Collected Notes
On Vectors
May 27, 2005

Larry Susanka

Preface

This text began as sets of notes provided to students to supplement the text in several of my Math classes at Bellevue Community College. Somehow they grew to be rather more than that, but still serve that original purpose as a "Swiss Army Knife" of supplementary notes. Trigonometry students might start at Chapter One, while Calculus students could start at Chapter Three. Multivariable Calculus students would be interested in Chapters Four and Five.

I use the first ten sections of these notes when I teach our Math 120 class which is devoted primarily to Trigonometry. We don't presume these students have any experience with mathematically precise presentations, and concentrate on creating intuition about vectors, linking vector concepts to experiences students have likely had. We create a structure that can be used to deal with common problems from the Sciences. There are a number of applications but just a few topics that I emphasize in this light exposure. In two dimensions, our goal is to:

- Understand the "tip-to-tail" method of drawing the picture of a combination of vectors such as $2A - 3B$.
- Distinguish between a point in the plane with coordinates (a, b) and the vector $\langle a, b\rangle$ which, when represented as an arrow in standard position, "points at" (a, b).
- Write a vector V as $|V|\ \frac{V}{|V|}$ and thereby separate cleanly the two things that characterize a vector: **magnitude** and **direction**.
- Use dot products to calculate the angle between two vectors. Verify with a picture.
- Decompose a vector into the sum of two vectors, one of which is a multiple of a specified vector W, while the other is perpendicular to W. Both drawing the picture and the calculation are key here.
- Understand the idea of parametric constant velocity motion. Again, drawing pictures and creating equations are equally important. If a student is comfortable with linear motion he or she will find that the parametric curves in Calculus and Physics are easier to understand.

We then extend these ideas to three dimensions.

In the second chapter we expand on these ideas so we can understand planes and other surfaces in space, alternative coordinate systems and we begin working with parametric vector functions which are not constant velocity.

At this point one would want to begin exploring a computer graphing utility, such as Maple or Mathematica. I do not include instructions on how to use such a utility in the main part of the text because this type of thing is a rapidly moving target and will no doubt change (and get easier and cheaper) each year. However I

do include Maple 8 instructions to graph some of the illustrations in the text in the endnotes. These utilities are so powerful and useful that it simply makes no sense not to learn how to access that power. These and other endnotes are referenced by numbered superscript at the relevant spot in the text. Topics are often placed in these endnotes not because they are unimportant but to avoid breaking up a discussion with a side issue. The reader should at least glance at them.

A large fraction of the potential readership has had exposure to vectors before and will be able to start immediately with Chapter Three. For these folks the first two chapters should be skimmed to establish notation and provide reference points for a small library of examples and techniques.

In Chapter Three we look at a collection of topics related to curves. In this treatment curves correspond to the range of differentiable functions from the real line into the plane or space. Part of this chapter could be used in conjunction with a first course in Differential Calculus and the rest saved for a later quarter, after Integral Calculus is familiar to the reader. If all the sections are used, there is more depth in some areas than one would expect to see devoted to these topics during a typical first year Calculus sequence.

Chapters Four and Five contain ideas usually explored in Multivariable Calculus or beyond.

Chapter Four involves a collection of ideas which are in some sense "dual" to those introduced in Chapter Three. We study real valued functions from the plane or space, and their derivatives.

In Chapter Five we discuss integration involving two or three variables.

Exercises form the heart of any math book to be used by students to learn new material. Sometimes authors create hundreds of problems, many of which are nearly identical, and gather them at the end of each chapter. In this work you will encounter far fewer exercises, and these are placed in context throughout the text. Individual instructors might want to create "drill" type problem sets for several of the topics. There are good reasons for doing that. In my classes I make sure students have worked on several of each type of problem I might have on an exam by adding extra problems this way.

Every problem you find in the text has been included for some reason or other. That reason might not be obvious until considerably later. The student should try them all and think about each and every one. Don't be shocked if a problem takes a while to understand completely. I have put "stars" by the longer ones and if they are both longer and require a bit of sophistication I put two stars. You may use earlier exercises in proving later ones and I sometimes refer to these exercises in later text.

