for an instant but is typically what we would think of as a surface, again persisting and possibly changing through time. There is no vector-calculus form of these spacetime equations; neither F nor M can be described by vectors, even ones with 4 components (although there is a concept of bivector or antisymmetrized dyad, a kind of tensor, that could be used here if you really insist).

It's worth looking specifically at the components that would go into F, M, and j. We have

$$F = \mathcal{B} + \mathcal{E}\,\mathrm{d}t = \mathbf{B}\cdot\eth\mathbf{S} + \mathbf{E}\cdot\mathrm{d}\mathbf{r}\,\mathrm{d}t = B_1\,\mathrm{d}y\,\mathrm{d}z + B_2\,\mathrm{d}z\,\mathrm{d}x + B_3\,\mathrm{d}x\,\mathrm{d}y + E_1\,\mathrm{d}x\,\mathrm{d}t + E_2\,\mathrm{d}y\,\mathrm{d}t + E_3\,\mathrm{d}z\,\mathrm{d}t;$$

this has 6 coefficients, containing all of the information in both $\mathcal{E}$ and $\mathcal{B}$ (so nothing is lost by combining the two equations into one). Similarly,

$$M = \mathcal{D} - \mathcal{H}\,\mathrm{d}t = \mathbf{D}\cdot\eth\mathbf{S} - \mathbf{H}\cdot\mathrm{d}\mathbf{r}\,\mathrm{d}t = D_1\,\mathrm{d}y\,\mathrm{d}z + D_2\,\mathrm{d}z\,\mathrm{d}x + D_3\,\mathrm{d}x\,\mathrm{d}y - H_1\,\mathrm{d}x\,\mathrm{d}t - H_2\,\mathrm{d}y\,\mathrm{d}t - H_3\,\mathrm{d}z\,\mathrm{d}t,$$

and

$$j = \mathcal{Q} - \mathcal{J}\,\mathrm{d}t = \rho\,\eth V - \mathbf{J}\cdot\eth\mathbf{S}\,\mathrm{d}t = \rho\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z - J_1\,\mathrm{d}y\,\mathrm{d}z\,\mathrm{d}t - J_2\,\mathrm{d}z\,\mathrm{d}x\,\mathrm{d}t - J_3\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t.$$

(The information in the four-current form can be put into a four-dimensional vector, but I won't bother, since everything works already with forms.)

Special relativity

We have not dealt with the constitutive relations in four dimensions. That is, what is the relationship between F and M? (To keep things simple, work in a vacuum, with ϵ_0 and μ_0.) At the level of components, we already know that $E_i = D_i/\epsilon_0$ and $B_i = \mu_0 H_i$. I'd like to say that $*F$ is a constant times M, but how does the Hodge dual work in spacetime? This is not an easy question, but the great thing about Maxwell's equations is that they tell us how it must work!

We can make our jobs a little easier by using (instead of SI units) units of measurement in which ϵ_0 and μ_0 are 1. Then we have $E_i = D_i$ and $B_i = H_i$ exactly, and we also should have $*F = M$ exactly. This immediately gives us these rules:

$$\begin{aligned}
*(\mathrm{d}x\,\mathrm{d}t) &= \mathrm{d}y\,\mathrm{d}z,\\
*(\mathrm{d}y\,\mathrm{d}t) &= \mathrm{d}z\,\mathrm{d}x,\\
*(\mathrm{d}z\,\mathrm{d}t) &= \mathrm{d}x\,\mathrm{d}y,\\
*(\mathrm{d}y\,\mathrm{d}z) &= -\mathrm{d}x\,\mathrm{d}t,\\
*(\mathrm{d}z\,\mathrm{d}x) &= -\mathrm{d}y\,\mathrm{d}t,\\
*(\mathrm{d}x\,\mathrm{d}y) &= -\mathrm{d}z\,\mathrm{d}t.
\end{aligned}$$

If you try to make a consistent mnemonic for these rules along the lines of the mnemonic that I gave for the Hodge dual in space (where the Hodge dual is whatever is left afterwards in the volume form), then you will fail if you try it directly; in particular, the first rule suggests $\mathrm{d}x\,\mathrm{d}t\,\mathrm{d}y\,\mathrm{d}z$, but this equals $\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z\,\mathrm{d}t$ (two reversals), so there's no explanation for the minus sign in the last rule.

However, we can make it work if we use imaginary numbers! I will put things back in SI units just for the sake of giving the full answer; if you think of the volume form as

$$\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z\,\mathrm{d}(\mathrm{i}ct),$$

where $c = 1/\sqrt{\epsilon_0\mu_0}$, then we get these specific rules:

$$\begin{aligned}
*(\mathrm{d}x\,\mathrm{d}t) &= *\Big(-\frac{\mathrm{i}}{c}\,\mathrm{d}x\,\mathrm{d}(\mathrm{i}ct)\Big) = -\frac{\mathrm{i}}{c}\,\mathrm{d}y\,\mathrm{d}z,\\
*(\mathrm{d}y\,\mathrm{d}t) &= *\Big(-\frac{\mathrm{i}}{c}\,\mathrm{d}y\,\mathrm{d}(\mathrm{i}ct)\Big) = -\frac{\mathrm{i}}{c}\,\mathrm{d}z\,\mathrm{d}x,\\
*(\mathrm{d}z\,\mathrm{d}t) &= *\Big(-\frac{\mathrm{i}}{c}\,\mathrm{d}z\,\mathrm{d}(\mathrm{i}ct)\Big) = -\frac{\mathrm{i}}{c}\,\mathrm{d}x\,\mathrm{d}y,\\
*(\mathrm{d}y\,\mathrm{d}z) &= \mathrm{d}x\,\mathrm{d}(\mathrm{i}ct) = \mathrm{i}c\,\mathrm{d}x\,\mathrm{d}t,\\
*(\mathrm{d}z\,\mathrm{d}x) &= \mathrm{d}y\,\mathrm{d}(\mathrm{i}ct) = \mathrm{i}c\,\mathrm{d}y\,\mathrm{d}t,\\
*(\mathrm{d}x\,\mathrm{d}y) &= \mathrm{d}z\,\mathrm{d}(\mathrm{i}ct) = \mathrm{i}c\,\mathrm{d}z\,\mathrm{d}t.
\end{aligned}$$

