Multivariable Calculus

Toby Bartels
MATH-2080-LN01
2026 Fall

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; you should also have the official course textbook, which is the 4th 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. Conversely, there are parts of Chapters 1 and 2 that won't be assigned, since they apply only to Calculus 2.

Contents

Detailed contents

1 Vector algebra

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.

1.1 *Points*

In this class, we look at spaces with up to 3 dimensions, but most of the ideas in this course 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 think of that space geometrically.

If we assign rectangular coordinates to a space of n dimensions, then the result is called $\mathbb{R}^n$ (or $\mathbf{R}^n$); in particular, a coordinate space of 1 dimension is $\mathbb{R}^1$ or simply $\mathbb{R}$, 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 $\mathbb{R}$; then to use x and y as the coordinates in $\mathbb{R}^2$; then x, y, and z in $\mathbb{R}^3$; and finally x_1, x_2, ..., and x_n in $\mathbb{R}^n$ generally. But there are other systems of notation; as long as you list n independent variables in a row, then you have a valid list of coordinates for $\mathbb{R}^n$.

A **point** in $\mathbb{R}^n$ 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 $\mathbb{R}$, while (x, y) gives a point in the coordinate plane $\mathbb{R}^2$, (x, y, z) gives a point in the coordinate space $\mathbb{R}^3$, and $(x_1, x_2, \ldots, x_n)$ gives a point in $\mathbb{R}^n$ (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 terribly confusing convention, so I won't follow it, but you will see it sometimes, even in the textbook.

1.2 *Vectors*

A **vector** is a movement between points. For example, to move in $\mathbb{R}^2$ 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 $\mathbb{R}^n$ 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 $[1, -2]$, $\begin{bmatrix} 1 \\ -2 \end{bmatrix}$, and $\langle 1, -2 \rangle$. I will use the last of these, since it's what the textbook uses. (There's another notation, which the textbook uses even more often than $\langle 1, -2 \rangle$, and that is $\mathbf{i} - 2\mathbf{j}$. However, I'll save that for Section 1.4 below.) 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 points, 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,

$$\Delta P = P_2 - P_1,$$

or

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

When people want a symbol for the vector from P_1 to P_2 but don't want to refer to subtraction of points, then they'll sometimes write $\overrightarrow{P_1 P_2}$ instead of $P_2 - P_1$, but I won't do that.

When you give a vector a name of its own, without reference to points, it's common to use a boldface lowercase letter, such as $\boldsymbol{v}$ (or $\mathbf{v}$). Thus, if I use $\boldsymbol{v}$ to refer to the vector that I've been using as an example throughout this section, then I would write $\boldsymbol{v} = \langle 1, -2 \rangle$. 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.

1.3 *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 $\boldsymbol{u}$ and $\boldsymbol{v}$ are vectors in n dimensions, both representing some motion within $\mathbb{R}^n$, then $\boldsymbol{u} + \boldsymbol{v}$ represents the motion of $\boldsymbol{u}$ followed by the motion of $\boldsymbol{v}$. 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 $\boldsymbol{v}$ is a vector, then $-\boldsymbol{v}$ 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 $\boldsymbol{u} - \boldsymbol{v}$ just means $\boldsymbol{u} + (-\boldsymbol{v})$.

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 $\mathbb{R}^n$:

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

Besides adding and subtracting vectors, you can multiply or divide them by real numbers. For example, if $\boldsymbol{v}$ is a vector representing some motion, then $2\boldsymbol{v}$ represents doing that motion twice, $1/2\,\boldsymbol{v}$ or $\boldsymbol{v}/2$ represents performing half of that motion, $-2\boldsymbol{v}$ 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 $\mathbb{R}^n$:

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

This operation is called **scalar multiplication** (or *scalar division* in the case of $\boldsymbol{v}/a$), 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, $2\boldsymbol{u} + 3\boldsymbol{v} - 5\boldsymbol{w}$ is a linear combination of the vectors $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{w}$. Geometrically, this represents moving twice according to $\boldsymbol{u}$, then moving 3 times according to $\boldsymbol{v}$, and then moving 5 times the reverse of the motion given by $\boldsymbol{w}$.

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, while 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 $\boldsymbol{u}$ and $\boldsymbol{v}$ are vectors, then $2A - 3B + 2C + 4\boldsymbol{u} - 5\boldsymbol{v}$ is a point (because $2-3+2=1$), while $2A - 3B + C + 4\boldsymbol{u} - 5\boldsymbol{v}$ is a vector (because $2-3+1=0$). Geometrically, $2A - 3B + 2C + 4\boldsymbol{u} - 5\boldsymbol{v}$ 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 $\boldsymbol{u}$, and then moving 5 times the reverse of the motion given by $\boldsymbol{v}$. (That is, think of it as $A + (A-B) + 2(C-B) + 4\boldsymbol{u} - 5\boldsymbol{v}$.) Similarly, $2A - 3B + C + 4\boldsymbol{u} - 5\boldsymbol{v}$ 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 $\boldsymbol{u}$, and then moving 5 times the reverse of the motion given by $\boldsymbol{v}$. (That is, think of it as $2(A-B)+(C-B)+4\boldsymbol{u}-5\boldsymbol{v}$.)

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, $\boldsymbol{u}+\boldsymbol{v}=\boldsymbol{v}+\boldsymbol{u}$, $(A+\boldsymbol{u})+\boldsymbol{v}=A+(\boldsymbol{u}+\boldsymbol{v})$, $2(\boldsymbol{u}+\boldsymbol{v})=2\boldsymbol{u}+2\boldsymbol{v}$, 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, either directly or 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, in Sections 1.6–1.12.

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

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

It obeys algebraic rules analogous to those obeyed by the real number 0, such as $\mathbf{0}+\boldsymbol{v}=\boldsymbol{v}$, $\boldsymbol{v}-\boldsymbol{v}=\mathbf{0}$, 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.)

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

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

In $\mathbb{R}^3$, 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 $\mathbb{R}^n$, 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 $\mathbb{R}^2$ or in $\mathbb{R}^3$, then you know what they mean.)

If a component of a vector happens to be 1, then you can leave out that component from 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 out that term entirely. For example, $\langle 1,-2\rangle=1\mathbf{i}+(-2)\mathbf{j}=\mathbf{i}-2\mathbf{j}$. In $\mathbb{R}^3$, $\langle 1,-2,0\rangle$ 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 $\mathbb{R}^n$. 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. I'll generally continue to use list notation for both vectors and points, for consistency, but you may use either notation.

1.5 *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 $\mathbb{R}^n$ 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 $\boldsymbol{v}$ as $\|\boldsymbol{v}\|$, even though the textbook writes this as simply $|\boldsymbol{v}|$ instead. This is because people occasionally use $|\boldsymbol{v}|$ to mean the vector that's the componentwise absolute value of $\boldsymbol{v}$.) As a statement about distances, this is the n-dimensional generalization of the Pythagorean Theorem.

One basic algebraic property of lengths is

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

(Note that I still write $|a|$ when a is a scalar, even though I you use $\|\boldsymbol{v}\|$ when $\boldsymbol{v}$ 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,

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

Also,

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

conversely, if $\|\boldsymbol{v}\| = 0$, then it must be that $\boldsymbol{v} = \mathbf{0}$. (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 $\|\boldsymbol{u} + \boldsymbol{v}\|$; however, we can say

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

This is called the **triangle inequality**, since if you draw a triangle whose sides are $\boldsymbol{u}$, $\boldsymbol{v}$, and their sum $\boldsymbol{u} + \boldsymbol{v}$, 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 $\boldsymbol{v} \neq \boldsymbol{0}$ (so that you can divide by $\|\boldsymbol{v}\|$), then $\boldsymbol{v}/\|\boldsymbol{v}\|$ is a vector whose own magnitude is 1. (This is because

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

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

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

For some reason, the textbook 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 $\boldsymbol{v} = \|\boldsymbol{v}\|\hat{\boldsymbol{v}}$; this expresses the common slogan that a vector has both a length and a direction. (However, the zero vector has only a length, which is 0, and no way to pick out any unit vector as its direction vector.)

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

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

(I'll skip this derivation too, but it's based on first replacing $\boldsymbol{v}$ with $-\boldsymbol{v}$ in the triangle inequality, squaring both sides, and rearranging terms to derive one half of this result, then going back to the beginning and replacing $\boldsymbol{u}$ with $\boldsymbol{u} - \boldsymbol{v}$, squaring both sides again, and rearranging terms to derive the other half of the result.) If you draw a triangle whose sides are $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{u} - \boldsymbol{v}$ (so that $\boldsymbol{u}$ and $\boldsymbol{v}$ 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 $\boldsymbol{u}$ and $\boldsymbol{v}$, and the inequality verifies that this lies within the possible range of values for a cosine. (If either $\boldsymbol{u}$ or $\boldsymbol{v}$ 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 (aka inverse cosine), so taking the arccosine of the expression above gives you this angle:

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

(In $\mathbb{R}^2$, and only in $\mathbb{R}^2$, it's possible to distinguish clockwise and counterclockwise angles, which I'll come back to when I discuss the scalar cross product in Section 1.11.) 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 at the end of Section 1.7, using the dot product, but it's nice to know that angles can be calculated from lengths alone.)

Two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ 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 $\boldsymbol{u} \perp \boldsymbol{v}$. In terms of lengths, $\boldsymbol{u}$ and $\boldsymbol{v}$ are perpendicular when $\|\boldsymbol{u} - \boldsymbol{v}\|^2 = \|\boldsymbol{u}\|^2 + \|\boldsymbol{v}\|^2$ (or replacing $\boldsymbol{v}$ with $-\boldsymbol{v}$, which would also be perpendicular to $\boldsymbol{u}$, $\|\boldsymbol{u} + \boldsymbol{v}\|^2 = \|\boldsymbol{u}\|^2 + \|\boldsymbol{v}\|^2$). Similarly, $\boldsymbol{u}$ and $\boldsymbol{v}$ are **parallel** if the angle between them is the zero angle (0, or 0°), whose cosine is 1; the symbol for this is $\boldsymbol{u} \parallel \boldsymbol{v}$. However, people sometimes use this symbol (or even the word 'parallel') to include the case where $\boldsymbol{u}$ and $\boldsymbol{v}$ are **antiparallel**, meaning that the angle between them is a straight angle (π, or 180°), whose cosine is -1. In terms of lengths, $\boldsymbol{u}$ and $\boldsymbol{v}$ are parallel if $\|\boldsymbol{u} - \boldsymbol{v}\|^2 = \big(\|\boldsymbol{u}\| - \|\boldsymbol{v}\|\big)^2$ (so $\|\boldsymbol{u} - \boldsymbol{v}\| = \big|\|\boldsymbol{u}\| - \|\boldsymbol{v}\|\big|$, or $\|\boldsymbol{u} + \boldsymbol{v}\| = \|\boldsymbol{u}\| + \|\boldsymbol{v}\|$), and $\boldsymbol{u}$ and $\boldsymbol{v}$ are antiparallel if $\|\boldsymbol{u} - \boldsymbol{v}\|^2 = (\|\boldsymbol{u}\| + \|\boldsymbol{v}\|)^2$ (so $\|\boldsymbol{u} - \boldsymbol{v}\| = \|\boldsymbol{u}\| + \|\boldsymbol{v}\|$, or $\|\boldsymbol{u} + \boldsymbol{v}\| = \big|\|\boldsymbol{u}\| - \|\boldsymbol{v}\|\big|$).

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: $\boldsymbol{u}$ and $\boldsymbol{v}$ are parallel if there is a scalar $k > 0$ such that $\boldsymbol{u} = k\boldsymbol{v}$; they're antiparallel if there is a scalar $k < 0$ such that $\boldsymbol{u} = k\boldsymbol{v}$.

1.6 *Projections*

If you have two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, 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 $A + \boldsymbol{v}$ to the line through A and $A + \boldsymbol{u}$ 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 $\boldsymbol{v}$ onto $\boldsymbol{u}$, denoted $\operatorname{proj}_{\boldsymbol{u}} \boldsymbol{v}$. Sometimes people also consider the projection of $\boldsymbol{v}$ *perpendicular* to $\boldsymbol{u}$; this is the vector from B to $A + \boldsymbol{v}$:

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

(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 $\boldsymbol{v}$ in the direction of $\boldsymbol{u}$, denoted $\operatorname{comp}_{\boldsymbol{u}} \boldsymbol{v}$; this is a scalar chosen so that

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

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

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

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

I have not allowed $\boldsymbol{v}$ to be the zero vector, because then $A + \boldsymbol{v}$ is simply A, right on the line through A and $A + \boldsymbol{u}$, 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 $\operatorname{proj}_{\boldsymbol{u}} \boldsymbol{v}$, which is $B - A$, is also $\mathbf{0}$. Thus, we have these results:

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

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

$$\operatorname{proj}_{\boldsymbol{u}} (\boldsymbol{v}_1 + \boldsymbol{v}_2) = \operatorname{proj}_{\boldsymbol{u}} \boldsymbol{v}_1 + \operatorname{proj}_{\boldsymbol{u}} \boldsymbol{v}_2, \text{ so } \operatorname{comp}_{\boldsymbol{u}} (\boldsymbol{v}_1 + \boldsymbol{v}_2) = \operatorname{comp}_{\boldsymbol{u}} \boldsymbol{v}_1 + \operatorname{comp}_{\boldsymbol{u}} \boldsymbol{v}_2;$$
$$\operatorname{proj}_{\boldsymbol{u}} (a\boldsymbol{v}) = a \operatorname{proj}_{\boldsymbol{u}} \boldsymbol{v}, \text{ so } \operatorname{comp}_{\boldsymbol{u}} (a\boldsymbol{v}) = a \operatorname{comp}_{\boldsymbol{u}} \boldsymbol{v}.$$

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 $\boldsymbol{v}$ forms the hypotenuse of a right triangle, one of whose legs is $\operatorname{proj}_{\boldsymbol{u}} \boldsymbol{v}$, and whose angle next to that leg is $\angle(\boldsymbol{u}, \boldsymbol{v})$ if $\boldsymbol{u}$ and $\boldsymbol{v}$ have roughly the same direction or $\pi - \angle(\boldsymbol{u}, \boldsymbol{v})$ if they have roughly opposite directions. In the first case,

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

in the other case,

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

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

$$\operatorname{comp}_{\boldsymbol{u}} \boldsymbol{v} = \|\boldsymbol{v}\| \cos\angle(\boldsymbol{u}, \boldsymbol{v})$$

as long as $\boldsymbol{v} \neq \boldsymbol{0}$. (If $\boldsymbol{v} = \boldsymbol{0}$, then the angle $\angle(\boldsymbol{u}, \boldsymbol{v})$ 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 on page 8 how to express this cosine using only $\|\boldsymbol{u}\|$, $\|\boldsymbol{v}\|$, and $\|\boldsymbol{u} - \boldsymbol{v}\|$, but for now, let's just leave it as $\cos\angle(\boldsymbol{u}, \boldsymbol{v})$.

1.7 *The dot product*

To treat $\boldsymbol{u}$ and $\boldsymbol{v}$ equally, the disussion of projections and components above suggests that we'll get an interesting operation if we define

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

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

$$\begin{gathered}\boldsymbol{u} \cdot (\boldsymbol{v}_1 + \boldsymbol{v}_2) = (\boldsymbol{u} \cdot \boldsymbol{v}_1) + (\boldsymbol{u} \cdot \boldsymbol{v}_2), \\ \boldsymbol{u} \cdot (a\boldsymbol{v}) = a(\boldsymbol{u} \cdot \boldsymbol{v}).\end{gathered}$$

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

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

and then these properties also follow:

$$\begin{gathered}(\boldsymbol{u}_1 + \boldsymbol{u}_2) \cdot \boldsymbol{v} = (\boldsymbol{u}_1 \cdot \boldsymbol{v}) + (\boldsymbol{u}_2 \cdot \boldsymbol{v}), \\ (a\boldsymbol{u}) \cdot \boldsymbol{v} = a(\boldsymbol{u} \cdot \boldsymbol{v}).\end{gathered}$$

The definition $\|\boldsymbol{u}\| \operatorname{comp}_{\boldsymbol{u}} \boldsymbol{v}$ allows $\boldsymbol{v}$ to be $\boldsymbol{0}$, but not $\boldsymbol{u}$. However, since the operation is symmetric when the vectors are nonzero, we can define it so that it continues to be symmetric, so that $\boldsymbol{0} \cdot \boldsymbol{v} = 0$ as well as $\boldsymbol{u} \cdot \boldsymbol{0} = 0$. In particular, we define $\boldsymbol{0} \cdot \boldsymbol{0}$ to be 0. (Thus, it remains true in a way that $\boldsymbol{u} \cdot \boldsymbol{v} = \|\boldsymbol{u}\| \, \|\boldsymbol{v}\| \cos\angle(\boldsymbol{u}, \boldsymbol{v})$, even when $\angle(\boldsymbol{u}, \boldsymbol{v})$ 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**. (But don't confuse *scalar multiplication*, describing the operation for $a\boldsymbol{v}$, with the *scalar product*, describing the result of the operation $\boldsymbol{u} \cdot \boldsymbol{v}$.) The properties above state

that the dot product distributes 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

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

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

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

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

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, the cosine between any two different ones is 0, so 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:

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

Even lengths can be expressed using the dot product:

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

1.8 *Row vectors*

I developed the dot product geometrically, and we've seen that it's closely related to lengths and angles. At the end of Section 1.5, I remarked that lengths and angles don't always make sense in context, 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 in the previous paragraph 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, as you'll see in Section 1.13. Sometimes row vectors are distinguished from column vectors by choosing a different notation for vectors from the common notations listed on page 3. When column vectors are written $\begin{bmatrix} a \\ b \end{bmatrix}$, row vectors are usually written $\begin{bmatrix} a & b \end{bmatrix}$. This notation also comes from matrix theory, as you'll also see in Section 1.13.

Row vectors obey the same rules of arithmetic as column vectors; here is a list of operations with these that make sense even when lengths and angles do not:

- 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 no useful notion of 'row point' that can interact with row vectors in the way that points interact with column vectors.

1.9 *Area*

Now let's go back to a geometric conception of vectors. If you take two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ 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 $\boldsymbol{u}$, one running from A to $A + \boldsymbol{u}$, the other running from $A + \boldsymbol{v}$ to $A + \boldsymbol{v} + \boldsymbol{u}$; the other two opposite sides are $\boldsymbol{v}$, one running from A to $A + \boldsymbol{v}$, the other running from $A + \boldsymbol{u}$ to $A + \boldsymbol{u} + \boldsymbol{v}$ (which of course is the same as $A + \boldsymbol{v} + \boldsymbol{u}$).

This question can be asked in any number of dimensions, and the answer may be written $\|\boldsymbol{u} \times \boldsymbol{v}\|$. This notation suggests that this area will be the magnitude of something more fundamental, which is $\boldsymbol{u} \times \boldsymbol{v}$ 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 $\|\boldsymbol{u} \times \boldsymbol{v}\|$. However, I can give you the terminology: whatever $\boldsymbol{u} \times \boldsymbol{v}$ 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,

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

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

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

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

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

(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, even though he didn't use vectors; he expressed it directly using the distances between the points. (Hero also invented the steam engine, the windmill, and the vending machine, although none of those caught on at the time.)

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

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

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

(The last of these assumes that $\boldsymbol{u} \neq \boldsymbol{0}$, so that projection perpendicular to $\boldsymbol{u}$ 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.

1.10 *The cross product in three dimensions*

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

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

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

$$\boldsymbol{u} \times (\boldsymbol{v}_1 + \boldsymbol{v}_2) = \boldsymbol{u} \times \boldsymbol{v}_1 + \boldsymbol{u} \times \boldsymbol{v}_2,$$
$$\boldsymbol{u} \times a\boldsymbol{v} = a(\boldsymbol{u} \times \boldsymbol{v}).$$

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

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

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:

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

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.

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

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

This can be calculated with determinants as well:

$$\langle a, b, c\rangle \cdot \langle d, e, f\rangle \times \langle g, h, i\rangle = \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix}.$$

Geometrically, this represents a volume; more precisely, $|\boldsymbol{u} \cdot \boldsymbol{v} \times \boldsymbol{w}|$ is the volume of a parallelepiped whose edges are $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{w}$, and $\boldsymbol{u} \cdot \boldsymbol{v} \times \boldsymbol{w}$ is positive if you can curl the fingers of your right hand from $\boldsymbol{u}$ to $\boldsymbol{v}$ and stick out your thumb along $\boldsymbol{w}$ but negative if your thumb points the wrong way.

1.11 *The cross product in two dimensions*

For vectors in $\mathbb{R}^2$, we can interpret $\boldsymbol{u} \times \boldsymbol{v}$ as a scalar, so this is sometimes called the *scalar cross product*. The absolute value $|\boldsymbol{u} \times \boldsymbol{v}|$ is the $\|\boldsymbol{u} \times \boldsymbol{v}\|$ from Section 1.9; $\boldsymbol{u} \times \boldsymbol{v}$ itself is positive if you turn counterclockwise to go from $\boldsymbol{u}$ to $\boldsymbol{v}$ 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 $\boldsymbol{u}$ and $\boldsymbol{v}$ are parallel (or antiparallel), or if either of them is the zero vector $\mathbf{0}$, then $\boldsymbol{u} \times \boldsymbol{v}$ is just 0.

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

$$\begin{aligned} \boldsymbol{u} \times (\boldsymbol{v}_1 + \boldsymbol{v}_2) &= \boldsymbol{u} \times \boldsymbol{v}_1 + \boldsymbol{u} \times \boldsymbol{v}_2, \\ \boldsymbol{u} \times a\boldsymbol{v} &= a(\boldsymbol{u} \times \boldsymbol{v}), \\ \boldsymbol{u} \times \boldsymbol{v} &= -\boldsymbol{v} \times \boldsymbol{u}. \end{aligned}$$

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

Another way to think of the scalar cross product is to embed $\mathbb{R}^2$ within $\mathbb{R}^3$; 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 $\boldsymbol{u} \times \boldsymbol{v}$, you can speak of the triple scalar product $\mathbf{k} \cdot \boldsymbol{u} \times \boldsymbol{v}$, although it's simpler to use the 2-dimensional cross product directly. Yet another way to think of it is as a dot product; much as $a - b$ is the sum of a and $-b$, so $\boldsymbol{u} \times \boldsymbol{v}$ is the dot product of $\boldsymbol{u}$ and $\times\boldsymbol{v}$, where $\times\boldsymbol{v}$ is the result of rotating $\boldsymbol{v}$ clockwise through a right angle. (In general, the result of rotating $\boldsymbol{v}$ clockwise by θ radians is $\cos\theta\, \boldsymbol{v} + \times \sin\theta\, \boldsymbol{v}$; if you rotate counterclockwise, as is more common, then the result is $\cos\theta\, \boldsymbol{v} - \times \sin\theta\, \boldsymbol{v}$.)