If you get stuck on a hard exercise you have two choices: persevere or accept that exercise as fact and go on. It would be a pretty rare student who would take the time to go through every single exercise in detail. But one would want to, at least, understand the statements of all the exercises before proceeding. Mathematics, in general, is not for spectators. That is particularly true in this book after Chapter Two and an inescapable feature of Chapters Four and Five.

One way of handling these problems in a classroom setting might be to assign "no star problems" as individual assignments and "one star problems" as group problems for those who like that sort of thing. If done this way a group should cast aside standard group behavior and guarantee, at least, that all of its members actually understand the statements of these problems. The "two starred problems" are, I think, best solved by motivated individuals who have either had solid Calculus classes or who are willing to go back and fill in any gaps on their own.

If you see a word you don't recognize check the index, which I have tried to make very complete. The topic might be discussed in the endnotes and referenced somewhere earlier in the text.

You will no doubt see other more sophisticated (and more complete) treatments in later courses. The first course in Linear Algebra—the one we offer here at Bellevue Community College—concentrates on the vector structure from ordered n-tuples of real numbers and real matrices. A second course in Linear Algebra would concentrate on the algebraic structure of vectors from a more abstract standpoint. In still later courses, tensors and sections of vector bundles on manifolds await you. These classes provide increasingly precise mathematical definitions and discuss more subtle properties of the ideas we begin to work with here.

In another direction, Physics and Engineering students deal with all kinds of "vector ideas" at levels of rigor tailored to each topic. In fact, many Math departments downplay vector ideas before Linear Algebra, leaving science students to learn about vectors elsewhere. Optics uses vectors—discussions of polarization and refraction are examples. An Electricity and Magnetism class would discuss vector fields and words like "curl" and "divergence" will be on every page—vector creatures! Surface tension and viscosity from Fluid Mechanics are phrased in terms of tensors. Elasticity and stress from a Solid State class are defined in terms of tensors. Discussions in Quantum Mechanics and Special Relativity use vectors throughout—not to mention General Relativity and its use of the curvature tensor and four dimensional manifolds! The list goes on and on.

Suffice it to say that anyone who hopes to understand current thinking about the physical behavior of the world at the junior or senior undergraduate level must become a "vector expert." Many of the ideas that serve as a foundation for intuition and calculation have been gathered together in these notes.

I have tried to reconcile in this text three conflicting goals: (i) I wanted to create a text so that the student would find, down the road in later classes with their particular and more complete approaches, that there were only a few "gaping holes" in proof structure, and those present were as clearly identified as possible. (ii) I wanted to include a range of tools with a lot of immediate utility in the classes for which this text could be used. (iii) I wanted the whole thing to weigh in at under two hundred pages of text.

In achieving some kind of balance certain things were sacrificed, and some might feel these choices to be a bit eccentric. To give a few examples offhand, there is no mention of a "right hand rule." Curvature is not discussed. We do not introduce differential forms. We don't talk about re-parameterization by arclength. There are only one or two problems in the text associated with each idea, rather than dozens.

In place of these missing items the reader will find my attempt to create a thorough introduction to the most basic ideas of vectors for absolute beginners in the early chapters. The later chapters build rapidly to create a rather detailed and self-contained description, including proofs often reserved for "Advanced Calculus" classes, of part of the Vector Calculus involving curves and surfaces and volumes. An instructor using this work as a supplement to a more conventional text would likely dip in and out of these later sections, picking and choosing rather than working straight through. Done this way, some thought regarding the dependency of text material on earlier exercises will be necessary.

The next steps for the student would be a Linear Algebra course followed by a more advanced Calculus class using a book such as Michael Spivak's beautiful *Calculus on Manifolds* or Walter Rudin's durable *Principles of Mathematical Analysis.* On the less rigorous but eminently practical side I like H. M. Shey's *Div Grad Curl and All That.* Those with interests in the Physical Sciences should gain, as early as possible, an understanding of Differential Equations and the Linear-Algebraic basics of Tensors.

This text, like most, is a compromise. I can only hope the reader finds it to be useful and that it conveys some sense of the elegance and beauty of the ideas which underlie it all. And, perhaps, a few will decide to dig deeper!