Then if you work it through, you get

$$*F = -\mathrm{i}\sqrt{\frac{\mu_0}{\epsilon_0}}M.$$

I have essentially split the rogue minus sign (which appeared only when $\mathrm{d}t$ was on one side of the equation) into i in each place where $\mathrm{d}t$ appears.

It is more fashionable these days to use only real numbers and to use directly the rules for the Hodge dual that I first wrote down. To do this, you think of the volume form as $\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z\,\mathrm{d}(ct)$ and remember to throw in a minus sign whenever applying the Hodge to a term with $\mathrm{d}t$ in it. (This is particularly nice when using units in which $c = 1$.) But I have always preferred the formulation with imaginary numbers.

This has implications for the notion of length in 4-dimensional spacetime. Whereas

$$đs^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2$$

in 3-dimensional space, the corresponding form in spacetime is

$$đ\tau^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 + \mathrm{d}(\mathrm{i}ct)^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 - c^2\mathrm{d}t^2 = \mathrm{d}s^2 - c^2\mathrm{d}t^2.$$

(Unlike everywhere else in these notes, the multplication with which I'm squaring these differentials is ordinary multiplication, rather than exterior multiplication, and so these differential forms are not exterior or pseudoexterior forms. I'm only following the usual practice in this; it usually doesn't cause confusion, since $\mathrm{d}x \wedge \mathrm{d}x = 0$, so $\mathrm{d}x^2$ is unlikely to mean that.) Properly interpreted, this gives us the entire theory of special relativity.

Einstein's key insight in that theory was that time is a feature of the geometry of the world as much as space is. In particular, whether two events happen at the same time depends on your own motion through space and time, just as much as whether two events happen at the same place. However, time's role in spacetime geometry is different from space's role. If you naively use $\mathrm{d}s^2 + \mathrm{d}t^2$, then this doesn't make sense, because the units don't match, so something like c^2 must appear there to convert between units of space and time. If you use $\mathrm{d}s^2 + c^2\,\mathrm{d}t^2$, then space and time play the same role in geometry, just measured in different units. But since $\mathrm{d}s^2 - c^2\,\mathrm{d}t^2$ is correct, the roles of time and space are different.

It's in this way that there is an absolute notion of speed, because if $\mathrm{d}s/\mathrm{d}t = c$, then $\mathrm{d}\tau^2$ comes out to 0. Also, $\mathrm{d}\tau^2$ can sometimes be negative, so that $\mathrm{d}\tau$ itself is imaginary; this happens for motion that (like the motion of ordinary matter) is travelling at a speed slower than c. (If you put the minus sign on the other term, then you get something which is positive for ordinary matter, so sometimes people do this; it makes no difference in the end to the physical predictions of the theory.) Of course, this special speed c is the speed of light in a vacuum, although I haven't explained yet why that is so.

Everything which we regard as a vector in space must now be seen as merely the space part of a vector in spacetime, and there is some scalar which also serves as its time part. We'll find that this scalar, while previously thought to be an absolute notion, in fact depends on your frame of reference in special relativity. An important example is the momentum of an object, whose corresponding scalar is (more or less) its energy. Much of this relationship was known before special relativity; if p_x, p_y, and p_z are the

components of momentum in the three dimensions of space, then $[p_x, p_y, p_z, -E]$ is a row vector that can be multiplied by the column vector $\langle \mathrm{d}x, \mathrm{d}y, \mathrm{d}z, \mathrm{d}t \rangle$ to produce the *action differential* $p_x\,\mathrm{d}x + p_y\,\mathrm{d}y + p_z\,\mathrm{d}z - E\,\mathrm{d}t$, which has been used to study mechanics since before Einstein was born. If we now use $\mathrm{i}c\,\mathrm{d}t$ in place of $\mathrm{d}t$, then this means that we must use $\mathrm{i}E/c$ in place of $-E$, and the square of the magnitude of the resulting vector is

$$p_x^2 + p_y^2 + p_z^2 - \frac{E^2}{c^2}.$$

Again this is negative for ordinary matter, and it has units of mass squared times speed squared, so if you divide this by $-c^2$ before taking the square root, then you'll get a real value with units of mass:

$$m = \sqrt{\frac{E^2}{c^4} - \frac{{p_z}^2}{c^2} - \frac{{p_y}^2}{c^2} - \frac{{p_x}^2}{c^2}}.$$

What is this mass? Einstein realized that is simply the mass of the object. In particular, for an object at rest (so that p_x, p_y, and p_z are zero), $m = \sqrt{E^2/c^4} = E/c^2$ (assuming that the energy E is positive); equivalently,

$$E = mc^2.$$

Accordingly, mass is a form of energy that can be converted into other forms, as in a nuclear explosion.