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

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

where the bar over the angle symbol indicates this signed angle. There's even a version of the signed angle in 3 dimensions, making this same equation for $\boldsymbol{u} \times \boldsymbol{v}$ true; $\bar{\angle}(\boldsymbol{u}, \boldsymbol{v})$ is now a vector whose magnitude is the (positive) angle between $\boldsymbol{u}$ and $\boldsymbol{v}$ and whose direction is given by the right-hand rule, while the sine of this vector is a vector defined by $\sin \boldsymbol{w} = \sin \|\boldsymbol{w}\|\, \hat{\boldsymbol{w}}$. (The cosine of a vector remains a scalar: $\cos \boldsymbol{w} = \cos \|\boldsymbol{w}\|$.) This paragraph is not important for this course, but it has uses in physics related to angular momentum.

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*. (Similarly, $\times\boldsymbol{v}$ alone is a tensor in more than 2 dimensions.) While the cross product in 4 dimensions (and $\times\boldsymbol{v}$ alone in 3 dimensions) can be interpreted as a matrix (see Section 1.13), more general tensors are even more complicated, and we will not be using these in this course.

1.12 *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 contexts where lengths and angles don't make sense, so we can distinguish *axial* vectors from *polar* vectors in contexts 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 $\mathbb{R}^3$ may be fully described as $\langle -2, -1, 6\rangle$, an axial vector in $\mathbb{R}^3$ 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,

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

(There is still a convention in play here, however; in a left-handed coordinate system, you would write $\langle 1, -2, 0\rangle \times \langle 2, 2, 1\rangle = \langle -2, -1, 6\rangle$ 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 $\mathbb{R}^2$, the cross product of two vectors is a pseudoscalar; in $\mathbb{R}^3$, 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 $\mathbb{R}^3$:

- 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 $\mathbb{R}^2$, 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 variance (row vs column) and chirality (polar vs axial) can prevent you from accidentally doing meaningless calculations.

1.13 *Matrices*

This material isn't necessary right now, but it will be useful in Section 4.9. Recall from Section 1.8 above that we sometimes want to distinguish *column vectors* (the usual kind) from *row vectors* (which appear in dot products with column vectors). One way to distinguish them that I mentioned is to write $\langle a, b\rangle$ as $\begin{bmatrix} a \\ b \end{bmatrix}$ when it's a column vector but as $\begin{bmatrix} a & b \end{bmatrix}$ when it's a row vector. A **matrix** (plural *matrices*) generalizes both of these; it's an array of entries arranged in both columns *and* rows. A typical example of a matrix is

$$\begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix};$$

this matrix has 2 rows and 3 columns, so we call it a 2-*by*-3 matrix. Thus, a column vector in n dimensions is a matrix with n rows and 1 column (also called a *column matrix*), while a row vector in n dimensions is a matrix with 1 row and n columns (also called a *row matrix*). But for matrices in general, the number of rows and the number of columns could each be any whole number.

You can multiply any matrix by a scalar, and you can add or subtract two matrices if they have the same size to get another matrix of that size. This works entry by entry, just as with vectors. But there is also an operation that generalizes the dot product: **matrix multiplication**. In general, you can multiply an m-by-n matrix and an n-by-o matrix to get an m-by-o matrix. The entry in row i and column j in the product matrix is obtained as the dot product of row i of the first matrix and column j of the second matrix, each thought of a vector in n dimensions. (This operation distributes over matrix addition and associates with scalar multiplication, which is why we can think of it as a kind of multiplication.) For example, multiply the 2-by-3 matrix above by the 3-by-1 matrix that is the vector $\langle x, y, z\rangle$ thought of as a column matrix, to get a 2-by-1 matrix:

$$\begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \langle a, b, c\rangle \cdot \langle x, y, z\rangle \\ \langle d, e, f\rangle \cdot \langle x, y, z\rangle \end{bmatrix} = \begin{bmatrix} ax + by + cz \\ dx + ey + fz \end{bmatrix},$$

which is the vector $\langle ax + by + cz, dx + ey + fz\rangle$ thought of as a column matrix. (In a way, this is the most fundamental thing that matrices do: they transform vectors in a homogeneous-linear way.) This operation has many uses, but the reason that I'm bringing it up here is that you can use it in the multivariable Chain Rule in Section 4.9.

There are a couple of special matrices that you should know. Given any whole numbers m and n, there is an m-by-n **zero matrix** whose entries are all the constant 0. Adding a zero matrix to another matrix (of the same size) doesn't change it, and multiplying a zero matrix by another matrix (if their sizes allow them to be multiplied) results in another zero matrix (of the appropriate size). In this way, the zero matrices behave like the number zero. The m-by-n zero matrix can be denoted $\mathbf{0}_{m,n}$ or $\mathbf{O}_{m,n}$, or just $\mathbf{0}$ or $\mathbf{O}$ if the size is clear from context. Also, given any whole number n, there is an n-by-n **identity matrix** whose entries are mostly 0 except along the main diagonal (where the row number equals the column number); there the entries are 1. Mutliplying an identity matrix by another matrix (if their sizes allow them to be multiplied) doesn't change the other matrix. The n-by-n identity matrix can be denoted $\mathbf{1}_n$ or $\mathbf{I}_n$, or just $\mathbf{1}$ or $\mathbf{I}$ if the size is clear from context. (I've put these in boldface, but not everybody does that.) Also, if A and B are matrices that can be multiplied in either order, and AB and BA are both identity matrices, then we say that A and B are **inverse matrices** (and that each is the *inverse matrix* of the other one) and write $B = A^{-1}$ (and $A = B^{-1}$); an inverse matrix is unique if it exists. I mention this because it comes up a lot, but we're not going to need inverse matrices in this course.

Another common operation on matrices is the **transpose**, in which the rows and columns are swapped. Although we won't really need it in this course, there's some notation related to it that you might see even

in material about vectors, so it's worth mentioning here. The transpose of a matrix A is denoted $A^\top$; for example,

$$\begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix}^\top = \begin{bmatrix} a & d \\ b & e \\ c & f \end{bmatrix}.$$

The upper left and lower right positions are unchanged, while the lower left and upper right switch places. When applied to vectors, this operation turns row vectors into column vectors and column vectors into row vectors. For this reason, yet another notation for the column vector $\langle a, b \rangle$, which is really a column matrix $\begin{bmatrix} a \\ b \end{bmatrix}$, is $\begin{bmatrix} a & b \end{bmatrix}^\top$. (That way, you can write it all in one line.) Conversely, an alternative notation for the dot product $\boldsymbol{u} \cdot \boldsymbol{v}$ of two vectors (themselves thought of as column matrices) is $\boldsymbol{u}^\top \boldsymbol{v}$; here, the transpose turns the column matrix $\boldsymbol{u}$ into a row matrix, this 1-by-n row matrix $\boldsymbol{u}^\top$ is multiplied by the n-by-1 column matrix $\boldsymbol{v}$ to produce a 1-by-1 matrix (with only one entry), and this is then interpreted as a scalar (its only entry). This is kind of a complicated way to think of the dot product, but some people like to write it like that.

2 Parametrized curves

Besides individual points and vectors, one can also consider variable points and vectors, which are the outputs of point- and vector-valued functions and can be interpreted geometrically as parametrized curves.

2.1 *Point- and vector-valued functions*

A **point-valued function** in $\mathbb{R}^n$ consists of n ordinary functions, all with the same domain. For example, a point-valued function in $\mathbb{R}^2$ 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 point-valued function (f, g), which takes a real number t as input and produces the point as output:

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

(in this example), where O is the origin of a coordinate system and $\mathbf{i}$ and $\mathbf{j}$ are its standard basic vectors. A **vector-valued function** in $\mathbb{R}^n$ also consists of n ordinary functions, all with the same domain. But now we think of the output as a vector:

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

(in this 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 the point-valued function (f, g) and the vector-valued function $\langle f, g\rangle$.

The textbook often doesn't distinguish between a point P and its position vector $\boldsymbol{r} = P - \mathrm{O}$ (which it writes upright as $\mathbf{r}$). 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 you don't pick a point and call it the origin. 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 $P - \mathrm{O}$ to get $\boldsymbol{r}$, you find that the components of $\boldsymbol{r}$ 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 (a location) or to a vector (a movement).

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 a point O deemed to be the origin, but that is how the textbook insists on doing it starting in Chapter 12. However, this isn't an issue in Chapter 10, since the textbook isn't discussing vectors there.

2.2 *Velocity and acceleration*

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 $\boldsymbol{v}$ for $\mathrm{d}P/\mathrm{d}t$ (which can also be written as $\mathrm{d}\boldsymbol{r}/\mathrm{d}t$), then $\mathrm{d}\boldsymbol{v}/\mathrm{d}t$ 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 $\mathrm{d}\|\boldsymbol{v}\|/\mathrm{d}t$, 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 $\boldsymbol{C}$ 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:

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

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 if we already know an indefinite integral:

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

This is very handy when solving initial-value problems. Since $\boldsymbol{v} = \langle 3, 0, 4\rangle$ when $t = 0$, I was doing the same operation to both sides of the equation in the first step in which I introduced semidefinite integrals; similarly, the second introduction of semidefinite 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.

2.4 *Standard parametrizations*

The simplest curve to parametrize is a straight line. A line through the origin along a vector $\langle a, b\rangle$ can be parametrized with

$$x = at,$$
$$y = bt.$$

If you want a *ray* (half-line) starting at the origin and travelling in the direction of this vector, then use the same formulas for x and y but add the restriction $t \geq 0$ for a closed ray (including the endpoint at the origin) or $t > 0$ for an open ray (*not* including that endpoint). For a line *segment* running along the length of that vector, use the restriction $0 \leq t \leq 1$ for a closed line segment (including both endpoints) or $0 < t < 1$ for an open line segment (including neither endpoint). (It's also possible to consider half-open half-closed line segments.) For both rays and line segments, the closed version is the usual standard, although there are times when another version is needed instead.

If the line (or ray or line segment) doesn't go through the origin, then you'll need some point (x_1, y_1) that it *does* go through. Then you can use

$$x = x_1 + at,$$
$$y = y_1 + bt.$$

Again, without any restriction on t, this is a line; but you can restrict t as above to get a ray or a line segment. Or if you have two points on the line, then you can subtract them to get the relevant vector. Then the parametrization becomes

$$x = x_1 + (x_2 - x_1)t,$$
$$y = y_1 + (y_2 - y_1)t.$$

All of this works in any number of dimensions; the line through P_1 along the vector $\boldsymbol{v}$ has the parametrization

$$P = P_1 + t\boldsymbol{v},$$

and the line through P_1 and P_2 is

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

The same restrictions on t as before will turn these into rays or line segments.

Going back to 2 dimensions, the unit circle (whose radius is 1 and whose centre is at the origin) is usually parametrized like this:

$$x = \cos t,$$
$$y = \sin t.$$

If there are no restrictions on t, then you are effectively going around and around the circle forever, counterclockwise (in a counterclockwise coordinate system). If you want the parametrization to be one-to-one, so that every point on the circle is covered exactly once, then you need a restriction on t; the usual one is $0 \leq t < 2\pi$. It's even more common to use

$$0 \leq t \leq 2\pi;$$

this is *almost* one-to-one (since only the point $(1, 0)$ is covered twice, once when $t = 0$ and once when $t = 2\pi$), and it has a compact domain (which is helpful for some things). So this restriction is the standard one for a circle.

If the radius of the circle is r, then the parametrization becomes

$$x = r\cos t,$$
$$y = r\sin t.$$

If the circle is centred at (h, k) instead of at the origin, then the parametrization becomes

$$x = h + r\cos t,$$
$$y = k + r\sin t.$$

You use the same restrictions as before to make the parametrization one-to-one or almost one-to-one.

For something a little fancier than a circle, consider a helix, that is a circular spiral that extends into a third dimension. This can be parametrized by

$$\begin{aligned} x &= h + r\cos t, \\ y &= k + r\sin t, \\ z &= mt + b. \end{aligned}$$

So x and y are going around in a circle, while z is moving up in a straight line. You could also swap some of the variables around for a horizontal helix instead of a vertical one; if the helix is going off at an angle, then it's more complicated. A helix continues up and down forever, so even though x and y will repeat with a period of 2π, there is no restriction on the parameter t (unless you only want part of the helix).

Another useful parametrization is the graph of a function f. For this, you can use x itself as the parameter:

$$\begin{aligned} x &= t, \\ y &= f(t). \end{aligned}$$

Since you can always call the parameter something else instead of t, you can even call it x:

$$\begin{aligned} x &= x, \\ y &= f(x). \end{aligned}$$

If you only want the graph of the function restricted to an interval $[a, b]$, then place this restriction on the parameter:

$$a \le t \le b$$

(or $a \le x \le b$ if you are calling the parameter x instead of t). This works more generally any time you have an equation that you can solve for y; if you get a unique solution, then this equation defines a function, and the equation $y = f(x)$ in the parametrization above is the equation that you get when you solve for y.

If you solve for x instead of for y, then you can say that x is some function g of y. This isn't the graph of that function exactly, since the variables come in the wrong order, but you can still parametrize the curve using y as the parameter:

$$\begin{aligned} x &= g(t), \\ y &= t. \end{aligned}$$

Again, you can put a restriction on t if you only want certain values of the independent variable, which is now y.

2.5 *Linear geometry*

There are some 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 2-dimensional cross product (instead of only the 3-dimensional cross product as the book does).

As we saw in the last section, a parametric equation for the line through a point P_0 in the direction of a nonzero vector $\boldsymbol{v}$ is

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

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 $\boldsymbol{v}$ in 2 dimensions is

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

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

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

(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 $\boldsymbol{v}$ is

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

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 nonzero vector $\boldsymbol{n}$ is

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

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

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

2.6 *Derivatives and parametrized curves*

If x and y are given as functions of t, as happens with a parametrized curve in 2 dimensions, then the formulas for derivatives of y with respect to x, in terms of the derivatives of x and y with respect to t, ought to fall directly out of the notation. Unfortunately, the usual notation for higher derivatives prevents this.

To see how this should work, consider the first derivative. There, the formula is

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\mathrm{d}y/\mathrm{d}t}{\mathrm{d}x/\mathrm{d}t}.$$

That is, simply divide both sides of the fraction by $\mathrm{d}t$. Another even slicker way to do this would be to reinterpret the differentials as derivatives with respect to t; that is, writing a dot above a quantity to indicate differentiation with respect to t, write

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\dot{y}}{\dot{x}}.$$

But If you try to do this with second derivatives, based on the usual notation for them, then you get a formula which is *wrong*:

$$\frac{\mathrm{d}^2y}{\mathrm{d}x^2} \neq \frac{\ddot{y}}{\dot{x}^2} = \frac{\mathrm{d}^2y/\mathrm{d}t^2}{(\mathrm{d}x/\mathrm{d}t)^2}.$$

(Here, I've written '$\neq$' to show that '$=$' would have been wrong, but it's possible that these may happen to be equal in certain examples.)

To get the correct formula instead, we simply need to differentiate $\dot{y}/\dot{x}$ using the Quotient Rule:

$$\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)\left(\frac{\dot{y}}{\dot{x}}\right) = \frac{\dot{x}\,\mathrm{d}\dot{y}/\mathrm{d}t - \dot{y}\,\mathrm{d}\dot{x}/\mathrm{d}t}{\dot{x}^2} = \frac{\dot{x}\ddot{y} - \dot{y}\ddot{x}}{\dot{x}^2}.$$

Dividing by $\dot{x}$ to change $\mathrm{d}/\mathrm{d}t$ to $\mathrm{d}/\mathrm{d}x$, the second derivative of y with respect to x is

$$(\mathrm{d}/\mathrm{d}x)^2 y = \frac{\dot{x}\ddot{y} - \dot{y}\ddot{x}}{\dot{x}^3} = \frac{\ddot{y}}{\dot{x}^2} - \frac{\dot{y}}{\dot{x}} \cdot \frac{\ddot{x}}{\dot{x}^2};$$

in other words, the naïve formula is only the first term of a two-term expression. This formula is a little long, but it will correctly give you the second derivative of y with respect to x using the first and second derivatives of x and y with respect to t.

There is a symbol for the second derivative using differentials that can serve as a mnemonic for this. To get it, we again differentiate $\mathrm{d}y/\mathrm{d}x$ using the Quotient Rule, only now using the Quotient Rule for differentials rather than the Quotient Rule for derivatives:

$$\mathrm{d}\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right) = \frac{\mathrm{d}x\,\mathrm{d}(\mathrm{d}y) - \mathrm{d}y\,\mathrm{d}(\mathrm{d}x)}{(\mathrm{d}x)^2} = \frac{\mathrm{d}x\,\mathrm{d}^2y - \mathrm{d}y\,\mathrm{d}^2x}{\mathrm{d}x^2}.$$

Dividing by $\mathrm{d}x$ to turn d into $\mathrm{d}/\mathrm{d}x$, the second derivative of y with respect to x is

$$(\mathrm{d}/\mathrm{d}x)^2 y = \frac{\mathrm{d}x\,\mathrm{d}^2y - \mathrm{d}y\,\mathrm{d}^2x}{\mathrm{d}x^3} = \frac{\mathrm{d}^2y}{\mathrm{d}x^2} - \frac{\mathrm{d}y}{\mathrm{d}x}\cdot\frac{\mathrm{d}^2x}{\mathrm{d}x^2}.$$

As you can see, replacing 'd's with dots throughout gives the formula from the previous paragraph again.

For this reason, I don't like to write $\mathrm{d}^2y/\mathrm{d}x^2$ for the second derivative of y with respect to x. Of course, nobody wants to write the formula from the previous paragraph when they just want a symbol for the second derivative; fortunately, you can write $(\mathrm{d}/\mathrm{d}x)^2 y$ for that. This simply means that you apply the operation $\mathrm{d}/\mathrm{d}x$ (find the differential and then divide by $\mathrm{d}x$, or equivalently find the derivative with respect to x) twice to get the second derivative, which is certainly correct. You can even use this as a mnemonic for finding this second derivative: instead of interpeting $\mathrm{d}/\mathrm{d}x$ as taking the differential and then dividing by $\mathrm{d}x$, interpret it as taking the derivative with respect to t and then dividing by $\dot{x}$. This is essentially how the textbook tells you to take the second derivative.

Finally, whether you use either $(\mathrm{d}x\,\mathrm{d}^2y - \mathrm{d}y\,\mathrm{d}^2x)/\mathrm{d}x^3$ or $(\mathrm{d}/\mathrm{d}x)^2 y$, either way you can perform practical caclulations by interpreting the differentials literally. You simply have to write everything in terms of t, put $\mathrm{d}t$ and d^2t in where they naturally appear, and find that the differentials of t cancel in the final answer. Alternatively, anticipating that the differentials of t will cancel, you can ignore them, which turns taking differentials into taking derivatives with respect to t again.

I'll do Example 10.2.2 starting on page 589 of the textbook to illustrate all of these approaches. (In that example, $x = t - t^2$, $y = t - t^3$, and you are asked to find $(\mathrm{d}/\mathrm{d}x)^2 y$.) First, $\mathrm{d}x = \mathrm{d}t - 2t\,\mathrm{d}t$, or $\dot{x} = \mathrm{d}x/\mathrm{d}t = 1 - 2t$. Next, $\mathrm{d}^2x = \mathrm{d}^2t - 2\,\mathrm{d}t^2 - 2t\,\mathrm{d}^2t$ (whre I've applied the Product Rule to the second term of $\mathrm{d}x$), while $\ddot{x} = -2$. Similarly, $\mathrm{d}y = \mathrm{d}t - 3t^2\,\mathrm{d}t$, or $\dot{y} = 1 - 3t^2$. Next, $\mathrm{d}^2y = \mathrm{d}^2t - 6t\,\mathrm{d}t^2 - 3t^2\,\mathrm{d}^2t$, while $\ddot{y} = -6t$.

Now, to find $(\mathrm{d}/\mathrm{d}x)y = \mathrm{d}y/\mathrm{d}x$, either directly divide $(\mathrm{d}t - 3t^2\,\mathrm{d}t)/(\mathrm{d}t - 2t\,\mathrm{d}t)$ and simplify this (by cancelling factors of $\mathrm{d}t$) to $(1 - 3t^2)/(1 - 2t)$, or instead divide $\dot{y}/\dot{x}$, which again gives $(1 - 3t^2)/(1 - 2t)$. (This is pretty much the same process, no matter how you go about it.)

Then to find $(\mathrm{d}/\mathrm{d}x)^2 y$, one way is to differentiate $(\mathrm{d}/\mathrm{d}x)y$ with respect to x again. Either take

$$\frac{\mathrm{d}\left(\frac{1-3t^2}{1-2t}\right)}{\mathrm{d}x} = \frac{\frac{2\,\mathrm{d}t - 6t\,\mathrm{d}t + 6t^2\,\mathrm{d}t}{(1-2t)^2}}{\mathrm{d}t - 2t\,\mathrm{d}t}$$

and simplify by cancelling factors of $\mathrm{d}t$, or take

$$\frac{(\mathrm{d}/\mathrm{d}t)\left(\frac{1-3t^2}{1-2t}\right)}{\dot{x}} = \frac{\frac{2-6t+6t^2}{(1-2t)^2}}{1-2t};$$

either way, you get

$$(\mathrm{d}/\mathrm{d}x)^2 y = \frac{2 - 6t + 6t^2}{(1-2t)^3}.$$

This is essentially how the textbook does this problem. Alternatively, using the formula

$$(\mathrm{d}/\mathrm{d}x)^2 y = \frac{\mathrm{d}x\,\mathrm{d}^2y - \mathrm{d}y\,\mathrm{d}^2x}{\mathrm{d}x^3},$$

we immediately get

$$\frac{(\mathrm{d}t - 2t\,\mathrm{d}t)(\mathrm{d}^2t - 6t\,\mathrm{d}t^2 - 3t^2\,\mathrm{d}^2t) - (\mathrm{d}t - 3t^2\,\mathrm{d}t)(\mathrm{d}^2t - 2\,\mathrm{d}t^2 - 2t\,\mathrm{d}^2t)}{(\mathrm{d}t - 2t\,\mathrm{d}t)^3},$$

which simplifies drastically to the same answer as above. (Notice that there is no need to work out $\mathrm{d}y/\mathrm{d}x$ first!) Or using

$$(\mathrm{d}/\mathrm{d}x)^2 y = \frac{\dot{x}\ddot{y} - \dot{y}\ddot{x}}{\dot{x}^3},$$

you immediately get

$$\frac{(1-2t)(-6t) - (1-3t^2)(-2)}{(1-2t)^3},$$

which simplifies (somewhat less drastically) to the same answer once again. I prefer this last method, which gets the answer in one step after the preliminary calculations and doesn't require quite as much algebra to simplify as the corresponding method using differentials.

2.7 *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 multivariable Calculus, sometimes hidden just under the surface, sometimes out in the open without being acknowledged. I'll try to point 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 in Section 5.4). 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 $\mathrm{d}\boldsymbol{r}$ (where $\boldsymbol{r} = P - \mathrm{O}$), so you can also think of $\mathrm{d}s$ as $\|\mathrm{d}\boldsymbol{r}\|$, the magnitude of the differential of the position vector $\boldsymbol{r}$. In 2 dimensions, where $P = (x, y)$ and $\boldsymbol{r} = \langle x, y\rangle$, $\mathrm{d}\boldsymbol{r} = \mathrm{d}P = \langle \mathrm{d}x, \mathrm{d}y\rangle$, whose magnitude is the arclength element that I wrote down in the previous paragraph. In 3 dimensions, $\mathrm{d}P = \langle \mathrm{d}x, \mathrm{d}y, \mathrm{d}z\rangle$, 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 $\boldsymbol{r}$) 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 $\|\boldsymbol{v}\|\,|\mathrm{d}t| = \|\boldsymbol{v}\|\,\mathrm{d}t$, where $\boldsymbol{v} = \mathrm{d}P/\mathrm{d}t = \mathrm{d}\boldsymbol{r}/\mathrm{d}t$ 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. (In Chapter 5, 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 any parametrization.)

For example, to find the length of the helix with $x = 3\cos t$, $y = 3\sin t$, and $z = 4t$ from $t = 0$ to $t = 6\pi$ (so 3 complete revolutions), you can calculate $\mathrm{d}x = -3\sin t\,\mathrm{d}t$, $\mathrm{d}y = 3\cos t\,\mathrm{d}t$, and $\mathrm{d}z = 4\,\mathrm{d}t$, so that

$$\begin{aligned}\mathrm{d}s &= \sqrt{\mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2} = \sqrt{(-3\sin t\,\mathrm{d}t)^2 + (3\cos t\,\mathrm{d}t)^2 + (4\,\mathrm{d}t)^2} \\ &= \sqrt{9\sin^2 t\,\mathrm{d}t^2 + 9\cos^2 t\,\mathrm{d}t^2 + 16\,\mathrm{d}t^2} = \sqrt{9\,\mathrm{d}t^2 + 16\,\mathrm{d}t^2} = \sqrt{25\,\mathrm{d}t^2} = 5\,|\mathrm{d}t| = 5\,\mathrm{d}t\end{aligned}$$

(with the last step because we set up the integral so that t is increasing, so that $\mathrm{d}t$ is positive). Alternatively, you can calculate $\mathrm{d}x/\mathrm{d}t = -3\sin t$, $\mathrm{d}y/\mathrm{d}t = 3\cos t$, and $\mathrm{d}z/\mathrm{d}t = 4$, so that (since t is increasing)

$$\frac{\mathrm{d}s}{\mathrm{d}t} = \sqrt{\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2 + \left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2 + \left(\frac{\mathrm{d}z}{\mathrm{d}t}\right)^2} = \sqrt{(-3\sin t)^2 + (3\cos t)^2 + 4^2}$$
$$= \sqrt{9\sin^2 t + 9\cos^2 t + 16} = \sqrt{9+16} = \sqrt{25} = 5,$$

so that $\mathrm{d}s = 5\,\mathrm{d}t$ again. Either way, you then integrate $\int_{t=0}^{6\pi} \mathrm{d}s = \int_{t=0}^{6\pi} 5\,\mathrm{d}t = 5t|_{t=0}^{6\pi} = 5(6\pi) - 5(0) = 30\pi$, and so 30π is the total length of 3 turns of this helix.

2.8 *Polar coordinates*

In addition to the usual rectangular coordinates, there are other ways to represent points in $\mathbb{R}^2$ or $\mathbb{R}^3$ as pairs or triples of real numbers. Polar coordinates are a widely used example.

Polar coordinates represent vectors more directly than points, so perhaps we should speak first of polar *components*. Polar components are based on lengths and angles. Specifically, if r and θ are any real numbers such that $\boldsymbol{v} = \langle r\cos\theta, r\sin\theta\rangle$, then we say that r and θ (in that order) are **polar components** of $\boldsymbol{v}$. When using polar components, sometimes people write $\langle r, \theta\rangle$ and make a note somewhere that this is polar instead of rectangular; this works if you *always* use polar components, but otherwise it's confusing, so I will never do this. Sometimes people write $\langle r; \theta\rangle$ instead; the semicolon is supposed to indicate polar components. You should *not* write $r\mathbf{i} + \theta\mathbf{j}$; if you want to write it out using operations on the standard basis vectors, then it has to be $r\cos\theta\,\mathbf{i} + r\sin\theta\,\mathbf{j}$ or $r(\cos\theta\,\mathbf{i} + \sin\theta\,\mathbf{j})$. Another method is to write $\angle\theta$ (the *direction vector* at angle θ) for $\langle\cos\theta, \sin\theta\rangle = \cos\theta\,\mathbf{i} + \sin\theta\,\mathbf{j}$; then you can write $r\angle\theta$ for the vector. This is probably the slickest method, but you have to know what $\angle\theta$ means.

Notice that the magnitude of $r\angle\theta$ is

$$\|r\angle\theta\| = \|\langle r\cos\theta, r\sin\theta\rangle\| = \sqrt{(r\cos\theta)^2 + (r\sin\theta)^2} = \sqrt{r^2\cos^2\theta + r^2\sin^2\theta} = \sqrt{r^2} = |r|.$$

It's tempting to simplify this one more step to r, but that assumes that $r \geq 0$, which might not be true! This highlights that polar components are *not* unique; if you are *choosing* r and θ for a given vector, then you can impose a convenient restriction, such as $r \geq 0$ (which is almost always imposed when possible) and $0 \leq \theta < 2\pi$ (which is often imposed but less often) or $-\pi < \theta \leq \pi$ (which is also fairly common). However, if you are *given* r and θ, then none of these restrictions is guaranteed to hold in general. Finally, if you have the zero vector $\mathbf{0} = \langle 0, 0\rangle$, then you can only have $r = 0$, but now θ can be anything at all!

If you start with a vector $\boldsymbol{v}$ and want to choose polar components for it, then you can always use $r = \|\boldsymbol{v}\|$, which will guarantee $r \geq 0$. Then to choose θ, you can use the signed angle from $\mathbf{i}$ to $\boldsymbol{v}$ described in the middle of Section 1.11, which is $\measuredangle(\mathbf{i}, \boldsymbol{v})$ in the notation of that paragraph. That is, if the vector is in the same direction as $\mathbf{i}$, then we use $\theta = 0$; if it's in the same direction as $\mathbf{j}$, then we use $\theta = \pi/2$; and so on. Turning from $\mathbf{i}$ towards $\mathbf{j}$ (which is usually oriented to be counterclockwise) is a positive angle; turning the other direction is a negative angle. More explicitly, given a vector $\boldsymbol{v} = \langle a, b\rangle$, you can use $\theta = \arccos(a/\|\boldsymbol{v}\|)$ if $b \geq 0$ and $\theta = -\arccos(a/\|\boldsymbol{v}\|)$ if $b < 0$ (with $\theta = 0$ by default if $\|\boldsymbol{v}\| = 0$). This will give $-\pi < \theta \leq \pi$ as the restriction on θ, but it's probably more common to use $\theta = 2\pi - \arccos(a/\|\boldsymbol{v}\|)$ if $b < 0$, so that $0 \leq \theta < 2\pi$ is the restriction. Other choices of restriction on θ are also possible. Having made such a choice, it's common to write $\operatorname{atan2}\boldsymbol{v}$ for this value of θ, called the *phase* of $\boldsymbol{v}$, so that $\hat{\boldsymbol{v}} = \angle\operatorname{atan2}\boldsymbol{v}$ and $\boldsymbol{v} = \|\boldsymbol{v}\|\angle\operatorname{atan2}\boldsymbol{v}$. This notation was first used in the Fortran programming language, and is because $\operatorname{atan2}\langle a, b\rangle = \arctan(b/a)$ when $a > 0$ and $b \geq 0$ (and possibly more often, depending on the restriction chosen for θ). But again, you *don't have to* impose any restrictions on polar components; we have $\boldsymbol{v} = r\angle\theta$ whenever $r = \|\boldsymbol{v}\|$ and $\theta = \operatorname{atan2}\boldsymbol{v} + 2\pi k$ for any integer k, and we *also* have $\boldsymbol{v} = r\angle\theta$ whenever $r = -\|\boldsymbol{v}\|$ and $\theta = \operatorname{atan2}\boldsymbol{v} + \pi + 2\pi k$ for any integer k. (And if $\boldsymbol{v}$ is the zero vector, then we have $\boldsymbol{v} = r\angle\theta$ whenever $r = 0$, no matter what θ is.)

Once you understand polar components of vectors, polar coordinates of points are straightforward. In rectangular coordinates, the point (x, y) is the origin O plus the vector $\langle x, y\rangle$, so now we just write this vector in polar components. That is, r and θ (in that order) are **polar coordinates** of a point P if

$P = \mathrm{O} + r\angle\theta$. Instead of $\mathrm{O} + r\angle\theta$, it's more common to write $(r;\theta)$ (with a semicolon again); or even just (r,θ) (with a note that these are polar coordinates). More explicitly, if $P = (x, y)$, then the requirement for r and θ is

$$x = r\cos\theta,$$
$$y = r\sin\theta.$$

Any numbers r and θ that satisfy these two equations give polar coordinates for the point (x, y). (If you don't learn anything else in this section, learn these two equations.)

Given r and θ, it's easy to calculate x and y, as above. To go back, you can use $r = \|\langle x, y\rangle\|$ and $\theta = \operatorname{atan2}\langle x, y\rangle$; that is:

$$r = \sqrt{x^2 + y^2},$$
$$\theta = \begin{cases} \arccos(x/r) & \text{for } y \geq 0,\ r > 0, \\ 2\pi - \arccos(x/r) & \text{for } y < 0,\ r > 0, \\ 0 & \text{for } r = 0. \end{cases}$$

But again, you can make these formulas true if you *want* to, as long as you are given x and y and are choosing which r and θ to use with them; but you *don't need* to use them, and you *can't assume* them if you're given r and θ by someone else. (You can only assume $x = r\cos\theta$ and $y = r\sin\theta$, nothing more.)

2.9 *Parametrized curves in polar coordinates*

It's very common to describe a parametrized curve by giving r as a function of θ. That is, θ is the parameter, and you have a function f such that $x = f(\theta)\cos\theta$ and $y = f(\theta)\sin\theta$. As a parametrized curve often comes with a restriction on the parameter, so this often comes with the restriction that $0 \leq \theta < 2\pi$; even more common is $0 \leq \theta \leq 2\pi$ (so that the domain of the parametrization will be compact). However, the restriction may be different, or they may be no restriction! This is also a situation where you *cannot* assume that $r \geq 0$; if the function f always takes nonnegative values, that's fine, but if it may take negative values, then you need to accept that.

Although these work like any other parametrized curve, it's also possible to develop specific formulas for this situation. These are based on $\mathrm{d}x = \cos\theta\,\mathrm{d}r - r\sin\theta\,\mathrm{d}\theta = f'(\theta)\cos\theta\,\mathrm{d}\theta - f(\theta)\sin\theta\,\mathrm{d}\theta$ and $\mathrm{d}y = \sin\theta\,\mathrm{d}r + r\cos\theta\,\mathrm{d}\theta = f'(\theta)\sin\theta\,\mathrm{d}\theta + f(\theta)\cos\theta\,\mathrm{d}\theta$. Dividing these and cancelling $\mathrm{d}\theta$,

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{f'(\theta)\sin\theta + f(\theta)\cos\theta}{f'(\theta)\cos\theta - f(\theta)\sin\theta}.$$

A particularly important case of this is at the origin, where $r = 0$; then substituting zero for $f(\theta)$ and cancelling $f'(\theta)$,

$$\left.\frac{\mathrm{d}y}{\mathrm{d}x}\right|_{r=0} = \tan\theta.$$

This result can be helpful just for graphing. Another fact useful for graphing, at least when there are no restrictions on θ, is that because extreme values of y can happen only when $\mathrm{d}y$ is zero or undefined, the highest and lowest points on the graph can only occur when $f'(\theta)\sin\theta + f(\theta)\cos\theta$ is zero or undefined; similarly, the leftmost and rightmost points can only occur when $f'(\theta)\cos\theta - f(\theta)\sin\theta$ is zero or undefined. (But if there is a restriction on θ, then you also have to check the points at the extreme values of θ.)

The arclength element is

$$\mathrm{d}s = \sqrt{\mathrm{d}x^2 + \mathrm{d}y^2} = \sqrt{\mathrm{d}r^2 + r^2\,\mathrm{d}\theta^2} = \sqrt{f'(\theta)^2 + f(\theta)^2}\ |\mathrm{d}\theta|,$$

where $\sin^2\theta + \cos^2\theta = 1$ is used to simplify the formula. Therefore, the length of the curve given in polar coordinates by $r = f(\theta)$ and $\alpha \leq \theta \leq \beta$ is

$$\int_{\theta=\alpha}^{\beta} \sqrt{f'(\theta)^2 + f(\theta)^2}\,\mathrm{d}\theta,$$

as long as $\alpha \leq \beta$ and f is differentiable on $[\alpha, \beta]$.

2.10 *Polar coordinates in higher dimensions*

The simplest way to use polar coordinates in 3 (or more) dimensions is to replace x and y with r and θ, then keep z (and any other rectangular coordinates in higher dimensions). These are called **cylindrical coordinates**. So the equations for cylindrical coordinates in 3 dimensions are

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

notice that z does double duty as both a rectangular coordinate and a cylindrical coordinate. If $\boldsymbol{v} = \langle x, y, z\rangle$, then we can write $\boldsymbol{v} = r\angle\theta + z\mathbf{k}$, where $\angle\theta = \cos\theta\,\mathbf{i} + \sin\theta\,\mathbf{j}$ as in Section 2.8, only now this is intepreted in 3 dimensions as $\angle\theta = \langle\cos\theta, \sin\theta, 0\rangle$. Similarly, if $P = (x, y, z)$, then $P = \mathrm{O} + r\angle\theta + z\mathbf{k}$.

Note that r here is *not* the magnitude of the vector $\langle x, y, z\rangle$, even assuming the restriction $r \geq 0$, unless z happens to be 0. Another way to use polar coordinates in higher dimensions is to use this magnitude, which we call ρ (at least in 3 dimensions), and more angles. Specifically, switch from (z, r) to (ρ, φ) in exactly the same way that polar coordinates switch from (x, y) to (r, θ). That is, $z = \rho\cos\varphi$, and $r = \rho\sin\varphi$. These are called **spherical coordinates**, with this final set of equations:

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

As with ordinary polar coordinates, you can impose the restriction that $\rho \geq 0$; if you also use $r \geq 0$, then you can impose $0 \leq \varphi \leq \pi$, which is especially convenient. In particular, you can calculate ρ and φ with

$$\begin{aligned} \rho &= \sqrt{r^2 + z^2} = \sqrt{x^2 + y^2 + z^2}, \\ \varphi &= \begin{cases} \arccos(z/\rho) & \text{for } \rho \neq 0, \\ 0 & \text{for } \rho = 0; \end{cases} \end{aligned}$$

although that last line is just an arbitrary default choice. Note that we still recover θ from x and y as before, but if you find ρ and φ first (instead of r), then you can use

$$\theta = \begin{cases} \arccos(x\csc\varphi/\rho) & \text{for } y \geq 0,\ 0 < \varphi < \pi, \\ 2\pi - \arccos(x\csc\varphi/\rho) & \text{for } y < 0,\ 0 < \varphi < \pi, \\ 0 & \text{for } \varphi = 0, \pi. \end{cases}$$

But again, if you are given spherical coordinates from someone else, then you can't assume that they follow these restrictions! That said, $\rho \geq 0$ and $0 \leq \varphi \leq \pi$ are imposed extremely often.

To extend spherical coordinates to even higher dimensions, you can continue as before, converting the next rectangular coordinate and the most recently created radial coordinate to polar coordinates; but we won't need that in this course, so you can skip this paragraph if you want. Specifically, give $\mathbb{R}^n$ the rectangular coordinates $x_1, x_2, \ldots, x_n$, and begin with $x_1 = r_1\cos\theta_1$ and $x_2 = r_1\sin\theta_1$. (Notice that if $x = x_1$ and $y = x_2$, then $r = r_1$ and $\theta = \theta_1$.) Next, we have $x_3 = r_2\cos\theta_2$ and $r_1 = r_2\sin\theta_2$. (Notice that if also $z = x_3$, then $\rho = r_2$ and $\varphi = \theta_2$.) Continuing, $x_4 = r_3\cos\theta_3$ and $r_2 = r_3\sin\theta_3$. Then $x_5 = r_4\cos\theta_4$ and $r_3 = r_4\sin\theta_4$. And so on and so on until $x_n = r_{n-1}\cos\theta_{n-1}$ and $r_{n-2} = r_{n-1}\sin\theta_{n-1}$. At this point, we have the radial coordinate $r_{n-1} = \sqrt{x_1^2 + \cdots + x_n^2}$ (at least if we impose the restriction $r_{n-1} \geq 0$) and angular coordinates $\theta_1, \ldots, \theta_{n-1}$, for a total of n coordinates. (We can also impose $0 \leq \theta_1 < 2\pi$ and $0 \leq \theta_i \leq \pi$ for $i \geq 2$.)

While I'm on the subject, I should also warn you that different disciplines and different countries and languages use different standard symbols for the polar coordinates. It's common for φ and θ to be swapped completely (including using only φ in 2 dimensions), and *very* common to swap r and ρ, at least in 3 dimensions. It's pretty much only English-speaking mathematicians (especially American ones) 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 read another language or take a physics class!

3 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).

3.1 *The hierarchy of functions and relations*

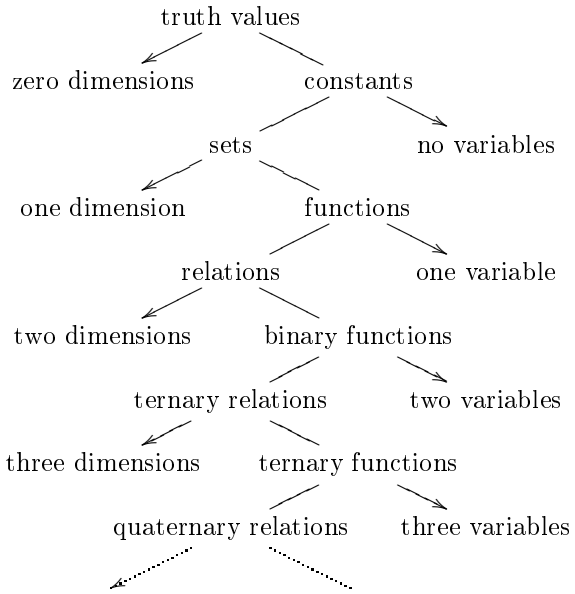
There are many different types of mathematical objects that we could study in this class. Some of them are relation-like objects:

- 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're 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. The graph of the true truth value is a solid dot, while the graph of the false truth value is a hollow circle; either way, this takes zero dimensions. 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 that in this course, but you'll do it constantly if you study Logic.

A **constant** is, in this class, a *real number*, such as -2. Any expression with no variables should evaluate to a constant (possibly undefined), and 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); given values of these variables, those values are related by the domain (or equivalently, the point whose coordinates are those values belongs to the domain) if and only if the function is defined at those values.
- The **range** of a function of any number of variables is a set (a relation with 1 variable, the output); given a value of that variable (a number), that number belongs to the range if and only if there is some point in the domain where the value of the function is that number.
- The graph of a function F 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:

$$\operatorname{gr} F = \{x_1, \dots x_n, u \mid F(x_1, \dots, x_n) = u\}.$$

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 $g(a,b) = c$, then $(a,b) \in \operatorname{dom} g$ (where $\operatorname{dom} g$ is the domain of g), $c \in \operatorname{ran} g$ (where $\operatorname{ran} g$ is the range of g), and $(a,b,c) \in \operatorname{gr} g$ (where $\operatorname{gr} g$ is the relation whose graph is the same as the graph of g). Conversely, if $(a,b,c) \in \operatorname{gr} g$, then $g(a,b) = c$, so the ternary relation $\operatorname{gr} g$ contains all of the information in the binary function g.

3.2 *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 textbook writes its definitions, but it's the way that I prefer.

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 u. 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 $(F \circ C)(t)$:

$$(F \circ C)(t) = F\Big(C(t)\Big) = F(P) = u.$$

I want to give a name to this composite function $F \circ C$; I call it F **along** C. This is not, in itself, standard terminology; however, there are associated terms that *are* standard terminology, such as the *limits* of F along C, the *derivatives* of F along C, etc; whose values are simply the limits, derivatives, etc of $F \circ C$. For a specific example, suppose $F(x,y) = 2x + 3y$ (so that $u = 2x + 3y$) and $C(t) = (\cos t, \sin t)$ (so that $x = \cos t$ and $y = \sin t$). Then $(F \circ C)(t) = 2\cos t + 3\sin t$; that is, $u = 2\cos t + 3\sin t$. In my terminology, F along C is the function from t to $2\cos t + 3\sin t$. Then the limit of F along C at 0 is

$$\lim_{t\to 0} (F \circ C)(t) = \lim_{t\to 0} (2\cos t + 3\sin t) = 2\cos 0 + 3\sin 0 = 2;$$

and the derivative of F along C at 0 is

$$(F \circ C)'(0) = \frac{\mathrm{d}}{\mathrm{d}t}(2\cos t + 3\sin t)\Big|_{t=0} = (-2\sin t + 3\sin t)\Big|_{t=0} = -2\sin 0 + 3\cos 0 = 3.$$

Sometimes instead of speaking of a limit, derivative, etc of F along C at a number t, people will speak of a limit, derivative, etc of F along C at a point P. Then you want to use a number t (if one exists) such that $C(t) = P$. For a *simple* curve, that is one that never repeats a point, this is unambiguous, because there can be only one number t such that $C(t) = P$. But in the example above, if I want (say) the limit of F along C at $(1,0)$, then I could use $t = 0$ again, but I could instead use $t = 2\pi$ or indeed $t = 2\pi k$

for any integer k, because these all give $C(t) = (1, 0)$. But perhaps surprisingly (or perhaps not), these all give the same result: $\lim_{t\to 2\pi k} (2\cos t + 3\sin t) = 2\cos(2\pi k) + 3\sin(2\pi k) = 2$. Now suppose we use a different curve that also goes through the point $(1, 0)$, say $C(t) = (t, \ln t)$, which goes through $(1, 0)$ when $t = 1$ (and only then). Then the limit of F along this C at $(1, 0)$ is $\lim_{t\to 1} F(t, \ln t) = \lim_{t\to 1} (2t + 3\ln t) = 2(1) + 3\ln(1) = 2$ again.

At this point, you might suspect that we could take *any* curve through $(1, 0)$ and the limit of F along that curve at $(1, 0)$ would be 2. But this is not true! For example, suppose $C(t) = (0, t)$ when $t < 0$ and $C(t) = (1, t)$ when $t \geq 0$; then $C(t) = (1, 0)$ when $t = 0$ (and only then). So the limit of F along C at $(1, 0)$ is $\lim_{t\to 0} F\big(C(t)\big)$ (if this exists), but $\lim_{t\to 0^-} F\big(C(t)\big) = \lim_{t\to 0^-} F(0, t) = \lim_{t\to 0^-} (2(0) + 3t) = 0$, while $\lim_{t\to 0^+} F\big(C(t)\big) = \lim_{t\to 0^+} F(1, t) = \lim_{t\to 0^+} (2(1) + 3t) = 2$, so overall, $\lim_{t\to 0} F\big(C(t)\big)$ does not exist. Or suppose $C(t) = (0, t)$ when $t \neq 0$ while $C(t) = (1, t)$ when $t = 0$; then the limit of F along C at $(1, 0)$ exists, but it's 0, not 2 like it should be.

But these examples are cheating, because the curve C is not continuous! It's easy to prove that if $C(t_0) = (1, 0)$ and C is continuous at t_0, then the limit of F along C at t_0 must be 2. To be explicit, write C as (f, g); that is, $C(t) = \big(f(t), g(t)\big)$, so that $f(t_0) = 1$, $g(t_0) = 0$, and $F\big(C(t)\big) = 2f(t) + 3g(t)$. Then to say that C is continuous at t_0 just means that f and g are continuous at t_0, and this means that

$$\lim_{t\to t_0} F\big(C(t)\big) = \lim_{t\to t_0} \big(2f(t) + 3g(t)\big) = \big(2f(t_0) + 3g(t_0)\big) = 2(1) + 3(0) = 2.$$

Even writing this all out with f and g is kind of overkill; the point is that if x is continuously approaching 1 and y is continuously approaching 0, then $2x + 3y$ must be approaching $2(1) + 3(0) = 2$, which is kind of obvious when you put it that way.

This will be taken as the definition of the statement that the limit of F at $(1, 0)$ is 2. We'll write this as $\lim_{(x,y)\to(1,0)} F(x, y) = 2$ or $\lim_{\substack{x\to 1\\ y\to 0}} F(x, y) = 2$, and secretly underneath that will be the idea that x and y are continuous functions of some other variable t, but we'll just think of it as x approaching 1 and y approaching 0 along some reasonable path, and seeing that the value of $2x + 3y$ approaches 2 as this happens.

I'll now look at several specific examples of definitions in this vein.

3.3 *Continuity*

Let F be a function of n variables, and let P_0 be a point in the domain of f. Suppose that C is a curve through P_0, that is, C is curve in $\mathbb{R}^n$ with $C(t_0) = P_0$ for some number t_0, and suppose that C is continuous at t_0. Then F is **continuous along** C at t_0 if $F \circ C$ is also continuous at t_0.

We say that F is **continuous at** P_0 if, for *every* curve C with $C(t_0) = P_0$ that's continuous at t_0, F is continuous along C at t_0. We say that F is **continuous on** a subset S of its domain if F is continuous at every point in S. Finally, we say that F is **continuous** if F is continuous on its domain, that is if F is continuous at every point in its domain.

An equivalent definition is to say that F is continuous at P_0 if F is defined at P_0 and, for each positive number ε, there is some positive number δ such that, whenever $\|P - P_0\| < \delta$ and F is defined at P, then $|F(P) - F(P_0)| < \varepsilon$. 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.)

For example, let

$$F(x, y) = \begin{cases} \dfrac{y^3}{x^2 + y^2} & \text{for } (x, y) \neq (0, 0), \\ 0 & \text{for } (x, y) = (0, 0). \end{cases}$$

Is F continuous at $(0, 0)$? To answer this, suppose that C is a continuous curve through $(0, 0)$, so that $C(t_0) = (0, 0)$ and C is continuous at t_0. More explicitly, let $C = (f, g)$, so $f(t_0) = 0$, $g(t_0) = 0$, and f and g are continuous at t_0. Then

$$F\big(C(t)\big) = \begin{cases} \dfrac{g(t)^3}{f(t)^2 + g(t)^2} & \text{for } C(t) \neq (0, 0), \\ 0 & \text{for } C(t) = (0, 0). \end{cases}$$

So the question is whether this is continuous at $t = t_0$. Since f and g are continuous, we could try plugging in $t = t_0$ to see what the limit is, but this gives 0/0. You might try L'Hôpital's Rule, but we don't know that f and g are differentiable (and besides, L'Hôpital's Rule rarely helps in multiple dimensions). Instead, change perspective; for each value of t, let $r(t)$ and $\theta(t)$ be polar coordinates that give the point $\big(f(t), g(t)\big)$. We can pick $r(t)$ to be $\sqrt{f(t)^2 + g(t)^2}$, so that r is also continuous at t_0 and $r(t_0) = 0$. We *cannot* assume that θ is continuous at t_0, because any formula for θ is piecewise-defined, but we won't need it to be. Then

$$F\big(C(t)\big) = \frac{g(t)^3}{f(t)^2 + g(t)^2} = \frac{r(t)^3 \sin^3\theta(t)}{r(t)^2\cos^2\theta(t) + r(t)^2\sin^2\theta(t)} = \frac{r(t)\sin^3\theta(t)}{\cos^2\theta(t) + \sin^2\theta(t)} = r(t)\sin^3\theta(t)$$

when $r(t) \neq 0$, and $F\big(C(t)\big) = 0$ when $r(t) = 0$. Now we can use the Squeeze Theorem (aka the Sandwich Theorem); since $-1 \leq \sin^3\theta(t) \leq 1$, we have $-r(t) \leq F\big(C(t)\big) \leq r(t)$, and the limit of these as $t \to t_0$ is $r(t_0) = 0$. Therefore, $\lim_{t\to t_0} F\big(C(t)\big) = 0 = F\big(C(t_0)\big)$, which means that $F \circ C$ is continuous at t_0. Since this works for every appropriate curve C, we conclude that F is continuous at $(0, 0)$.

But the paragraph above is really too pedantic. We can calculate this more easily by just talking about x and y. Introduce polar coordinates r and θ directly. (Technically, these aren't the same as the r and θ in the previous paragraph; those were functions whose outputs at t are the r and θ from this paragraph.) Then

$$F(x, y) = \frac{y^3}{x^2 + y^2} = \frac{r^3\sin^3\theta}{r^2\cos^2\theta + r^2\sin^2\theta} = \frac{r\sin^3\theta}{\cos^2\theta + \sin^2\theta} = r\sin^3\theta$$

when $r \neq 0$, while $F(x, y) = 0$ when $r = 0$. Then since $r \to 0$ as $(x, y) \to (0, 0)$ and $-1 \leq \sin^3\theta \leq 1$, we conclude that $F(x, y) \to 0$ as $(x, y) \to (0, 0)$, and since $F(0, 0) = 0$ too, we conclude that F is continuous at $(0, 0)$. Secretly underneath all of that, x and y (and so also r and θ) are functions of some independent parameter t, but we don't need to think about that explicitly.

For another example, consider

$$F(x, y) = \begin{cases} \dfrac{y^2}{x^2 + y^2} & \text{for } (x, y) \neq (0, 0), \\ 0 & \text{for } (x, y) = (0, 0). \end{cases}$$

This is very similar to the previous example, but if we try the same trick, then we end up with $F(x, y) = \sin^2\theta$, and this could take any value between 0 and 1. It looks like the limit doesn't exist and the function isn't continuous. But to be sure, we should find a specific curve along which F is not continuous; maybe something where $\sin^2\theta = 1$, which would mean a vertical line where $x = 0$. So let $C(t) = (0, t)$, so that $x = 0$ and $y = t$, with $t_0 = 0$ so that $C(t_0) = (0, 0)$. Then along this curve,

$$F\big(C(t)\big) = \begin{cases} \dfrac{t^2}{0^2 + t^2} = 1 & \text{for } t \neq 0, \\ 0 & \text{for } t = 0, \end{cases}$$

so F is not continuous along C at 0. Notice that if we'd chosen $C(t) = (t, 0)$ instead, then we'd have found that F *is* continuous along that curve, but that's not enough; if there's even a single continuous curve that F is not continuous along, then F is not continuous, no matter how many other curves it is continuous along.

If a function has a formula built out of the coordinate variables using only the usual operations, then it's continuous, that is continuous wherever it's 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, when defined in the real-number system. Some notable operations *not* on this list are piecewise definitions

and powers where the exponent varies and the base may be zero or negative. You'll notice that all of the tricky examples are piecewise-defined; this is why.)

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 $G\circ C$ 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:

$$\Big((F\circ C)+(G\circ C)\Big)(t) = (F\circ C)(t)+(G\circ C)(t) = F\Big(C(t)\Big)+G\Big(C(t)\Big), \text{ and}$$
$$\Big((F+G)\circ C\Big)(t) = (F+G)\Big(C(t)\Big) = F\Big(C(t)\Big)+G\Big(C(t)\Big).$$

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.)

3.4 *Limits*

I sort of defined limits in Section 3.2 and referred to them in the Section 3.3, but the general notion of limit in multiple dimensions is a little more complicated than in one-variable Calculus. To keep things simple otherwise, we'll only look at finite limits approaching a finite value; none of our limits will involve infinity in any role. Even so, things will be complicatd enough; where one-variable Calculus has both two-sided and one-sided limits (that is $x\to a$ and $x\to a^+$ or $x\to a^-$), multivariable Calculus has any number of directions from which to approach a point.

To handle this correctly, we need to deal with a technicality about limits that's often ignored in one-variable Calculus: the function whose limit you're taking must be defined at numbers arbitrarily close to the number that the limit 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 1 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 F approaching P_0 should be the limit of F along any curve through P_0. But 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 equal P_0 except when $t = t_0$. Finally, we only care what happens if the curve is continuous as it goes through P_0, since otherwise the limit could jump around because of what C is doing rather than what F is doing. So we're only looking at certain curves that are *appropriate* to the problem (not standard terminology), which is why the definition has all of this fine print. 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's 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 and P_0 is a limit point of $\operatorname{dom} F$, 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, $x \to 0^+$; that is, $x \to 0$ while $x > 0$.) 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 an undefined value if $x = y$ instead (so F is a partially-defined function, that is a piecewise-defined function with only one piece).

3.5 *Differentiability*

Given a function F of n variables, a parametrized curve C in $\mathbb{R}^n$, and a number t_0 in the domain of C, then following the general principle, the **derivative** of F **along** C **at** t_0 is $(F \circ C)'(t_0)$, if this exists. Of course, we wouldn't expect this to exist unless C is differentiable at t_0. So we might say that F is differentiable at a point P_0 if, whenever C is a differentiable curve with $C(t_0) = P_0$, the derivative of F along C at t_0 exists. But in fact, this is *not* enough!

The reason is that we want to get our hands on the derivative (of F at P_0) itself. Now, the way that differentiability fits in with composition of functions is the Chain Rule

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

If we replace f with the multivariable function F and replace g with the curve C, so that the values of the derivative of this ($C'(t)$, replacing $g'(t)$) are vectors, then what will $f'\big(g(t)\big)$ be replaced with? Since C' is a vector-valued function and the composite is an ordinary scalar-valued function, the derivative of F

should multiply by a vector to get a scalar. The general 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 geometrically even when lengths and angles don't apply, this vector is going to have to be a *row* vector; see Section 1.8.) There are actually several sorts of derivatives in higher dimensions, and we'll come back to this subject in Chapter 4; but the one which is a vector will provide the definition of differentiability.

So, we say that the function F is **differentiable** at the point P_0 if there exists a (row) vector $\boldsymbol{v}$ such that, whenever C is a parametrized curve, $C(t_0) = P_0$ for some number t_0, and C is differentiable at t_0, then $F \circ C$ is also differentiable at t_0 and furthermore $(F \circ C)'(t_0) = \boldsymbol{v} \cdot C'(t_0)$. In other words, F is differentiable at P_0 if (and only if) there is a vector $\boldsymbol{v}$ such that the derivative of F along C is the dot product of $\boldsymbol{v}$ with the velocity vector C'. As with continuity, we can talk about differentiability on a set instead of only at a point, and F is simply *differentiable* if F is differentiable at every point in its domain.

This vector $\boldsymbol{v}$ 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

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

Again, the left-hand side here is the derivative of F along C, while the right-hand side is now the dot product of the gradient of F and the velocity of C.

For example, suppose that $F(x, y) = 2x^2 + 3xy + 4y^2$, and let P_0 be $(5, 6)$. If $C(t) = (f(t), g(t))$, with $f(t_0) = 5$, $g(t_0) = 6$, and both f and g differentiable at t_0, then $(F \circ C)(t) = 2f(t)^2 + 3f(t)g(t) + 4g(t)^2$, so the derivative of F along C at t_0 is

$$\begin{aligned}
(F \circ C)'(t_0) &= 2\Big(2f(t_0)f'(t_0)\Big) + 3\Big(f'(t_0)g(t_0) + f(t_0)g'(t_0)\Big) + 4\Big(2g(t_0)g'(t_0)\Big) \\
&= 4f(t_0)f'(t_0) + 3g(t_0)f'(t_0) + 3f(t_0)g'(t_0) + 8g(t_0)g'(t_0) \\
&= \Big\langle 4f(t_0) + 3g(t_0), 3f(t_0) + 8g(t_0)\Big\rangle \cdot \langle f'(t_0), g'(t_0)\rangle \\
&= \langle 4(5) + 3(6), 3(5) + 8(6)\rangle \cdot C'(t_0) = \langle 38, 63\rangle \cdot C'(t_0).
\end{aligned}$$

Since $f'(t_0)$ and $g'(t_0)$ exist, this exists; but that's not enough, we also need to be able to express this as a dot product with $C'(t_0)$; and even that's not enough, but the vector that we're taking the dot product with should be the same for every curve. Fortunately it's always $\langle 38, 63\rangle$, and so F is differentiable at $(5, 6)$ with $\nabla F(5, 6) = \langle 38, 63\rangle$.

As usual, this level of detail in the calculation is overkill, but I'll look at how to calculate it efficiently in Chapter 4.

3.6 *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 in Chapter 4.) 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.)

A useful set of examples to keep in mind (adapted from one-variable Calculus) are

$$F_n(x, y) = \begin{cases} (x^2 + y^2)^{n/2} \sin\left(\dfrac{1}{\sqrt{x^2 + y^2}}\right) & \text{for } (x, y) \neq (0, 0), \\ 0 & \text{for } (x, y) = (0, 0). \end{cases}$$

Then f_0 is not even continuous at $(0, 0)$, while f_1 is continuous but not differentiable there, f_2 is differentiable but not continuously differentiable there, f_3 is continuously differentiable but not twice differentiable there, f_4 is twice differentiable but not continuously twice differentiable there, and so on. (But these results are not supposed to be obvious; it takes some work to calculate them.)

When differentiability goes on forever, the function is **infinitely differentiable**: it is differentiable, its partial derivatives are differentiable, their partial derivatives are differentiable, etc. For example, all of the functions in the list above are infinitely differentiable everywhere *except* at $(0, 0)$. In fact, any function built out of the usual operations* (which these examples are not, because they're piecewise-defined, but they are if you leave them undefined at $(0, 0)$) is infinitely differentiable except possibly 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.

* See the list on the bottom of page 33.

4 Differentiation

I already discussed differentiation in the previous chapter, but this chapter will go into more detail. But first, I need to discuss the sort of thing that you get as a result of differentiation: differential forms.

4.1 *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.

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 in Section 4.7.)

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 an infinitesimal change in 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 themselves and the ones of order 0 (the ordinary quantities with no differentials in them), the ones most likely to appear in a real problem are $\mathrm{d}x + \mathrm{d}y$ and $3\,\mathrm{d}x + x^2\,\mathrm{d}y + \mathrm{e}^y\,\mathrm{d}z$, which are called *linear differential forms*. 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, starting in Chapter 6; this chapter is only about differential forms of rank 1, or 1-forms for short. (Ordinary quantities count as rank 0, or 0-forms.)

4.2 *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 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 1-form, however, 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

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

(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 have small absolute values as in this example, 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 P.)

(To evaluate higher-order differential forms, which are 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.)

4.3 *Linear differential forms as vectors*

Often a differential form is a sum of terms (possibly only one term, or none if it's 0), each of which is an ordinary expression multiplied by a differential (or something algebraically equivalent to this). These are called *linear differential forms* (which are also called *exterior differential 1-forms*).

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

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

(We can recover $\boldsymbol{V}$ 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 $\boldsymbol{V}$ rather than on α as the object of study. In this case, we need to think of $\boldsymbol{V}$ as a *row* vector. Regardless of whether $\boldsymbol{V}$ 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 $\alpha = \boldsymbol{V} \cdot \mathrm{d}P$ that matters. It's more common to see $\boldsymbol{V} \cdot \mathrm{d}\boldsymbol{r}$, where as usual the vector $\boldsymbol{r} = P - \mathrm{O}$ (where O is $(0, 0)$ or $(0, 0, 0)$) satisfies $\mathrm{d}\boldsymbol{r} = \mathrm{d}P$. Sometimes $\boldsymbol{V} \cdot \mathrm{d}\boldsymbol{r}$ 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 $\boldsymbol{V} \cdot \mathrm{d}\boldsymbol{r}$ (even in the textbook), so it can be nice to know what it means!

In the textbook, they also sometimes write $\mathrm{d}\boldsymbol{r} = \boldsymbol{T}\,\mathrm{d}s$, where $\mathrm{d}s$ (which is not really the differential of anything in space as a whole) is the magnitude $\mathrm{d}s = \|\mathrm{d}\boldsymbol{r}\|$ and $\boldsymbol{T} = \widehat{\mathrm{d}\boldsymbol{r}}$, the unit vector in the direction of $\mathrm{d}\boldsymbol{r}$. 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 $\boldsymbol{V} = \langle 3, x^2\rangle$, then

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

(This requires that changes in x and y make sense as having a geometric length even when $\boldsymbol{V}$ 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 $\boldsymbol{u} \times \boldsymbol{v} = \boldsymbol{u} \cdot \times\boldsymbol{v}$, then you can write $\boldsymbol{V} \times \mathrm{d}\boldsymbol{r}$ as $\boldsymbol{V} \cdot \times\mathrm{d}\boldsymbol{r}$; the book sometimes writes this as $\boldsymbol{V} \cdot \boldsymbol{n}\,\mathrm{d}s$, where $\mathrm{d}s = \|\times\mathrm{d}\boldsymbol{r}\| = \|\mathrm{d}\boldsymbol{r}\|$ again, and now $\boldsymbol{n} = \widehat{\times\mathrm{d}\boldsymbol{r}} = \times\boldsymbol{T}$ is the direction perpendicular and clockwise from $\mathrm{d}\boldsymbol{r}$. Again, sometimes $\boldsymbol{n}$ and $\mathrm{d}s$ are useful when thinking about the geometry, but you don't need them for doing calculations.

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

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

and

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

So in Section 15.2 of the textbook, 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 $\boldsymbol{F}$ by integrating one of the two differential forms above (usually the first one). But even if you're *not* doing integrals (which we will not be doing for a while), the relationship between vector fields and differential forms is helpful for geometric visualization.

4.4 *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),

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

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 Root Rule: $\mathrm{d}(\sqrt[m]{u}) = \dfrac{\sqrt[m]{u}\,\mathrm{d}u}{mu}$ if m is constant.
- The Exponentiation Rule: $\mathrm{d}(\exp u) = \exp u\,\mathrm{d}u$ (where $\exp u$ is the same as 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: $\mathrm{d}(\arcsin u) = \dfrac{\mathrm{d}u}{\sqrt{1-u^2}}$ (where $\arcsin u$ is the same as $\sin^{-1} u$).
- The Arccosine Rule: $\mathrm{d}(\arccos u) = -\dfrac{\mathrm{d}u}{\sqrt{1-u^2}}$.
- The Arctangent Rule: $\mathrm{d}(\arctan u) = \dfrac{\mathrm{d}u}{u^2+1}$.
- The Arccotangent Rule: $\mathrm{d}(\operatorname{arccot} u) = -\dfrac{\mathrm{d}u}{u^2+1}$.
- The Arcsecant Rule: $\mathrm{d}(\operatorname{arcsec} u) = \dfrac{\mathrm{d}u}{|u|\sqrt{u^2-1}}$.
- The Arccosecant Rule: $\mathrm{d}(\operatorname{arccsc} u) = -\dfrac{\mathrm{d}u}{|u|\sqrt{u^2-1}}$.
- The Chain Rule: $\mathrm{d}\big(f(u)\big) = f'(u)\,\mathrm{d}u$ 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.)

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 $2x\,\mathrm{d}x$).

Notice that the result is a linear differential form, and indeed it can be written as a dot product with a vector field: $\big\langle 2xy, x^2, 2z\cos(z^2)\big\rangle \cdot \langle \mathrm{d}x, \mathrm{d}y, \mathrm{d}z\rangle$. This is because every rule of differentiation turns a differential into a linear differential form in the same quantities, which ultimately comes out to a linear differential form in the independent variables. If this is not how your answer ends up, then something went wrong!

4.5 *Partial derivatives*

If $f(x,y,z)$ (for example) can be expressed using the usual operations (so that we have rules for differentiating them), then its differential will come out as a linear differential form

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

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 $\mathrm{D}_1 f$, $\mathrm{D}_2 f$, and $\mathrm{D}_3 f$ (respectively); compare

the notation $\mathrm{D}f$ for f' that is sometimes used in one-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 in the last section), so

$$\begin{aligned}
\mathrm{D}_1 f(x,y,z) &= 2xy, \\
\mathrm{D}_2 f(x,y,z) &= x^2, \text{ and} \\
\mathrm{D}_3 f(x,y,z) &= 2z\cos(z^2).
\end{aligned}$$

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', but usually thought of as a curly 'd'. 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 $\left(\frac{\partial u}{\partial x}\right)_{y,z}$ 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 $\frac{\partial u}{\partial x}$, 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 $\mathrm{D}_x f$, f_x, $\frac{\partial f}{\partial x}$, and so on, as well as u_x, u_1, $\mathrm{D}_1 u$, 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.

The partial derivatives of f can also be viewed as the derivatives of f along certain curves (in fact straight lines), in the sense of Section 3.5. Specifically, $\mathrm{D}_1 f(x_0,y_0,z_0)$ is the derivative of f along the curve $(x,y,z) = (t,y_0,z_0)$ (so y and z are constant) at $t = x_0$. Similarly, $\mathrm{D}_2 f(x_0,y_0,z_0)$ is the derivative of f along the curve $(x,y,z) = (x_0,t,z_0)$ at $t = y_0$; and so on.

It's possible that these individual derivatives may exist even if f is not actually differentiable (because the derivatives along other curves might not exist, or might exist but not be given by a dot product). An example is $f(x,y) = \frac{y^3}{x^2+y^2}$ with $f(0,0) = 0$ to make it continuous. Because this is piecewise-defined, we can't differentiate it following the usual rules. Still, $\mathrm{D}_1 f(0,0)$ exists as the derivative of $\frac{0^3}{t^2+0^2}$ at $t = 0$,

which is 0, and $\mathrm{D}_2 f(0,0)$ exists as the deriavtive of $\frac{t^3}{0^2+t^2}$ at $t=0$, which is 1. If f were differentiable, then the derivative of f along any other differentiable curve C with $C(t_0)=(0,0)$ would be $\langle 0,1\rangle \cdot C'(t_0)$. However, if we differentiate along $C(t)=(t,t)$, so that $C(0)=(0,0)$ and $C'(0)=\langle 1,1\rangle$, tnen in fact we get the derivative of $\frac{t^3}{t^2+t^2}$ at $t=0$, which is $1/2$, even though $\langle 0,1\rangle \cdot \langle 1,1\rangle = 1$ instead. So this is not differentiable at $(0,0)$, even though the partial derivatives exist there.

However, there is a theorem that if the partial derivatives exist *and are continuous*, then the function must be differentiable. That doesn't apply in this situation, because we have (for example) $\mathrm{D}_1 f(x,y) = -\frac{2xy^3}{(x^2+y^2)^2}$ when $(x,y)\neq(0,0)$, and this doesn't have a limit as $(x,y)\to(0,0)$. But it does apply in many other cases.

4.6 *Higher-order partial derivatives*

We can continue to differentiate the partial derivatives, as described earlier in Section 3.6. If the partial derivatives are themselves differentiable, then their partial derivatives are the original function's *second partial derivatives*.

For example, if $u=f(x,y,z)=x^2y+\sin(z^2)$ again, then we saw in the last section that $\partial u/\partial x = \mathrm{D}_1 f(x,y,z) = 2xy$, $\partial u/\partial y = \mathrm{D}_2 f(x,y,z) = x^2$ and $\partial u/\partial z = \mathrm{D}_3 f(x,y,z) = 2z\cos(z^2)$. Then differentiating these in their turn,

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

The first of this gives

$$\frac{\partial^2 u}{\partial x^2} = \mathrm{D}_{1,1} f(x,y,z) = 2y,\ \frac{\partial^2 u}{\partial y\,\partial x} = \mathrm{D}_{1,2} f(x,y,z) = 2x, \text{ and } \frac{\partial^2 u}{\partial z\,\partial x} = \mathrm{D}_{1,3} f(x,y,z) = 0.$$

Then the middle one gives

$$\frac{\partial^2 u}{\partial x\,\partial y} = \mathrm{D}_{2,1} f(x,y,z) = 2x,\ \frac{\partial^2 u}{\partial y^2} = \mathrm{D}_{2,2} f(x,y,z) = 0, \text{ and } \frac{\partial^2 u}{\partial z\,\partial y} = \mathrm{D}_{2,3} f(x,y,z) = 0.$$

Finally, the last one gives

$$\frac{\partial^2 u}{\partial x\,\partial z} = \mathrm{D}_{2,1} f(x,y,z) = 0,\ \frac{\partial^2 u}{\partial y\,\partial z} = \mathrm{D}_{2,2} f(x,y,z) = 0, \text{ and } \frac{\partial^2 u}{\partial z^2} = \mathrm{D}_{2,3} f(x,y,z) = 2\cos(z^2) - 2z^2\sin(z^2).$$

Notice that in the ∂/∂ notation, the variables appear in the reverse order that they're being differentiated with respect to, so that $\frac{\partial^2 u}{\partial x\,\partial y}$ means $\frac{\partial}{\partial x}\frac{\partial u}{\partial y}$, while in the D notation, the numbers appear in order, so that $\mathrm{D}_{1,2} f$ means $\mathrm{D}_2\mathrm{D}_1 f$.

But does the order really matter? Look at these examples; even ignoring the ones that are 0, see how $\partial^2 u/\partial x\,\partial y = \partial^2 u/\partial y\,\partial x$, or equivalently how $\mathrm{D}_{1,2} f = \mathrm{D}_{2,1} f$. This is no coincidence! It's a theorem that, as long as a function is twice differentiable, then these *mixed second-partial derivatives* will be equal. There are examples, however, such as $f(x,y) = xy(x^2-y^2)/(x^2+y^2)$ with $f(0,0)=0$, where the second partial derivatives (defined as derivatives along straight-line curves of the first partial derivatives) exist but are not equal. Nevertheless, if the second partial derivatives are continuous, then (by the theorem cited at the end of the last section) the first partial derivatives must be differentiable, so the mixed partial derivatives must be equal after all.

Besides higher-order partial derivatives, it's also possible to consider higher-order differentials such as $\mathrm{d}^2 u$. The differential of order n contains all of the information about all of the partial derivatives up to order n. We won't get into that in this course, although there was a bit in the one-variable context in Section 2.6.

4.7 *Defining differentials*

Recall from Section 3.5 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 grad f for ∇f.)

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

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

where $\boldsymbol{r}$ is $P - \mathrm{O}$ (that is 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 in Section 4.4 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 of differentiability.

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

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

when $u = f(P)$. This quantity $\nabla f(P_0) \cdot \boldsymbol{v}_0$ is called the **derivative** of f at P_0 along $\boldsymbol{v}_0$. If $\boldsymbol{v}_0$ happens to be a unit vector (a *direction*), then $\nabla f(P_0) \cdot \boldsymbol{v}_0$ is also called the **directional derivative** of f at P_0 in the direction of $\boldsymbol{v}_0$. In general, the directional derivative in the direction of $\boldsymbol{v}_0$ is $\nabla f(P_0) \cdot \hat{\boldsymbol{v}}_0$ (where $\hat{\boldsymbol{v}} = \boldsymbol{v}/\|\boldsymbol{v}\|$ is the unit vector in the direction of $\boldsymbol{v}$); however, be careful, because some people use the term 'directional derivative' for $\nabla f(P_0) \cdot \boldsymbol{v}_0$ in the general case. 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 $\mathrm{d}\big(f(P)\big) = \nabla f(P) \cdot \mathrm{d}P = \nabla f(P) \cdot \mathrm{d}\boldsymbol{r}$, the value of the gradient may also be written as $\nabla f(P) = \mathrm{d}\big(f(P)\big)\big/\mathrm{d}P = \mathrm{d}\big(f(P)\big)\big/\mathrm{d}\boldsymbol{r}$ (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 in Chapter 8, there will be another reason that it's convenient to have a symbol ∇ that we can manipulate more thoroughly than the tiny tick mark on f'.)

4.8 *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

$$\begin{aligned} \mathrm{d}\big(f(x,y,z)\big) &= \mathrm{D}_1 f(x,y,z)\,\mathrm{d}x + \mathrm{D}_2 f(x,y,z)\,\mathrm{d}y + \mathrm{D}_3 f(x,y,z)\,\mathrm{d}z \\ &= \langle \mathrm{D}_1 f(x,y,z), \mathrm{D}_2 f(x,y,z), \mathrm{D}_3 f(x,y,z) \rangle \cdot \langle \mathrm{d}x, \mathrm{d}y, \mathrm{d}z \rangle. \end{aligned}$$

In other words,

$$\nabla f(x,y,z) = \langle \mathrm{D}_1 f(x,y,z), \mathrm{D}_2 f(x,y,z), \mathrm{D}_3 f(x,y,z) \rangle = \left\langle \frac{\partial\big(f(x,y,z)\big)}{\partial x}, \frac{\partial\big(f(x,y,z)\big)}{\partial y}, \frac{\partial\big(f(x,y,z)\big)}{\partial z} \right\rangle.$$

Put more simply,

$$\nabla f = \langle \mathrm{D}_1 f, \mathrm{D}_2 f, \mathrm{D}_3 f \rangle,$$

or even

$$\nabla = \langle \mathrm{D}_1, \mathrm{D}_2, \mathrm{D}_3 \rangle.$$

The gradient has the same information as the differential, and the partial derivatives are the components 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 $u = f(x, y, z)$:

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

$$\mathrm{d}u = \mathrm{D}_1 f(x, y, z)\,\mathrm{d}x + \mathrm{D}_2 f(x, y, z)\,\mathrm{d}y + \mathrm{D}_3 f(x, y, z)\,\mathrm{d}z;$$

$$\mathrm{D}_1 f(x, y, z) = \left(\frac{\partial u}{\partial x}\right)_{y,z},\ \mathrm{D}_2 f(x, y, z) = \left(\frac{\partial u}{\partial y}\right)_{x,z},\ \mathrm{D}_3 f(x, y, z) = \left(\frac{\partial u}{\partial z}\right)_{x,y};$$

$$\nabla f(x, y, z) = \langle \mathrm{D}_1 f(x, y, z), \mathrm{D}_2 f(x, y, z), \mathrm{D}_3 f(x, y, z) \rangle;$$

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

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

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

4.9 *Jacobian matrices*

If you have m functions of n variables each, or equivalently a function that takes a point in n dimensions as input and returns a point in m dimensions as output, then you can put their partial derivatives into an array with m rows and n columns, that is an m-by-n *matrix* (see Section 1.13). For example, if you have 2 functions of 3 variables each, say $u = f(x, y, z)$ and $v = g(x, y, z)$ (or in other words, $(u, v) = (f, g)(x, y, z)$), then the partial derivatives fit into a 2-by-3 matrix

$$\begin{bmatrix} \partial u/\partial x & \partial u/\partial y & \partial u/\partial z \\ \partial v/\partial x & \partial v/\partial y & \partial v/\partial z \end{bmatrix};$$

we may call this matrix $\mathrm{d}(u, v)/\mathrm{d}(x, y, z)$. You can also think of this as the result of applying the matrix of functions

$$\begin{bmatrix} \mathrm{D}_1 f & \mathrm{D}_2 f & \mathrm{D}_3 f \\ \mathrm{D}_1 g & \mathrm{D}_2 g & \mathrm{D}_3 g \end{bmatrix},$$

which may be called $\mathrm{D}(f, g)$, to the point (x, y, z). That is, we have $\mathrm{d}(u, v)/\mathrm{d}(x, y, z) = \mathrm{D}(f, g)(x, y, z)$. Or writing P for (x, y, z), Q for (u, v), and F for (f, g), so that $Q = F(P)$, we have $\mathrm{d}Q/\mathrm{d}P = \mathrm{D}F(P)$, where $\mathrm{D}F$ is the same matrix of functions as before. (You could justifiably write $\mathrm{D}F(P)$ as $F'(P)$, but this is not usually done in multiple dimensions.) In particular:

- If you have an ordinary function $y = f(x)$, you can think of this as a group of only 1 function of only 1 variable each, so that $\mathrm{d}(y)/\mathrm{d}(x) = \mathrm{D}(f)(x)$ is a 1-by-1 matrix, consisting of a single entry, which is the usual derivative $\mathrm{d}y/\mathrm{d}x = f'(x)$. That is, $\mathrm{d}(y)/\mathrm{d}(x) = [\,\mathrm{d}y/\mathrm{d}x\,]$, and $\mathrm{D}(f) = [\,\mathrm{D}f\,] = [\,f'\,]$.
- If you have a parametrized curve in 3 dimensions, say $P = (x, y, z) = \big(f(t), g(t), h(t)\big)$, then this is a group of 3 functions of 1 variable each, so that $\mathrm{d}(x, y, z)/\mathrm{d}(t) = \mathrm{D}(f, g, h)(t)$ is a 3-by-1 matrix, consisting of a single column with 3 entries, which are the components of the velocity vector $\mathrm{d}P/\mathrm{d}t =$

$\langle f'(t), g'(t), h'(t)\rangle$. That is, $\mathrm{d}(x,y,z)/\mathrm{d}(t) = \begin{bmatrix} \mathrm{d}x/\mathrm{d}t \\ \mathrm{d}y/\mathrm{d}t \\ \mathrm{d}z/\mathrm{d}t \end{bmatrix}$. It is for this reason that ordinary vectors that represent change *of* a point (such as velocity vectors) are sometimes called *column vectors*.

- If you have a function of 3 variables, say $u = F(x,y,z)$, then you can think of this as a group of 1 function of 3 variables each, so that $\mathrm{d}(u)/\mathrm{d}(x,y,z) = \mathrm{D}(F)(x,y,z)$ is a 1-by-3 matrix, consisting of a single row with 3 entries, which are the components of the gradient vector $\langle \partial u/\partial x, \partial u/\partial y, \partial u/\partial z\rangle = \nabla F(x,y,z)$. That is, $\mathrm{d}(u)/\mathrm{d}(x,y,z) = [\,\partial u/\partial x \quad \partial u/\partial y \quad \partial u/\partial z\,]$, and $\mathrm{D}(F) = [\,\mathrm{D}_1 f \quad \mathrm{D}_2 f \quad \mathrm{D}_3 F\,]$. For this reason, vectors that represent change *with respect to* a point, such as gradient vectors, are sometimes called *row vectors*.

In this way, Jacobian matrices include all of the kinds of derivatives that we have seen before.

Every form of the Chain Rule in Section 13.4 of the textbook can be expressed using *matrix multiplication*. (See Section 1.13 again.) Recall that you can multiply an m-by-n matrix and an n-by-o matrix to get an m-by-o matrix. Then if a point R in m dimensions is a function of a point Q in n dimensions, which is itself a function of a point P in o dimensions, then you can multiply the m-by-n matrix $\mathrm{d}R/\mathrm{d}Q$ and the n-by-o matrix $\mathrm{d}Q/\mathrm{d}P$ to obtain the m-by-p matrix $\mathrm{d}R/\mathrm{d}P$.

In particular, if you have both a parametrized curve $(x,y,z) = (f(t), g(t), h(t))$ and a multivariable function $u = F(x,y,z)$, then composition makes u an ordinary function of t; specifically, $u = F\big(f(t), g(t), h(t)\big) = \big(F \circ (f,g,h)\big)(t)$. Recall the defining property of the gradient from Section 3.5:

$$\big(F \circ (f,g,h)\big)'(t) = \nabla F(f(t), g(t), h(t)) \cdot \langle f'(t), g'(t), h'(t)\rangle;$$

or $\mathrm{d}u/\mathrm{d}t = \langle \partial u/\partial x, \partial u/\partial y, \partial u/\partial z\rangle \cdot \langle \mathrm{d}x/\mathrm{d}t, \mathrm{d}y/\mathrm{d}t, \mathrm{d}z/\mathrm{d}t\rangle$. The same thing can be expressed using matrix multiplication as

$$\frac{\mathrm{d}(u)}{\mathrm{d}(t)} = \frac{\mathrm{d}(u)}{\mathrm{d}(x,y,z)} \frac{\mathrm{d}(x,y,z)}{\mathrm{d}(t)},$$

because a matrix row is mutliplied by a matrix column using the same method as the dot product.

Even the relationship between derivatives and differentials may be expressed using matrices. In general, if $Q = F(P)$, then the column matrix $\mathrm{d}Q$ may be obtained by mutiplying the matrix $\mathrm{d}Q/\mathrm{d}P = \mathrm{D}F(P)$ by the column matrix $\mathrm{d}P$. For example, if $(u,v) = (f,g)(x,y,z)$, then the 2-by-1 matrix $\mathrm{d}(u,v)$ (a column vector in 2 dimensions being thought of as a matrix) is the result of multiplying the 2-by-3 matrix $\mathrm{d}(u,v)/\mathrm{d}(x,y,z)$ by the 3-by-1 matrix $\mathrm{d}(x,y,z)$ (a column vector in 3 dimensions being thought of as a matrix). More explicitly,

$$\begin{bmatrix} \mathrm{d}u \\ \mathrm{d}v \end{bmatrix} = \begin{bmatrix} \partial u/\partial x & \partial u/\partial y & \partial u/\partial z \\ \partial v/\partial x & \partial v/\partial y & \partial v/\partial z \end{bmatrix} \begin{bmatrix} \mathrm{d}x \\ \mathrm{d}y \\ \mathrm{d}z \end{bmatrix}.$$

If you're only interested in one differential at a time, then you really don't need any of this, or any form of the Chain Rule. For example, in the situation in the previous paragraph, if you only want to know $\mathrm{d}u$, then you can reason that

$$\mathrm{d}u = \frac{\partial u}{\partial x}\,\mathrm{d}x + \frac{\partial u}{\partial y}\,\mathrm{d}y + \frac{\partial u}{\partial z}\,\mathrm{d}z,$$

which we've seen before. And then if x, y, and z are functions of t, then you can continue:

$$\mathrm{d}u = \frac{\partial u}{\partial x}\,\mathrm{d}x + \frac{\partial u}{\partial y}\,\mathrm{d}y + \frac{\partial u}{\partial z}\,\mathrm{d}z = \frac{\partial u}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t}\,\mathrm{d}t + \frac{\partial u}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t}\,\mathrm{d}t + \frac{\partial u}{\partial z}\frac{\mathrm{d}z}{\mathrm{d}t}\,\mathrm{d}t.$$

Therefore,

$$\frac{\mathrm{d}u}{\mathrm{d}t} = \frac{\partial u}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t} + \frac{\partial u}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t} + \frac{\partial u}{\partial z}\frac{\mathrm{d}z}{\mathrm{d}t}.$$

Every application of the Chain Rule in Section 13.4 of the textbook can also be done in this way, using no fancy rules at all. The purpose of the Jacobian matrices (and the gradient vectors and velocity vectors, for that matter) is largely simply to organize the partial derivatives into convenient tables.

4.10 *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 formula for evaluating a differential in Section 4.7, 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!

4.11 *Taylor's Theorem in several variables*

In one-variable Calculus, Taylor's Theorem is about this sequence of polynomial approximations of a given function:

$f(x) \approx f(x_0)$, a constant approximation, valid for $x \approx x_0$ if f is continuous at x_0;
$f(x) \approx f(x_0) + f'(x_0)(x - x_0)$, a linear approximation, valid for $x \approx x_0$ if f is differentiable at x_0;
$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2$,
a quadratic approximation, valid for $x \approx x_0$ if f is twice differentiable at x_0;
etc.

To make these even slicker, I'll write Δx for $x - x_0$, so:

$$f(x) \approx f(x_0);$$
$$f(x) \approx f(x_0) + f'(x_0)\,\Delta x;$$
$$f(x) \approx f(x_0) + f'(x_0)\,\Delta x + \frac{1}{2}f''(x_0)\,\Delta x^2;$$
etc.

Then the general approximation is

$$f(x) \approx \sum_{n=0}^{k} \frac{1}{n!} f^{(n)}(x_0)\,\Delta x^n,$$

a degree-k approximation, valid for $x \approx x_0$ if f is differentiable k times at x_0. (Finally, we can say

$$f(x) = \sum_{n=0}^{\infty} \frac{1}{n!} f^{(n)}(x_0)\,\Delta x^n,$$

expressing f exactly as an infinite power series, valid for $x \approx x_0$ if f is analytic at x_0, but we're not really going to need that here.)

To write down the general statement in several variables requires some advanced combinatorial notation (called *multi-index notation*) that I don't want to get into, but here are the first few statements when f is a function of 2 variables:

$f(x,y) \approx f(x_0,y_0)$, a constant approximation, valid for $(x,y) \approx (x_0,y_0)$ if f is continuous at (x_0,y_0);
$f(x,y) \approx f(x_0,y_0) + \mathrm{D}_1 f(x_0,y_0)\,\Delta x + \mathrm{D}_2 f(x_0,y_0)\,\Delta y$,
a linear approximation, valid for $(x,y) \approx (x_0,y_0)$ if f is differentiable at (x_0,y_0);
$f(x,y) \approx f(x_0,y_0) + \mathrm{D}_1 f(x_0,y_0)\,\Delta x + \mathrm{D}_2 f(x_0,y_0)\,\Delta y$
$+ \frac{1}{2}\mathrm{D}_{1,1} f(x_0,y_0)\,\Delta x^2 + \frac{1}{2}\mathrm{D}_{1,2} f(x_0,y_0)\,\Delta x\,\Delta y + \frac{1}{2}\mathrm{D}_{2,1} f(x_0,y_0)\,\Delta y\,\Delta x + \frac{1}{2}\mathrm{D}_{2,2} f(x_0,y_0)\,\Delta y^2$,
a quadratic approximation, valid for $(x,y) \approx (x_0,y_0)$ if f is twice differentiable at (x_0,y_0);
etc.

(Here, $\Delta x := x - x_0$ and $\Delta y := y - y_0$.) Written this way, it's hopefully clear that the degeree-k approximation has one term for each partial derivative up to order k. The individual statements can be made shorter if we use the equality of mixed partial derivatives:

$$f(x,y) \approx f(x_0,y_0);$$
$$f(x,y) \approx f(x_0,y_0) + \mathrm{D}_1 f(x_0,y_0)\,\Delta x + \mathrm{D}_2 f(x_0,y_0)\,\Delta y;$$
$$f(x,y) \approx f(x_0,y_0) + \mathrm{D}_1 f(x_0,y_0)\,\Delta x + \mathrm{D}_2 f(x_0,y_0)\,\Delta y + \frac{1}{2}\mathrm{D}_{1,1} f(x_0,y_0)\,\Delta x^2 + \mathrm{D}_{1,2} f(x_0,y_0)\,\Delta x\,\Delta y + \frac{1}{2}\mathrm{D}_{2,2} f(x_0,y_0)\,\Delta y^2;$$
etc.

This makes the quadratic approximation slightly shorter and, although I haven't tried to write them out, the subsequent approximations even shorter still (which is good, since they otherwise get very long indeed). But the pattern is not as obvious when it's put this way.

Now, to make Taylor's Theorem a *theorem*, we need to say something about how good these approximations are. Write R_k for the difference from the polynomial approximation to the original function (called the k-th **remainder** function):

$$\begin{aligned}
R_0(x,y) &:= f(x,y) - f(x_0,y_0);\\
R_1(x,y) &:= f(x,y) - f(x_0,y_0) - \mathrm{D}_1 f(x_0,y_0)\,\Delta x - \mathrm{D}_2 f(x_0,y_0)\,\Delta y;\\
R_2(x,y) &:= f(x,y) - f(x_0,y_0) - \mathrm{D}_1 f(x_0,y_0)\,\Delta x - \mathrm{D}_2 f(x_0,y_0)\,\Delta y\\
&\qquad - \frac{1}{2}\mathrm{D}_{1,1} f(x_0,y_0)\,\Delta x^2 - \mathrm{D}_{1,2} f(x_0,y_0)\,\Delta x\,\Delta y - \frac{1}{2}\mathrm{D}_{2,2} f(x_0,y_0)\,\Delta y^2;\\
&\text{etc.}
\end{aligned}$$

The simplest way to say that this approximation is valid for $(x,y) \approx (x_0,y_0)$ is the say that the limit of $R_k(x,y)$ is 0 as $(x,y) \to (x_0,y_0)$. But in fact we can say something stronger:

$$\lim_{(x,y)\to(x_0,y_0)} \frac{R_k(x,y)}{\|\langle \Delta x, \Delta y\rangle\|^k} = 0.$$

That is, not only does the remainder approach 0, but (except for the constant approximation where $k = 0$) it's approaching 0 so quickly that even if you divide it by the distance between (x_0,y_0) and (x,y) (which is itself approaching 0), in fact even if you divide it by that distance k times, the quotient is *still* approaching 0.

For the degree-k approximation, the function f has to be differentiable k times, but it only has to be differentiable so much (or even continuous) at the point (x_0,y_0). But most functions in practice are differentiable more often and at more than just one point, and we can get an exact expression for the remainder if f is differentiable $k+1$ times on the entire line segment between (x_0,y_0) and (x,y):

$$\begin{aligned}
R_0(x,y) &= \Delta x \int_{t=0}^{1} \mathrm{D}_1 f(x_0 + t\,\Delta x, y_0 + t\,\Delta y)\,\mathrm{d}t + \Delta y \int_{t=0}^{1} \mathrm{D}_2 f(x_0 + t\,\Delta x, y_0 + t\,\Delta y)\,\mathrm{d}t;\\
R_1(x,y) &= \Delta x^2 \int_{t=0}^{1} (1-t)\,\mathrm{D}_{1,1} f(x_0 + t\,\Delta x, y_0 + t\,\Delta y)\,\mathrm{d}t\\
&\quad + 2\,\Delta x\,\Delta y \int_{t=0}^{1} (1-t)\,\mathrm{D}_{1,2} f(x_0 + t\,\Delta x, y_0 + t\,\Delta y)\,\mathrm{d}t\\
&\quad + \Delta y^2 \int_{t=0}^{1} (1-t)\,\mathrm{D}_{2,2} f(x_0 + t\,\Delta x, y_0 + t\,\Delta y)\,\mathrm{d}t;\\
R_2(x,y) &= \frac{1}{2}\,\Delta x^3 \int_{t=0}^{1} (1-t)^2\,\mathrm{D}_{1,1,1} f(x_0 + t\,\Delta x, y_0 + t\,\Delta y)\,\mathrm{d}t\\
&\quad + \frac{3}{2}\,\Delta x^2\,\Delta y \int_{t=0}^{1} (1-t)^2\,\mathrm{D}_{1,1,2} f(x_0 + t\,\Delta x, y_0 + t\,\Delta y)\,\mathrm{d}t\\
&\quad + \frac{3}{2}\,\Delta x\,\Delta y^2 \int_{t=0}^{1} (1-t)^2\,\mathrm{D}_{1,2,2} f(x_0 + t\,\Delta x, y_0 + t\,\Delta y)\,\mathrm{d}t\\
&\quad + \frac{1}{2}\,\Delta y^3 \int_{t=0}^{1} (1-t)^2\,\mathrm{D}_{2,2,2} f(x_0 + t\,\Delta x, y_0 + t\,\Delta y)\,\mathrm{d}t;\\
&\text{etc.}
\end{aligned}$$

Unfortunately, it's pointless to evaluate the integrals that appear here; if you know the exact value of the derivatives of f at the points between (x_0,y_0) and (x,y), then you can probably just evaluate f at (x,y) 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 (x_0, y_0) and (x, y), then you can use M to get an estimate of the remainder:

$$|R_k(x,y)| \le \frac{1}{(k+1)!} M(|\Delta x| + |\Delta y|)^{k+1}.$$

That is:

$|R_0(x,y)| \le M_1(|\Delta x| + |\Delta y|)$ if $|\mathrm{D}_1 f|$ and $|\mathrm{D}_2 f|$ are never greater than M_1 between (x_0, y_0) and (x, y);

$|R_1(x,y)| \le \frac{1}{2} M_2(|\Delta x| + |\Delta y|)^2$

if $|\mathrm{D}_{1,1} f|$, $|\mathrm{D}_{1,2} f|$, and $|\mathrm{D}_{2,2} f|$ are never greater than M_2 between (x_0, y_0) and (x, y)

$|R_2(x,y)| \le \frac{1}{6} M_3(|\Delta x| + |\Delta y|)^3$

if $|\mathrm{D}_{1,1,1} f|$, $|\mathrm{D}_{1,1,2} f|$, $|\mathrm{D}_{1,2,2} f|$, and $|\mathrm{D}_{2,2,2} f|$ are never greater than M_3 between (x_0, y_0) and (x, y)

etc.

So if you want the approximation to be good to within a certain maximum tolerable error, if you have an upper bound on the next order of partial derivatives, then you can calculate how big $|\Delta x| + |\Delta y|$ can be before you can no longer trust the approximation.

4.12 *Linear approximation with vectors*

Everything in the previous section is in 2 dimensions (or 1). To see how this works in higher dimensions, we can write things as much as possible using points and vectors, although there is a limit to how far this will take us. So P might be (x, y), (x, y, z), or potentially even a point in a higher-dimensional space; P_0 is (x_0, y_0), (x_0, y_0, z_0), or so on; and $\Delta P := P - P_0$ (so that $P = P_0 + \Delta P$) is $\langle \Delta x, \Delta y \rangle$, $\langle \Delta x, \Delta y, \Delta z \rangle$, or so on. Then the approximations are:

$f(P) \approx f(P_0)$, a constant approximation, valid for $P \approx P_0$ if f is continuous at P_0;

$f(P) \approx f(P_0) + \nabla f(P_0) \cdot \Delta P$, a linear approximation, valid for $P \approx P_0$ if f is differentiable at P_0;

$f(P) \approx f(P_0) + \nabla f(P_0) \cdot \Delta P + \frac{1}{2} \Delta P \cdot \nabla^2 f(P_0) \Delta P$,

a quadratic approximation, valid for $P \approx P_0$ if f is twice differentiable at P_0;

etc.

For the quadratic approximation, I had to introduce $\nabla^2 f(P_0)$, a matrix (called the *Hessian matrix* of f at P_0) whose entries are the second partial derivatives of f at P_0. To write down the general case (to any order) involves a massive generalization of vectors and matrices called *tensors*, so I'm not going to try to do it. (But if you follow the pattern for 2 variables, you can still write down any specific order in any specific number of variables.)

If we want to write the error estimate, then we need the 1-*norm* of a vector $\boldsymbol{v}$, written $\|\boldsymbol{v}\|_1$, which is the sum of the absolute values of its components; so $\|\Delta P\|_1$ is $|\Delta x| + |\Delta y|$, $|\Delta x| + |\Delta y| + |\Delta z|$, or so on. Similarly, we can write $\|\boldsymbol{v}\|_\infty$ for the largest absolute value of any component of $\boldsymbol{v}$, called the *infinity-norm* of $\boldsymbol{v}$, and this easily generalizes to matrices (and other tensors); so $\|\nabla f\|_\infty$ is the largest absolute value of the partial derivatives of f (or more precisely the function giving the largest of these at a given point), $\|\nabla^2 f\|_\infty$ is the largest absolute value of the second partial derivatives of f, $\|\nabla^3 f\|_\infty$ is the largest absolute value of the third partial derivatives of f, etc. (The usual magnitude $\|\boldsymbol{v}\| = \|\boldsymbol{v}\|_2$ is then called the 2-norm, because these absolute values are raised to the power of 2 before they're added and then the principal root of index 2 is extracted; in general, you can consider the p-norm $\|\boldsymbol{v}\|_p$ for any positive real

number p, or even other values of p if you're sufficiently clever. This can also be generalized to matrices and other tensors, although not in a straightforward way.)

Then writing R_k for the remainder as before:

$$|R_0(P)| \leq M_1 \left\|\Delta P\right\|_1 \text{ if } \left\|\nabla f\right\|_\infty \text{ is never greater than } M_1 \text{ between } P_0 \text{ and } P;$$

$$|R_1(P)| \leq \frac{1}{2} M_2 \left\|\Delta P\right\|_1^{\,2} \text{ if } \left\|\nabla^2 f\right\|_\infty \text{ is never greater than } M_2 \text{ between } P_0 \text{ and } P;$$

$$|R_2(P)| \leq \frac{1}{6} M_3 \left\|\Delta P\right\|_1^{\,3} \text{ if } \left\|\nabla^3 f\right\|_\infty \text{ is never greater than } M_3 \text{ between } P_0 \text{ and } P;$$

etc.

In this way, you can estimate the error of the approximation (or conversely, see how close you need to get to make the error small enough) to any order in any number of variables.

The linear approximation in particular can be thought of as approximating a difference with a differential. Since $P \approx P_0$, we could think of the difference between them as a change ΔP that happens to be small, or we could think of it as a very small change $\mathrm{d}P$. To be neutral, I'll write it as $\boldsymbol{v}$. Then if $u = f(P)$, we're approximating Δu with $\mathrm{d}u$. Specifically, as P changes from P_0 to $P_0 + \boldsymbol{v}$, then u changes from $f(P_0)$ to $f(P_0 + \boldsymbol{v})$:

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

Then the constant approximation says that

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

while the linear approximation says (more precisely) that

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

So in the end, the linear approximation approximates differences with differentials; any equation that holds exactly with d must hold approximately with Δ. (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

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

and

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

4.13 *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 a fixed differentiable function f of (say) 2 variables x and y whose range of possible values you already understand (typically intervals), then $\mathrm{d}u = \mathrm{D}_1 f(x,y)\,\mathrm{d}x + \mathrm{D}_2 f(x,y)\,\mathrm{d}y$, 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 the partial derivatives of f are undefined. Another way is if both (not just one) of the 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 situation 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, 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 is undefined,
- when all partial derivatives 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 in only 2 dimensions),
- 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. (In this case, it may still be called a *supremum* or *infimum*.)

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 the legs of the triangle, in feet. (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 as 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 $2lh\,\mathrm{d}l + 2l^2\,\mathrm{d}h = 0$ (after multiplying both sides by $4l$), while $2l\,\mathrm{d}l + 2h\,\mathrm{d}h = 0$ becomes $2lh\,\mathrm{d}l + 2h^2\,\mathrm{d}h = 0$ (after 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.

Differential 1-forms (that is differential forms without the wedge product that we'll get to in Chapter 6) 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 Section 4.7 earlier) is ultimately about curves, this is a very natural operation.

5.1 *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 or 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

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

or

$$\int_C \alpha$$

for short.

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 304–306 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 in the next section 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 $C \circ \varphi$ 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 includes the entire domain of C (so that we cover the entire curve), then the theorem is that $\int_C \alpha = \int_{C \circ \varphi} \alpha$. The proof is that any Riemann sum for C is also a Riemann sum for $C \circ \varphi$; 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.

5.2 *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 each subinterval to handle the differentials, whereas 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 positive 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.

5.3 *Integrating vector fields*

If you are asked to integrate a vector field $\boldsymbol{F}$ along an oriented curve, then they really want you to integrate the differential form $\boldsymbol{F}(x,y) \cdot \langle \mathrm{d}x, \mathrm{d}y\rangle$, or more generally $\boldsymbol{F}(P) \cdot \mathrm{d}P$, where P is (x,y) or (x,y,z). If you write $\boldsymbol{r}$ for the vector $P - \mathrm{O}$ (where O is the origin $(0,0)$ or $(0,0,0)$), then $\mathrm{d}P = \mathrm{d}\boldsymbol{r}$, and this is the reason for the traditional notation $\int_C \boldsymbol{F}\cdot\mathrm{d}\boldsymbol{r}$, which is used in the textbook. (Sometimes $\mathrm{d}\boldsymbol{r}$ is written $\mathrm{d}\boldsymbol{x}$, $\mathrm{d}\boldsymbol{s}$, or $\mathrm{d}\boldsymbol{l}$ instead; conventions are inconsistent. You may also see $\int_C \boldsymbol{F}\cdot\boldsymbol{T}\,\mathrm{d}s$, where $\mathrm{d}s$ is the đs that appears in Section 5.4 and $\boldsymbol{T}$ is defined to be $\mathrm{d}\boldsymbol{r}/đs$. This is usually completely pointless; if you see $\boldsymbol{T}\,\mathrm{d}s$, just think of it as $\mathrm{d}\boldsymbol{r}$.)

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

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

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. On the other hand, 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.

5.4 *Integrating scalar fields*

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 $\|\mathrm{d}\boldsymbol{r}\|$, 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 the differential of anything.

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.

5.5 *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 $\boldsymbol{F}\times\mathrm{d}\boldsymbol{r}$ in analogy with the usual $\boldsymbol{F}\cdot\mathrm{d}\boldsymbol{r}$. The textbook sometimes writes $\boldsymbol{F}\cdot\boldsymbol{n}\,\mathrm{d}s$, where $\boldsymbol{n} = \times\boldsymbol{T}$ and $\mathrm{d}\boldsymbol{r} = \boldsymbol{T}\,đs$ (as in Section 5.3), but this just results in $\boldsymbol{F}\cdot\times\mathrm{d}\boldsymbol{r} = \boldsymbol{F}\times\mathrm{d}\boldsymbol{r}$.

This is the 2-dimensional cross product, so the result is still a scalar. Geometrically, however, it is actually a **pseudoscalar**, because its sign depends on how you orient the plane (counterclockwise as is the usual 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 $\times\boldsymbol{v}$ of a vector $\boldsymbol{v}$ 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):

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

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. On the other hand, 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 even before doing the calculation.

5.6 *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

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

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 $\nabla f(P) \cdot \mathrm{d}\boldsymbol{r}$, so

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

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 in Chapter 8.

A differential form is called **exact** if there exists a quantity u such that $\alpha = \mathrm{d}u$. Similarly, a vector field $\boldsymbol{F}$ is called **conservative** if there is a scalar field f such that $\boldsymbol{F} = \nabla f$. The connection between these is that $\boldsymbol{F}$ is conservative if and only if $\boldsymbol{F}(P) \cdot \mathrm{d}\boldsymbol{r}$ is exact. (After all, if $\boldsymbol{F} = \nabla f$, then $\boldsymbol{F}(P) \cdot \mathrm{d}\boldsymbol{r} = \mathrm{d}\big(f(P)\big)$.) 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 along a closed curve is zero, because

$$\oint_{C} \alpha = \int_{a}^{a} \mathrm{d}u = u|_{a}^{a} = u|_{a} - u|_{a} = 0.$$

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. If the undefined region is sufficiently small, then this can be handled with improper integrals or other methods, but we don't consider that in this class.)

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 indefinite 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 (its differential does not equal α) 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 Chapter 8. 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

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

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

$$\mathrm{D}_2\boldsymbol{F}_1 = \mathrm{D}_1\boldsymbol{F}_2$$

(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 dozens of specific terms like these.) But a full discussion of the reasons for this must wait until Chapter 8.

In multivariable Calculus, we can also integrate with respect to more than one variable at a time. (But in practice, you usually work out these integrals by treating each variable in turn, using the theorems from Section 6.2.)

6.1 *Notation*

If you have a relation R (between 2 variables), you can think of this as a set of ordered pairs, defining a region in the coordinate plane; if you also have a function f of 2 variables, then you can try to integrate f on R. Similarly, if R is relation between 3 variables, then you can think of this as a region in 3-dimensional space; if f is now a function of 3 variables, then you can again try to integrate f on R.

These may be written $\int_R f$, or $\int_{(x,y)\in R} f(x,y)$ (in 2 dimensions) or $\int_{(x,y,z)\in R} f(x,y,z)$ (in 3 dimensions) for more detail. But to really specify what is being integrated, the proper notation is

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

(in 2 dimensions) or

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

(in 3 dimensions), which will be explained in Section 6.5. Most people don't write all of this out, however; in particular, the textbook writes $\iint_R f(x,y)\,\mathrm{d}x\,\mathrm{d}y$ (in 2 dimensions) or $\iiint_R f(x,y,z)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$ (in 3 dimensions); the repeated integral symbols are actually unnecessary in context, but otherwise this is a simplification of the proper notation that I wrote above.

Note that in any specific example (say in 2 dimensions), the statement that $(x,y)\in R$ and the expression $f(x,y)$ will be replaced with a more explicit statement and a more explicit expression. For example, if $R = \{x, y \mid x^2 + y^2 \le 1\}$ and $f = (x, y \mapsto 2x + 3y)$ (that is, $f(x,y) = 2x + 3y$ for all x and y), then $\int_{(x,y)\in R} f(x,y)\ |\mathrm{d}x \wedge \mathrm{d}y|$ becomes

$$\int_{x^2+y^2\le 1} (2x+3y)\ |\mathrm{d}x \wedge \mathrm{d}y|,$$

and you would often write this without directly mentioning either R or f. It's also common to keep the subscript R in the notation and describe R off to the side somewhere, to avoid having a complicated inequality squeezed into a subscript.

Another notation is to write $\mathrm{d}A$ and $\mathrm{d}V$ in place of $|\mathrm{d}x \wedge \mathrm{d}y|$ and $|\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z|$ respectively, but note that these are *not* the differentials of any quantities A and V; you can write đ in place of d to avoid this misleading impression, but hardly anybody ever does that. In any case, whether you write it $|\mathrm{d}x \wedge \mathrm{d}y|$, $\mathrm{d}x\,\mathrm{d}y$, $\mathrm{d}A$, or $đA$, this part of the integrand is called the **area element**; similarly, $|\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z|$, $\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$, $\mathrm{d}V$, and $đV$ are all ways to write the **volume element**.

The textbook defines these integrals formally as a limit of Riemann sums created by dividing the region R into rectangles with horizontal and vertical sides. I prefer to define them by dividing R into triangles with sides in arbitrary directions. Either way, you tag such a partition of R so that each part (each rectangle or triangle) is tagged with a specific point, evaluate f at that point, multiply by the area of the part (since the areas of rectangles and triangles are easy to calculate), and add these up. The limit as the length of the largest side of any part goes to zero, if it exists, is the value of the Riemann integral. (I am speaking here as if we are in 2 dimensions; in 3 dimensions, replace rectangles with boxes, triangles with tetrahedrons, and areas with volumes. This can also be generalized to higher dimensions.)

The two definitions are equivalent, basically because rectangles can always be divided further into triangles by cutting them in half (and boxes can be divided into tetrahedrons by cutting them into sixths, etc). Proving equivalence in the other direction is trickier, because you can't divide triangles into rectangles (much less ones parallel to the coordinate axes as the textbook requires); however, you can divide any triangle almost completely into small rectangles, with only a small part left over, and this small leftover part becomes arbitrarily small with sufficiently small rectangles. This is enough to make the proof work when it's written out in full, but I won't get into that level of detail here.

6.2 *The Fubini theorems*

As in one-variable Calculus, it's not reasonable to evaluate an integral by looking at all possible ways to divide the region into tagged pieces, find a general formula for the Riemann sum, and take the limit of this as the largest length goes to zero. So as a practical matter, evaluating these integrals depends on these theorems:

1. The integral of a continuous function on a compact (that is closed and bounded) region always exists: $\int_D f$ exists if D is compact and f is continuous. That is, $\int_{P\in D} f(P)\,đA = \int_{(x,y)\in D} f(x,y)\,|dx\wedge dy|$ exists if D is a compact relation between 2 variables and f is a continuous function of 2 variables; and $\int_{P\in D} f(P)\,đV = \int_{(x,y,z)\in D} f(x,y,z)\,|dx\wedge dy\wedge dz|$ exists if D is a compact relation between 3 variables and f is a continuous function of 3 variables.
2. If two regions D_1 and D_2 are completely disjoint (no overlap at all), or if their overlap $D_1\cap D_2$ 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: $\int_{D_1\cup D_2} f = \int_{D_1} f + \int_{D_2} f$ if the integrals on the right-hand side exist and $D_1\cap D_2$ has a smaller dimension. That is,

$$\int_{P\in D_1\cup D_2} f(P)\,đA = \int_{P\in D_1} f(P)\,đA + \int_{P\in D_2} f(P)\,đA$$

in 2 variables (if $D_1\cap D_2$ is only 1-dimensional at most), and

$$\int_{P\in D_1\cup D_2} f(P)\,đV = \int_{P\in D_1} f(P)\,đV + \int_{P\in D_2} f(P)\,đV$$

in 3 variables (if $D_1\cap D_2$ is only 2-dimensional at most). (There's a more general version of this that we don't really need: $\int_{D_1\cup D_2} f = \int_{D_1} f + \int_{D_2} f - \int_{D_1\cap D_2} f$ if the integrals on the right-hand side exist. That way, if $D_1\cap D_2$ has a smaller dimension than the number of variables, that is if its area/volume/etc is zero, then $\int_{D_1\cap D_2} f$ is zero and can be ignored.)
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 within 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). That is, in 2 dimensions,

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

if either side exists, which means (by renaming x and y on the right-hand side) that

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

if either side exists. In other words, $đA = |dx\wedge dy| = |dy\wedge dx|$ just as well. Similarly, in 3 dimensions,

$$\begin{aligned} đV &= |dx\wedge dy\wedge dz| = |dx\wedge dz\wedge dy| = |dy\wedge dx\wedge dz| \\ &= |dy\wedge dz\wedge dx| = |dz\wedge dx\wedge dy| = |dz\wedge dy\wedge dx|; \end{aligned}$$

and so on in even higher dimensions. In summary, the variables can come in any order.
4. Given constants a and b with $a\le b$ and continuous functions g and h (each of 1 variable), if $g(x)\le h(x)$ whenever $a\le x\le b$, then we can form the region $D := \{x, y \mid a\le x\le b,\ g(x)\le y\le h(x)\}$ (sometimes called a region of *Type 1*). Then the integral of any continuous function f of 2 variables on D is the same as a corresponding *iterated integral*:

$$\int_{\substack{a\le x\le b,\\ g(x)\le y\le h(x)}} f(P)\,đA = \int_{x=a}^{b} \left(\int_{y=g(x)}^{h(x)} f(x,y)\,dy \right) dx.$$

Technically, the inner integral here is an integral along a curve (actually a straight line segment) in the (x, y)-plane, as in Chapter 5, with a constant value of x between a and b. (The grouping parentheses around the inner integral aren't really necessary, but I often include them to avoid confusion. So you work out the inner integral first with respect to y, getting a function of x that you integrate in the outer integral.) Because of Theorem 3, this works just as well for a region of Type 2, which is the same thing but with x and y swapped.

5. Given a compact region R in 2 (or more) dimensions and continuous functions g and h (each of the same number of variables as the dimension of R), if $g(P) \le h(P)$ whenever $P \in R$, then we can form the region $D := \{P, z \mid P \in R,\ g(P) \le z \le h(P)\}$ of 1 higher dimension. Then the integral of any continuous function f on D is the same as a corresponding iterated integral:

$$\int_{\substack{(x,y)\in R,\\ g(x,y)\le z\le h(x,y)}} f(x,y,z)\,đV = \int_{(x,y)\in R}\left(\int_{z=g(x,y)}^{h(x,y)} f(x,y,z)\,\mathrm{d}z\right) đA$$

(and similarly in more variables). Of course, you can then use Theorem 4 (or repeat Theorem 5 if you have more variables) to do the outer integral. (And thanks to Theorem 3, you can use any variable for the inner integral, not only z.)

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

By itself, the Fubini Theorems only work 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, just oriented however we wish. Theorem 2, at least in many cases, allows us to divide a region up into smaller regions appropriate for the Fubini Theorems; the only question is whether the integrals exist. Theorem 1 guarantees this existence for continuous functions, which are the only ones that this version of the Fubini Theorems applies to anyway.

So using these in order, if you want to integrate a continuous function over a crazy but compact region, then divide the region (if you can) into pieces of suitable shape with negligible overlap. Since the function is continuous and these smaller regions are still compact, you know that their integrals exist (Theorem 1); since the regions overlap negligibly, you can recover the answer to the original problem by adding them up (Theorem 2). Finally, to get the integrals on these small regions, think of the variables as coming in whichever order works best (Theorem 3), and use the Fubini Theorems (Theorems 4 and 5) to replace double and triple integrals with iterated one-variable integrals. Hopefully, these will be integrals that you can do!

6.3 *Systems of inequalities*

In order to set up multiple integrals, it is necessary to solve **systems of inequalities**, a topic that isn't given much attention in Algebra courses. It's also necessary to present the solutions in a specific form. For purposes of multiple integration, a system of inequalities is **solved** if it takes a form such as this:

$$\begin{aligned} a \le x \le b,\\ f(x) \le y \le g(x),\\ h(x,y) \le z \le k(x,y); \end{aligned}$$

where x, y, and z are the variables in the system, a and b are constants with $a \le b$, f and g are functions with the property that $f(x) \le g(x)$ whenever $a \le x \le b$, and h and k are functions of two variables such that $h(x, y) \le k(x, y)$ whenever $a \le x \le b$ and $f(x) \le y \le g(x)$. (In other words, as you go through the inequalities in the list, if you have values of the variables so far that make all of the inequalities true so far, then there is at least one value of the next variable that also makes the next inequality true.)

This solution corresponds to an iterated integral of the form

$$\int_{x=a}^{b}\int_{y=f(x)}^{g(x)}\int_{z=h(x,y)}^{k(x,y)} \cdots \mathrm{d}z\,\mathrm{d}y\,\mathrm{d}x.$$

Of course, there could be more or fewer than 3 variables, and they don't have to come in alphabetical order. Also, we will consider the system solved if it's broken into cases, each of which takes the form above. (This corresponds to when you must write a sum of iterated integrals.) In principle, some or all of the inequalities in a system of inequalities (and hence, typically, in its solution) could be strict, although the ones that we need will always be weak (so that the domain of integration will be closed). Similarly, one side or the other of some or all of the compound inequalities could be left out, but ours will never do this (so that the domain of integration will be bounded, at least when the functions that appear in the solution are all continuous, so that the Extreme-Value Theorem applies).

One way to solve inequalities is to turn them into equations first, then test potential solutions on each side of the solutions to the equations. (This also requires the expressions involved to be continuous, so that the Intermediate-Value Theorem applies.) For compound inequalities such as we have here, the direction of the inequality is usually straightforward. Besides that, often a domain of integration is given as bounded by certain equations rather than by inequalities, and then you have no choice but to start with the equations.

Sometimes the relevant equations will have only one solution (or even none), and you'll find the other bound (or even both) by setting the two bounds on the next line equal. For example, in the solution template above, you might find a and/or b as the solutions to $f(x) = g(x)$ rather than directly from given equations or inequalities. (Another way to think of this is that you get $a \le x \le b$ as the solution to $f(x) \le g(x)$.) Similarly, you might find f and/or g by solving $h(x, y) = k(x, y)$ for y.

For example, let's solve this system of inequalities:

$$\begin{aligned} x &\ge 0, \\ y &\ge 0, \\ z &\ge 0, \\ x + y + z &\le 1. \end{aligned}$$

I'll start at the bottom and work my way up. I could start with any variable, but to match the pattern at the beginning of this section, I'll start with z. If I start with the equations $z = 0$ and $x + y + z = 1$ (by turning the inequalities that involve z into equations), then the solutions for z are 0 and $1 - x - y$. Setting these equal and solving for y, I get $y = 1 - x$; turning the only remaining inequality involving y into an equation, I also get $y = 0$. Setting $1 - x$ and 0 equal, I get $x = 1$; turning the last remaining inequality, involving only x, into an equation, I get $x = 0$. At this point, my results look like this:

$$\begin{aligned} x &= 1, 0; \\ y &= 1 - x, 0; \\ z &= 0, 1 - x - y. \end{aligned}$$

I still need to turn these into compound inequalities. Obviously, $1 > 0$. Choosing a number in between, such as $1/2$, for x, I see that $1 - x > 0$, because $1 - x = 1/2$ when $x = 1/2$. Keeping $x = 1/2$ and choosing a number between 0 and $1/2$, such as $1/4$, for y, I see that $0 < 1 - x - y = 1/4$. Therefore, the final solution is

$$\begin{aligned} 0 &\le x \le 1, \\ 0 &\le y \le 1 - x, \\ 0 &\le z \le 1 - x - y. \end{aligned}$$

So to integate over this region, I'd set up an integral of the form

$$\int_{x=0}^{1} \int_{y=0}^{1-x} \int_{z=0}^{1-x-y} \cdots \, \mathrm{d}z \, \mathrm{d}y \, \mathrm{d}x.$$

If the previous example were given simply as the region bounded by the coordinate planes and the plane with $x + y + z = 1$, then I would have to solve it pretty much as above, with equations. However, since it was given originally as a system of inequalities, I could also have solved it using entirely inequalities and no equations. Then I would solve the inequality $x + y + z \le 1$ for z to get $z \le 1 - x - y$,

then combine this with $z \geq 0$ to get the compound inequality $0 \leq z \leq 1 - x - y$. But this can only appear in the solution when $0 \leq 1 - x - y$; solving this for y, I get $y \leq 1 - x$. Combining this with $y \geq 0$, I get $0 \leq y \leq 1 - x$. And this is only valid when $0 \leq 1 - x$, so $x \leq 1$, which combines with $x \geq 0$ to produce $0 \leq x \leq 1$. At this point, the solution is complete:

$$\begin{aligned} &0 \leq x \leq 1, \\ &0 \leq y \leq 1 - x, \\ &0 \leq z \leq 1 - x - y. \end{aligned}$$

This is faster than using equations, since I don't have to test anything to get the correct directions. But it's easy to get turned around with inequalities, so I usually treat everything as equations first and then figure out the directions of the final inequalities afterwards.

You might also start with an integral and want to turn it into a system of inequalities (perhaps because you want to rearrange the order of the variables.) You can turn it directly into a system of compound inequalities, but when you look at the inequalities that make these up, some of them are redundant. For example, suppose that you start with

$$\int_{x=0}^{1} \int_{y=0}^{1-x} \int_{z=0}^{1-x-y} \cdots \, \mathrm{d}z \, \mathrm{d}y \, \mathrm{d}x.$$

This immediately becomes the system

$$\begin{aligned} &0 \leq x \leq 1, \\ &0 \leq y \leq 1 - x, \\ &0 \leq z \leq 1 - x - y; \end{aligned}$$

this can be further broken down into

$$\begin{aligned} &x \geq 0, x \leq 1, \\ &y \geq 0, y \leq 1 - x, \\ &z \geq 0, z \leq 1 - x - y. \end{aligned}$$

However, moving from the bottom up, you can see that $0 \leq z \leq 1 - x - y$ implies $y \leq 1 - x$, and $0 \leq y \leq 1 - x$ implies $x \leq 1$. Therefore, the only inequalities directly needed are

$$x \geq 0, \ y \geq 0, \ z \geq 0, \ z \leq 1 - x - y,$$

which is pretty much the system that I began with when I worked this example the other way. (You could also work with equations rather than inequalities to spot the redundant ones, as long as you can count on the original integral being set up properly.) Now you can solve with the variables in a different order if you wish.

In this way, all of the problems setting up integrals can be solved with Algebra even if you can't get a clear picture of the region of integration on a graph.

6.4 *Change of variables in single 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.

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 $u = \arcsin x$, so that $x = \sin u$ and $\mathrm{d}x = \cos u\,\mathrm{d}u$, and calculate:

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

(Incidentally, to integrate $\cos^2 u\,\mathrm{d}u$, use 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:

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

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 Chapter 5 before.)

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 integrals along oriented curves (as in Sections 5.2 and 5.3 of these notes and Section 15.2 of the textbook), while multiple integrals are like integrals on unoriented curves (as in Sections 2.7 and 5.4 of these notes and Sections 12.3 and 15.1 of the textbook). 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 integrals on unoriented curves in higher dimensions, 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:

$$\int_{0\le x\le 1} \sqrt{1-x^2}\,|\mathrm{d}x| = \int_{x=0}^{1} \sqrt{1-x^2}\,\mathrm{d}x = \frac{\pi}{4}, \text{ and}$$
$$\int_{0\le x\le 1} \sqrt{1-x^2}\,|\mathrm{d}x| = \int_{x=1}^{0} \sqrt{1-x^2}\,(-1)\,\mathrm{d}x = \frac{\pi}{4}.$$

Here, $|\mathrm{d}x| = \mathrm{d}x$ in the first calculuation, because x is increasing from 0 to 1, while $|\mathrm{d}x| = -\mathrm{d}x$ in the next calculation, because x is decreasing from 1 to 0; the final result is the same either way, but the first way is simpler.

Then to redo the subsitution $u = \arcsin x$, 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 $u = \arcsin x$ 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

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

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 $u = \arcsin x$, 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 one-to-one and has a differentiable inverse). To make it look more like the formulas for multiple

integrals, I'll write this as $x = g(u)$; 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 $g(u) \in R$ instead of either $g^{-1}(a) \le u \le g^{-1}(b)$ or $g^{-1}(b) \le u \le g^{-1}(a)$. Then the formula is

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

This is the complete analogue of the change-of-variables formulas for double integrals that appear in Section 6.6.

6.5 *The wedge product*

There is another complication that only appears with more than one variable. At the beginning of Section 6.1, I wrote the double integral of f on R as

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

You saw in the previous section 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. (Just as the cross product may also be called the outer product, so the wedge product may also be called the *exterior product*, but the term 'wedge product' is more common.) But instead of trying to interpret the result as another vector (or a scalar), it is (in any number of variables) a differential form of higher 'rank' than the original forms. I talked about evaluating differential forms in Section 4.2; those were forms of rank 1, and they were 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. In general, the rank of the wedge product of two differential forms is the sum of the original forms' ranks; for example, since $\mathrm{d}x$ and $\mathrm{d}y$ are 1-forms (meaning that they have rank 1), $\mathrm{d}x \wedge \mathrm{d}y$ is a 2-form (meaning that it has rank $1 + 1 = 2$).

The wedge product also involves subtracting one thing from another (like the cross product); if α and β are 1-forms, P_0 is a point, and $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are vectors, then you can evaluate $\alpha \wedge \beta$ at P_0, $\boldsymbol{v}_1$, and $\boldsymbol{v}_2$:

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

In words, you first evaluate α at P_0 and $\boldsymbol{v}_1$ and evaluate β at P_0 and $\boldsymbol{v}_2$, multiply the results, then swap which vector goes with which differential form, evaluate and multiply again, and finally subtract the two products.

For example, if $\alpha = x^2\,\mathrm{d}x + xy\,\mathrm{d}y$, $\beta = y^2\,\mathrm{d}x - xy\,\mathrm{d}y$, $P_0 = (2,3)$, $\boldsymbol{v}_1 = \langle 0.01, 0.04\rangle$, and $\boldsymbol{v}_2 = \langle -0.01, 0\rangle$, then

$$\begin{aligned}
\Big((x^2\,\mathrm{d}x + xy\,\mathrm{d}y) \wedge (y^2\,\mathrm{d}x - xy\,\mathrm{d}x)\Big)\Big|_{\substack{(x,y)=(2,3),\\ \mathrm{d}(x,y)=\langle 0.01,0.04\rangle,\langle -0.01,0\rangle}} \\
&= (x^2\,\mathrm{d}x + xy\,\mathrm{d}y)|_{\substack{(x,y)=(2,3),\\ \langle \mathrm{d}x,\mathrm{d}y\rangle=\langle 0.01,0.04\rangle}} (y^2\,\mathrm{d}x - xy\,\mathrm{d}y)|_{\substack{(x,y)=(2,3),\\ \langle \mathrm{d}x,\mathrm{d}y\rangle=\langle -0.01,0\rangle}} \\
&\quad - (x^2\,\mathrm{d}x + xy\,\mathrm{d}y)|_{\substack{(x,y)=(2,3),\\ \langle \mathrm{d}x,\mathrm{d}y\rangle=\langle -0.01,0\rangle}} (y^2\,\mathrm{d}x - xy\,\mathrm{d}y)|_{\substack{(x,y)=(2,3),\\ \langle \mathrm{d}x,\mathrm{d}y\rangle=\langle 0.01,0.04\rangle}} \\
&= \Big((2)^2(0.01) + (2)(3)(0.04)\Big)\Big((3)^2(-0.01) - (2)(3)(0)\Big) \\
&\quad - \Big((2)^2(-0.01) + (2)(3)(0)\Big)\Big((3)^2(0.01) - (2)(3)(0.04)\Big) \\
&= (0.28)(-0.09) - (-0.04)(-0.15) = -0.0312.
\end{aligned}$$

We will see below another way to evaluate this.

To see what $|\mathrm{d}x \wedge \mathrm{d}y|$ has to do with area, look at a parallelogram whose sides are given by vectors $\boldsymbol{v}_1 = \langle a, b\rangle$ and $\boldsymbol{v}_2 = \langle c, d\rangle$. If you evaluate $\mathrm{d}x \wedge \mathrm{d}y$ at any point and $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, then you get the scalar cross product of the two vectors:

$$\mathrm{d}x \wedge \mathrm{d}y|_{\langle \mathrm{d}x,\mathrm{d}y\rangle=\langle a,b\rangle,\langle c,d\rangle} = (\mathrm{d}x \wedge \mathrm{d}y)|_{\langle \mathrm{d}x,\mathrm{d}y\rangle=\langle a,b\rangle,\langle c,d\rangle} = (a)(d) - (b)(c) = ad - bc = \boldsymbol{v}_1 \times \boldsymbol{v}_2.$$

And the absolute value of this is the area of the parallelogram; that is, evaluating $|\mathrm{d}x \wedge \mathrm{d}y|$ at any point and those two vectors, really does give you the area. Ultimately, this is *why* the area element is $|\mathrm{d}x \wedge \mathrm{d}y|$. (Note that if you swap the order of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ or use $-\boldsymbol{v}_1$ and/or $-\boldsymbol{v}_2$ instead, then you'll get the same result. So this covers all of the ways to describe the parallelogram by giving vectors to represent its sides.)

A few basic properties of the wedge product follow immediately from the definition:

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

where α, β, and γ are 1-forms and u is a 0-form (that is an ordinary quantity with no differentials in it). This is because, if you evaluate each side at the same point and vectors, then you'll get the same result on both sides. So if you treat the wedge product as a kind of multiplication, 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 $P_0 = (2,3)$, $\boldsymbol{v}_1 = \langle 0.01, 0.04 \rangle$, and $\boldsymbol{v}_2 = \langle -0.01, 0 \rangle$. I get

$$\begin{aligned}(-xy(x^2+y^2)\,\mathrm{d}x \wedge \mathrm{d}y)\big|_{\substack{(x,y)=(2,3),\\ \mathrm{d}(x,y)=\langle 0.01,0.04\rangle,\langle -0.01,0\rangle}} \\ &= \Big(-xy(x^2+y^2)\Big)\Big|_{(x,y)=(2,3)} (\mathrm{d}x \wedge \mathrm{d}y)|_{\langle \mathrm{d}x,\mathrm{d}y\rangle=\langle 0.01,0.04\rangle,\langle -0.01,0\rangle} \\ &= -(2)(3)\Big((2)^2+(3)^2\Big)\Big((0.01)(0)-(0.04)(-0.01)\Big) = -0.0312,\end{aligned}$$

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 calculation in this section that is really useful for this course is the one with the asterisks on the previous page.

6.6 *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

$$\begin{aligned}\mathrm{d}x \wedge \mathrm{d}y &= (\mathrm{D}_1 g(u,v)\,\mathrm{d}u + \mathrm{D}_2 g(u,v)\,\mathrm{d}v) \wedge (\mathrm{D}_1 h(u,v)\,\mathrm{d}u + \mathrm{D}_2 h(u,v)\,\mathrm{d}v) \\ &= \mathrm{D}_1 g(u,v)\mathrm{D}_1 h(u,v)\,\mathrm{d}u \wedge \mathrm{d}u + \mathrm{D}_1 g(u,v)\mathrm{D}_2 h(u,v)\,\mathrm{d}u \wedge \mathrm{d}v \\ &\quad + \mathrm{D}_2 g(u,v)\mathrm{D}_1 h(u,v)\,\mathrm{d}v \wedge \mathrm{d}u + \mathrm{D}_2 g(u,v)\mathrm{D}_2 h(u,v)\,\mathrm{d}v \wedge \mathrm{d}v \\ &= 0 + \mathrm{D}_1 g(u,v)\mathrm{D}_2 h(u,v)\,\mathrm{d}u \wedge \mathrm{d}v - \mathrm{D}_2 g(u,v)\mathrm{D}_1 h(u,v)\,\mathrm{d}u \wedge \mathrm{d}v + 0 \\ &= \Big(\mathrm{D}_1 g(u,v)\mathrm{D}_2 h(u,v) - \mathrm{D}_2 g(u,v)\mathrm{D}_1 h(u,v)\Big)\,\mathrm{d}u \wedge \mathrm{d}v.\end{aligned}$$

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). (Notice that this is the determinant of the *Jacobian matrix* from Section 4.9. The Jacobian matrix is indicated with d/d, while the Jacobian determinant is indicated with ∂/∂.)

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

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

as long as (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, so that they're *almost* jointly one-to-one. (I'll explain this by way of example in the discussion of polar coordinates in Section 6.7.)

Besides its usual abbreviations (such as $\mathrm{d}u\,\mathrm{d}v$ for $|\mathrm{d}u\wedge\mathrm{d}v|$, the textbook's version of this formula (Theorem 3 on page 833 in Section 14.8) writes, in effect, $(u,v)\in G$ instead of $\big(g(u,v),h(u,v)\big)\in R$ for the domain of the integral on the right-hand side, where G is effectively defined to be $\left\{u,v \;\middle|\; \big(g(u,v),h(u,v)\big)\in R\right\}$. This G is called the *preimage* of R under (g,h). But this means that $(u,v)\in G$ precisely when $\big(g(u,v),h(u,v)\big)\in R$, so their integral says the same thing as mine, and there's no need to mention G explicitly. In practice, R is given by some inequalities involving x and y, and you just need to replace those two variables with $g(u,v)$ and $h(u,v)$ respectively, just like you do in the integrand $f(x,y)$.

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

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

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 (at least almost).

But you don't really need these formulas, because you can write $\mathrm{d}x$ and $\mathrm{d}y$ using u, v, $\mathrm{d}u$, and $\mathrm{d}v$, expand the wedge product $\mathrm{d}x\wedge\mathrm{d}y$ using Algebra (remembering that $\mathrm{d}u\wedge\mathrm{d}u = \mathrm{d}v\wedge\mathrm{d}v = 0$ and $\mathrm{d}v\wedge\mathrm{d}u = -\mathrm{d}u\wedge\mathrm{d}v$), and then take the absolute value to get đA; and similarly if you have more variables in more dimensions. That is how I usually do it, as you'll see in the next section.

6.7 *Polar coordinates*

Polar coordinates are a widely used example. (See Sections 2.8 and 2.10 if you need to review how these work, especially in 3 dimensions.) 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 $z = \rho\cos\varphi$ and $r = \rho\sin\varphi$) gives spherical coordinates:

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

if $r \geq 0$ and $\rho \geq 0$. (Since $r = \rho\sin\varphi$, if $r \geq 0$ and $\rho \geq 0$, then $\sin\varphi \geq 0$ too.) You can also calculate these using Jacobian determinants (and that's how the textbook does it), but as you can see, I prefer to use the wedge product.

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 (almost). The usual choices are $r \geq 0$, $0 \leq \theta \leq 2\pi$, $\rho \geq 0$, and $0 \leq \varphi \leq \pi$. (These are consistent, since $r = \rho\sin\varphi \geq 0$ when $\rho \geq 0$ and $0 \leq \varphi \leq \pi$.) If you don't use $r \geq 0$ and $\rho \geq 0$, then you need more absolute values in the formulas, and $0 \leq \varphi \leq \pi$ is the only good choice for φ (since it produces $\varphi = \arcsin(r/\rho)$ 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 $b - a = 2\pi$. In other words, if you want the map (g, h) from (r, θ) to (x, y) to be almost one-to-one, then you can't simply use $g(x, y) = r\cos\theta$ and $h(x, y) = r\sin\theta$, because that is far from one-to-one; instead, you must use $g(x, y) = r\cos\theta$ for $r \geq 0$, $0 \leq \theta \leq 2\pi$, and $h(x, y) = r\sin\theta$ for $r \geq 0$, $0 \leq \theta \leq 2\pi$, or something else with similar restrictions.

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. 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). This is what I mean when I say that the pair (g, h) of functions defining polar coordinates is *almost* one-to-one.

Therefore,

$$\int_{(x,y)\in R} f(x, y)\,|\mathrm{d}x \wedge \mathrm{d}y| = \int_{\substack{(r\cos\theta, r\sin\theta)\in R,\\ r\geq 0,\ 0\leq\theta\leq 2\pi}} f(r\cos\theta, r\sin\theta)\,r\,|\mathrm{d}r \wedge \mathrm{d}\theta|;$$

also,

$$\int_{(x,y,z)\in R} f(x, y, z)\,|\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z| = \int_{\substack{(r\cos\theta, r\sin\theta, z)\in R,\\ r\geq 0,\ 0\leq\theta\leq 2\pi}} f(r\cos\theta, r\sin\theta, z)\,r\,|\mathrm{d}z \wedge \mathrm{d}r \wedge \mathrm{d}\theta|;$$

finally,

$$\begin{aligned}\int_{(x,y,z)\in R} & f(x, y, z)\,|\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z| \\ &= \int_{\substack{(\rho\sin\varphi\cos\theta, \rho\sin\varphi\sin\theta, \rho\cos\varphi)\in R,\\ \rho\geq 0,\ 0\leq\varphi\leq\pi,\ 0\leq\theta\leq 2\pi}} f(\rho\sin\varphi\cos\theta, \rho\cos\varphi\sin\theta, \rho\cos\varphi)\,\rho^2\sin\varphi\,|\mathrm{d}\rho \wedge \mathrm{d}\varphi \wedge \mathrm{d}\theta|.\end{aligned}$$

The textbook has these same formulas (with somewhat different notation) in Sections 14.4 and 14.7, but it derives them by geometric arguments instead of by calculating them. (It does calculate them in Section 14.8, although it uses the Jacobian determinant instead of the wedge product.)

The restrictions on polar coordinates give us some default bounds on the integrals; in 2 dimensions, an iterated integral in polar coordinates usually looks like

$$\int_{\theta=0}^{2\pi}\int_{r=0}^{\cdots} \cdots r\,\mathrm{d}r\,\mathrm{d}\theta,$$

with the upper bound on r as the only bound with no default. However, this is *not always* correct, since the condition that $(r\cos\theta, r\sin\theta) \in R$ could always restrict the variables even further. Integrals in cylindrical coordinates usually look similar, except that there will be an integral with respect to z in there as well (and there are no default bounds on z). Finally, integrals in spherical coordinates usually look like

$$\int_{\theta=0}^{2\pi}\int_{\varphi=0}^{\pi}\int_{\rho=0}^{\cdots} \cdots \rho\sin\varphi\,\mathrm{d}\rho\,\mathrm{d}\varphi\,\mathrm{d}\theta,$$

with only one bound (out of a potential six) to find; but again, not always. So rather than *assuming* the default bounds, just remember that the default bounds are available if you don't get enough bounds from the description of the region R.

7 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 already covered in Chapter 6.)

Similarly, just as you can integrate a 2-dimensional vector field across a pseudooriented curve by taking a cross product with $\mathrm{d}\boldsymbol{r}$ 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 3-dimensional vector field across a pseudooriented surface by taking a dot product with $\mathrm{d}\boldsymbol{S}$ (defined in Section 7.5) 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 how to do that.

7.1 *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

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

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,

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

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 $\varphi = 0$, or when $\varphi = \pi$; 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.)

7.2 *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 oriented 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 in Section 7.6 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 (assuming that you're working in a right-handed coordinate system). So the textbook applies this right-hand rule whenever it needs an orientation but really has a pseudoorientation.

7.3 *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.

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

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

If you require the lengths of all of the sides of the triangles to all be less than δ and take the limit of all such Riemann sums as $\delta \to 0^+$, then the value of the Riemann 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.

7.4 *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 using spherical coordinates given at the beginning of Section 7.1:

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

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

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

Now I differentiate the parametrization:

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

Then

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

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

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

So, I am basically looking at

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

but I still need to think about the orientation. I really have $\mathrm{d}\varphi\wedge\mathrm{d}\theta$ rather than $\mathrm{d}\varphi\,\mathrm{d}\theta$, 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 $\mathrm{d}\theta\wedge\mathrm{d}\varphi$, or equivalently put in a minus sign wherever $\mathrm{d}\varphi\wedge\mathrm{d}\theta$ appears. So my real integral is

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

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 compact surface with a continuously differentiable parametrization, because this process will always leave you with a continuous double integral to do. In case you like formulas, if $x = f(u,v)$, $y = g(u,v)$, and $z = h(u,v)$, then

$$\begin{aligned}
\mathrm{d}x &= \frac{\partial x}{\partial u}\,\mathrm{d}u + \frac{\partial x}{\partial v}\,\mathrm{d}v = D_1 f(u,v)\,\mathrm{d}u + D_2 f(u,v)\,\mathrm{d}v,\\
\mathrm{d}y &= \frac{\partial y}{\partial u}\,\mathrm{d}u + \frac{\partial y}{\partial v}\,\mathrm{d}v = D_1 g(u,v)\,\mathrm{d}u + D_2 g(u,v)\,\mathrm{d}v, \text{ and}\\
\mathrm{d}z &= \frac{\partial z}{\partial u}\,\mathrm{d}u + \frac{\partial z}{\partial v}\,\mathrm{d}v = D_1 h(u,v)\,\mathrm{d}u + D_2 h(u,v)\,\mathrm{d}v,
\end{aligned}$$

so

$$\begin{aligned}
\mathrm{d}y \wedge \mathrm{d}z &= \left(\frac{\partial y}{\partial u}\frac{\partial z}{\partial v} - \frac{\partial y}{\partial v}\frac{\partial z}{\partial u}\right)\mathrm{d}u \wedge \mathrm{d}v = \Big(D_1 g(u,v) D_2 h(u,v) - D_2 g(u,v) D_1 h(u,v)\Big)\,\mathrm{d}u \wedge \mathrm{d}v,\\
\mathrm{d}z \wedge \mathrm{d}x &= \left(\frac{\partial z}{\partial u}\frac{\partial x}{\partial v} - \frac{\partial z}{\partial v}\frac{\partial x}{\partial u}\right)\mathrm{d}u \wedge \mathrm{d}v = \Big(D_1 h(u,v) D_2 f(u,v) - D_2 h(u,v) D_1 f(u,v)\Big)\,\mathrm{d}u \wedge \mathrm{d}v, \text{ and}\\
\mathrm{d}x \wedge \mathrm{d}y &= \left(\frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u}\right)\mathrm{d}u \wedge \mathrm{d}v = \Big(D_1 f(u,v) D_2 g(u,v) - D_2 f(u,v) D_1 g(u,v)\Big)\,\mathrm{d}u \wedge \mathrm{d}v;
\end{aligned}$$

thus, the most general exterior 2-form $U\,\mathrm{d}y \wedge \mathrm{d}z + V\,\mathrm{d}z \wedge \mathrm{d}x + W\,\mathrm{d}x \wedge \mathrm{d}y$ in 3 dimensions, where $U = F(x,y,z)$, $V = G(x,y,z)$, and $W = H(x,y,z)$, becomes

$$\begin{aligned}
&\left(U\frac{\partial y}{\partial u}\frac{\partial z}{\partial v} - U\frac{\partial y}{\partial v}\frac{\partial z}{\partial u} + V\frac{\partial z}{\partial u}\frac{\partial x}{\partial v} - V\frac{\partial z}{\partial v}\frac{\partial x}{\partial u} + W\frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - W\frac{\partial x}{\partial v}\frac{\partial y}{\partial u}\right)\mathrm{d}u \wedge \mathrm{d}v\\
&\qquad = F\Big(f(u,v), g(u,v), h(u,v)\Big)\Big(D_1 g(u,v) D_2 h(u,v) - D_2 g(u,v) D_1 h(u,v)\Big)\,\mathrm{d}u \wedge \mathrm{d}v\\
&\qquad\quad + G\Big(f(u,v), g(u,v), h(u,v)\Big)\Big(D_1 h(u,v) D_2 f(u,v) - D_2 h(u,v) D_1 f(u,v)\Big)\,\mathrm{d}u \wedge \mathrm{d}v\\
&\qquad\quad + H\Big(f(u,v), g(u,v), h(u,v)\Big)\Big(D_1 f(u,v) D_2 g(u,v) - D_2 f(u,v) D_1 g(u,v)\Big)\,\mathrm{d}u \wedge \mathrm{d}v.
\end{aligned}$$

But I prefer to calculate $\mathrm{d}x$, $\mathrm{d}y$, and $\mathrm{d}z$ directly and apply the wedge products to the results, rather than to use any of these formulas.

7.5 *Integrating vector fields*

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

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

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

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

(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 $\mathrm{d}P = \mathrm{d}\boldsymbol{r}$ by itself is not zero but rather twice something, and that something is what we mean by $đ\boldsymbol{S}$.)

The middle formula for $đ\boldsymbol{S}$ (the one without P or $\boldsymbol{r}$ in it) requires the use of the right-hand rule for the cross product. This is because $đ\boldsymbol{S}$ is really a **pseudoform**, meaning that it changes sign if you switch between right-hand and left-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 $đ\boldsymbol{S}$ or any simpler alternative; instead, it writes $\boldsymbol{n}\,đ\sigma$, or rather $\boldsymbol{n}\,\mathrm{d}\sigma$. But $đ\sigma$ is just $\|đ\boldsymbol{S}\|$, the magnitude of $đ\boldsymbol{S}$; and $\boldsymbol{n}$ is just $\widehat{đ\boldsymbol{S}}$, a unit vector in the direction of $đ\boldsymbol{S}$, that is a unit vector perpendicular to the surface pointing in the direction given by its pseudoorientation. So $\boldsymbol{n}\,\mathrm{d}\sigma$ is really just a complicated way of saying $đ\boldsymbol{S}$. To actually calculate $\boldsymbol{n}$ and $đ\sigma$ is a waste of time if $đ\boldsymbol{S}$ is all that you really want.)

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

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

on that hemisphere oriented clockwise when viewed from above, because turning the fingers of your right hand clockwise results in your thumb pointing downwards. In the example at the beginning of Section 7.4, 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 $đ\boldsymbol{S}$, and the final result would be the same.)

In case you like formulas, if $\boldsymbol{F}(x,y,z) = \langle U,V,W\rangle = \langle F(x,y,z), G(x,y,z), H(x,y,z)\rangle$, with $x = f(u,v)$, $y = g(u,v)$, and $z = h(u,v)$, then you can use the formula at the end of Section 7.4. But I prefer to remember only to use $\boldsymbol{F}(x,y,z)\cdot đ\boldsymbol{S}$, where $đ\boldsymbol{S} = \langle \mathrm{d}y\wedge\mathrm{d}z, \mathrm{d}z\wedge\mathrm{d}x, \mathrm{d}x\wedge\mathrm{d}y\rangle$, and then the rules for taking differentials and wedge products will make everything come out correctly.

7.6 *Integrating scalar fields*

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

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

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 $\mathrm{d}x\wedge\mathrm{d}y = \sin\varphi\cos\varphi\,\mathrm{d}\varphi\wedge\mathrm{d}\theta$ from the example in Section 7.4, 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 Section 7.4, so

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

and

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

Therefore, I am integrating

$$\cos\varphi\sqrt{\sin^4\varphi\cos^2\theta\,(\mathrm{d}\varphi\wedge\mathrm{d}\theta)^2+\sin^4\varphi\sin^2\theta\,(\mathrm{d}\theta\wedge\mathrm{d}\varphi)^2+\sin^2\varphi\cos^2\varphi\,(\mathrm{d}\varphi\wedge\mathrm{d}\theta)^2}$$
$$=\cos\varphi\sqrt{\sin^4\varphi\,(\mathrm{d}\varphi\wedge\mathrm{d}\theta)^2+\sin^2\varphi\cos^2\varphi\,(\mathrm{d}\varphi\wedge\mathrm{d}\theta)^2}=\cos\varphi\sqrt{\sin^2\varphi\,(\mathrm{d}\varphi\wedge\mathrm{d}\theta)^2}=\sin\varphi\cos\varphi\,|\mathrm{d}\varphi\wedge\mathrm{d}\theta|.$$

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

The value of the integral is now

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

(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 $|\mathrm{d}\varphi\wedge\mathrm{d}\theta|$ rather than $\mathrm{d}\varphi\wedge\mathrm{d}\theta$ 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

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

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

Again, in case you like formulas, if $x=f(u,v)$, $y=g(u,v)$, and $z=h(u,v)$, then

$$đ\sigma=\sqrt{\left(\frac{\partial y}{\partial u}\frac{\partial z}{\partial v}-\frac{\partial y}{\partial v}\frac{\partial z}{\partial u}\right)^2+\left(\frac{\partial z}{\partial u}\frac{\partial x}{\partial v}-\frac{\partial z}{\partial v}\frac{\partial x}{\partial u}\right)^2+\left(\frac{\partial x}{\partial u}\frac{\partial y}{\partial v}-\frac{\partial x}{\partial v}\frac{\partial y}{\partial u}\right)^2}\,|\mathrm{d}u\wedge\mathrm{d}v|$$
$$=\sqrt{\begin{pmatrix}\big(D_1g(u,v)D_2h(u,v)-D_2g(u,v)D_1h(u,v)\big)^2\\+\big(D_1h(u,v)D_2f(u,v)-D_2h(u,v)D_1f(u,v)\big)^2\\+\big(D_1f(u,v)D_2g(u,v)-D_2f(u,v)D_1g(u,v)\big)^2\end{pmatrix}}\,|\mathrm{d}u\wedge\mathrm{d}v|.$$

But again, I prefer not to use this formula.

However, in the case where x and y are the parameters, that is where $x=u$, $y=v$, and $z=h(u,v)=h(x,y)$, then this simplifies to

$$đ\sigma=\sqrt{(\partial z/\partial x)^2+(\partial z/\partial y)^2+1}\,đA=\sqrt{\|\nabla h(x,y)\|^2+1}\,đA$$

(where $đA=|\mathrm{d}x\wedge\mathrm{d}y|$), and this comes up fairly often (whenever the surface of integration is the graph of a differentiable function h).

The Stokes theorems generalize the (second) Fundamental Theorem of Calculus. The basic idea is that the integral of some differential form β on some manifold M (that is a curve, surface, etc) is equal to the integral of an antidifferential of β 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 β; one of its antidifferentials is $F(x)$ (because the differential of $F(x)$ is $\mathrm{d}\big(F(x)\big) = F'(x)\,\mathrm{d}x = f(x)\,\mathrm{d}x = \beta$). 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; 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.

8.1 *Exterior forms and exterior differentials*

You may recall from Chapter 7 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 arithmetic operations (addition etc) 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 $\langle \mathrm{d}x, \mathrm{d}y, \mathrm{d}z\rangle = \boldsymbol{v}_1$ and $\langle \mathrm{d}x, \mathrm{d}y, \mathrm{d}z\rangle = \boldsymbol{v}_2$, by the method described in Section 6.5 (although you never actually need to perform such an evaluation in this course).

You may also recall from Section 4.4 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 $\boldsymbol{v}_0$ by considering a parametrized curve through P_0 (say at some value t_0 of its parameter) with $\boldsymbol{v}_0$ as its velocity vector at t_0, 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 t of the curve, and the derivative of this function at t_0 is the result of evaluating $\mathrm{d}u$ at P_0 along $\boldsymbol{v}_0$.

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 $\boldsymbol{v}_0$, $\boldsymbol{v}_1$, ..., and $\boldsymbol{v}_p$. To evaluate this, consider a parametrized curve through P_0 (say at some value t_0 of its parameter) with $\boldsymbol{v}_0$ as its velocity vector at t_0. You can evaluate α at any point on that curve along the same p vectors $\boldsymbol{v}_1$, $\boldsymbol{v}_2$, ..., and $\boldsymbol{v}_p$ to get a number that is a function of the parameter t of the curve. The derivative of this function at t_0 is *sort*

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

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:

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

In words:

1. 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);
2. The exterior differential of one of these differentials is zero (ultimately because of the equality of mixed partial derivatives, so that the subtraction in the definition of exterior differential makes everything cancel);
3. The exterior differential obeys the Sum Rule (because every process in the definition obeys a Sum Rule);
4. 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
5. 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 (because if you trace the subtractions in the definition, you get an extra minus sign here).

(These are all special cases of more complicated rules that apply to differential forms of any rank, except that (1) really does only apply to 0-forms and (3) doesn't get any more complicated.)

Besides the Sum Rule (3), 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 $u\,\mathrm{d}v\wedge\mathrm{d}w\wedge\cdots$. A differential form that is a sum of such expressions 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's *really* useful in this course is how to perform calculations like the one above.

Notice that many of the integrals that we've studied in this course are integrals of exterior forms. In Section 5.2, we integrated linear 1-forms, which are the same as exterior 1-forms. In Section 5.3, we saw how integrating a vector field along a curve is equivalent to integrating a linear form along it. In Section 5.5, we saw how integrating a vector field across a curve is equivalent to integrating a linear pseudoform across it, which is essentially the same as integrating a linear form along the curve using the right-hand rule to orient the coordinate plane. An integral in one-variable Calculus is also the integral of a linear form along the real number line. (But in Sections 2.7 and 5.4, we integrated with respect to arclength,

which is *not* an exterior form or even somehow equivalent to one.) In Sections 7.3 and 7.4, we integrated exterior 2-forms along a surface. In Section 7.5, we saw how integrating a vector field across a surface is equivalent to integrating a pseudoform across it, which is essentially the same as integrating an exterior form along the surface using the right-hand rule to orient the coordinate space. (But in Section 7.6, we integrated with respect to surface area, which is *not* an exterior form or even somehow equivalent to one.) In Chapter 6, we integrated with respect to area and volume, which can also be thought of as pseudoforms and made equivalent to exterior forms using the right-hand rule to orient the coordinate plane or space. (This is less convenient when changing variables, because it requires us to keep track of orientation, which is why I did not treat them in this way in Chapter 6.) Even adding up the values of a function at some points is considered integrating a 0-form on those points. So we can form exterior differentials of all of these integrands and get versions of the Stokes Theorem for all of these integrals!

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 (which is probably why nobody calls it the wedge 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 Section 7.6 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 in the notation.) Another reason to include the wedges even with the differential, is that people sometimes take ordinary differentials of at least 1-forms, so that $\mathrm{d} \wedge \mathrm{d}u$ is always zero but $\mathrm{d}\mathrm{d}u$ usually is not. (In this case, $\mathrm{d}\mathrm{d}u$ can be abbreviated as d^2u, but $\mathrm{d}(u\,\mathrm{d}v)$ could be ambiguous.) 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.

8.2 *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 in any specific case), 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 is called **closed** if it has no boundary (but this is *different* from being a closed region in the sense of including all of its boundary 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 manifold 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 $\mathbb{R}^n$ 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, this can happen 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 (classifying holes) in this course; all that you really need to know are these facts:

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

The special case of this for 1-forms has already come up, in the discussion of the Fundamental Theorem of Calculus in Section 5.6.

8.3 *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 a region whose boundary may be parametrized as a curve.) 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

$$\int_{(x,y)\in C} (f(x,y)\,\mathrm{d}x + g(x,y)\,\mathrm{d}y) = \int_{(x,y)\in R} \big(D_1 g(x,y) - D_2 f(x,y)\big)\,đA.$$

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

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

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

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

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

$$\int_C \boldsymbol{F}(x,y) \cdot \mathrm{d}\boldsymbol{r} = \int_R (\nabla \times \boldsymbol{F})(x,y)\,đA,$$

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

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

On the other hand,

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

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

$$\int_C \boldsymbol{F}(x,y) \times \mathrm{d}\boldsymbol{r} = \int_R \nabla \cdot \boldsymbol{F}(x,y)\,đA,$$

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

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

These are both forms of Green's Theorem; thought of as theorems about vector fields, the first of these generalizes to Stokes's Theorem in Section 8.4 below, while the last of these generalizes to Gauss's Theorem in Section 8.5. 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 separated pieces, then we can still write Green's Theorem as

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

(or similarly for any of the versions involving scalar and vector fields), where you integrate along the boundary ∂R by integrating along each piece of that boundary and adding the integrals. But now only the outermost piece of the boundary is oriented clockwise overall; the inner pieces are oriented clockwise instead. This still matches the orientation of R, because an inner piece 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 triangular region, say

$$R = \{x, y \mid x \geq 0,\ y \geq 0,\ x + y \leq 1\}.$$

The boundary of this region is a triangle, running from $(x, y) = (0, 0)$ to $(1, 0)$ to $(0, 1)$ back to $(0, 0)$, which I will divide into three line segments, each parametrized by x or y itself. Then the integral along the boundary is

$$\begin{aligned}\int_{x=0}^{1} f(x,0)\,\mathrm{d}x + \int_{x=1}^{0} f(x,1-x)\,\mathrm{d}x + \int_{y=0}^{1} g(1-y,y)\,\mathrm{d}y + \int_{y=1}^{0} g(0,y)\,\mathrm{d}y \\ = \int_{x=0}^{1} \Big(f(x,0) - f(x,1-x)\Big)\,\mathrm{d}x + \int_{y=0}^{1} \Big(g(1-y,y) - g(0,y)\Big)\,\mathrm{d}y.\end{aligned}$$

Meanwhile, the integral on the triangular region itself is

$$\begin{aligned}\int_{y=0}^{1}\int_{x=0}^{1-y} \frac{\partial\big(g(x,y)\big)}{\partial x}\,\mathrm{d}x\,\mathrm{d}y - \int_{x=0}^{1}\int_{y=0}^{1-x} \frac{\partial\big(f(x,y)\big)}{\partial y}\,\mathrm{d}y\,\mathrm{d}x \\ = \int_{y=0}^{1} \Big(g(1-y,y) - g(0,y)\Big)\,\mathrm{d}y - \int_{x=0}^{1} \Big(f(x,1-x) - f(x,0)\Big)\,\mathrm{d}x.\end{aligned}$$

So, Green's Theorem is definitely true for this triangular region. But any triangle looks like this under an appropriate change of coordinates, and the area integral on any region R is a limit of the sum of the integrals on such triangular regions. Meanwhile, the corresponding sums of integrals on the triangles' 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, and the limit of this is the integral along the boundary itself. This is ultimately why all of these theorems apply only to exterior differential forms; there would be no cancellation for something like arclength $đs$.

(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, so I'm not going to do that.)

8.4 *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, the Curl Theorem, or some combination of these; the general Stokes Theorem is named after this particular case.)

If $\boldsymbol{F} = \langle M, N, P\rangle$ 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

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

where $\nabla\times\boldsymbol{F}$ is a vector field called the **curl** of $\boldsymbol{F}$:

$$\nabla\times\boldsymbol{F} = \langle D_1, D_2, D_3\rangle \times \langle M, N, P\rangle = \langle D_2P - D_3N, D_3M - D_1P, D_1N - D_2M\rangle;$$

or more explicitly,

$$\begin{aligned}&(\nabla\times\boldsymbol{F})(x,y,z) \\ &\quad= \left\langle \frac{\partial\big(P(x,y,z)\big)}{\partial y} - \frac{\partial\big(N(x,y,z)\big)}{\partial z}, \frac{\partial\big(M(x,y,z)\big)}{\partial z} - \frac{\partial\big(P(x,y,z)\big)}{\partial x}, \frac{\partial\big(N(x,y,z)\big)}{\partial x} - \frac{\partial\big(M(x,y,z)\big)}{\partial y} \right\rangle.\end{aligned}$$

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 includes a surface whose boundary consists of several pieces, but it won't apply to a surface whose boundary cannot be parametrized as a curve 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 a curve, 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 $P(x,y,z)$,

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

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

8.5 *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, the Divergence Theorem, or some combination.)

If $\boldsymbol{G} = \langle M,N,P\rangle$ 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

$$\int_S \boldsymbol{G}(x,y,z)\cdot\text{đ}\boldsymbol{S} = \int_D (\nabla\cdot\boldsymbol{G})(x,y,z)\,\text{đ}V,$$

where $\nabla\cdot\boldsymbol{G}$ is a scalar field called the **divergence** of $\boldsymbol{G}$:

$$\nabla\cdot\boldsymbol{G} = \langle D_1,D_2,D_3\rangle\cdot\langle M,N,P\rangle = D_1M + D_2N + D_3P;$$

or more explicitly,

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

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 includes a region whose boundary consists of several pieces, but it won't apply to a region whose boundary cannot be parametrized as a surface 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 tetrahedron splits into four pieces, three with one term each and one with three terms, so six terms grouped into three pairs, while the integral on the tetrahedron splits into three

terms whose partial integrals lead to the same six expressions. But I'm not going to write that all out like I did for Green's Theorem.

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 $P(x,y,z)$,

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

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

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

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