Have Fun!

Contents

Preface iii

Chapter I. Basic Vector Facts May 27, 2005 1
1. First Steps 2
2. Position Vectors and Constant Velocity Motion: Part 1 6
3. Decomposition of Vectors: Part 1 8
4. Vectors in the Plane 10
5. Dot Products 13
6. Problems in the Plane: Displacements, Forces and Velocity 15
7. Position Vectors and Constant Velocity Motion: Part 2 17
8. Decomposition of Vectors: Part 2 20
9. Problems in the Plane: Work, the Inclined Plane and a Robot Arm 22
10. Vectors in Three Dimensions 26
11. Angles in Higher Dimensions 32
12. The Cross Product 34

Chapter II. Graphing: A Few Surfaces and Vector Functions May 27, 2005 41
13. Surfaces in Three Dimensions and Representations of Planes 42
14. Parametric Planes and Translating Among Coordinate Systems 50
15. Vector Functions 55

Chapter III. Vector Calculus of Curves May 27, 2005 65
16. Limits and Continuity for Vector Functions 66
17. Derivatives of Vector Functions 67
18. Integrals of Vector Functions 71
19. When is a Curve Confined to a Line or a Plane? 72
20. Tangent Lines 76
21. A Cycloid and Bezier Curves 79
22. Line Integrals 84
23. Orientation of Curves and Line Integrals 88
24. Flux Past a Curve in Two Dimensions 92
25. Calculus in Polar Coordinates 94

Chapter IV. The Gradient May 27, 2005 103
26. Functions of Two Variables: Continuity 104
27. Functions of Two Variables: Differentiability 106
28. The Chain Rule and The Tangent Plane 109
29. Functions of Three or More Variables 118
30. Implicit Functions 121
31. Derivatives as Matrices 127

32. Potentials 132

Chapter V. Integration Involving Surfaces and Volumes May 27, 2005 137
33. Area and Integrals in the Plane 138
34. Area of a Curved Surface and Surface Integrals 142
35. Parametric Descriptions of a Plane Set 147
36. Change of Variable in the Plane 148
37. Parametric Descriptions of a Surface 152
38. Change of Variable on a Surface 154
39. Surface Integrals Over Composite Surfaces 159
40. Orientation of Surfaces and Integrals Involving Vector Fields 161
41. Volume 163
42. Change of Variable for $3D$ Integrals 168
43. Divergence Theorem and Stokes' Theorem 174

Endnotes 185

Index 213

CHAPTER I

Basic Vector Facts May 27, 2005

1. First Steps

A **vector** is an object completely characterized by two quantities, which we call **magnitude** and **direction**. The physical meaning of these quantities in an application of vectors comes from experience and varies from application to application.

Vectors can be represented as **arrows** with the direction given by the "tail-to-tip" direction of the arrow and the magnitude given by its length.

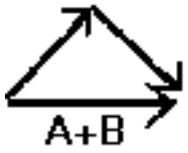
We take the point of view that two arrows located anywhere are merely **instances** of the "same" vector so long as each has the same length and direction. So the arrow to the left represents the "same" vector as the one on the right above, even though it is located at a different place.

You have plenty of experience with "fuzzing out" the distinctions among things which are manifestly different but which exhibit similarities upon which we wish to focus. For example, the fractions 3/4 and 6/8 represent different ideas. In the first, you break the "whole" into 4 equal pieces, and you have 3 of them. In the second, you break the "whole" into 8 equal pieces, and you have 6 of them. With these differences, there is something important that is similar about these two fractions: namely, I am just as full if I eat 3/4 of a pizza or if I eat 6/8 of a pizza. We choose to focus on that and we say $3/4 = 6/8$. We gather together all the fractions "equal" to 3/4 and refer to the entire pile of them by picking any convenient representative, such as that in lowest terms or with some specified denominator.

Similarly, the arrows above are different on the page, but by picking one of them we refer to both, and any other arrow with the same magnitude and direction as well!

Two vectors $A =$  and $B =$ 

are added by finding the copy of B which has its tail on the nose of a copy of A. The sum $A + B$ is the arrow that starts with its tail at the tail of this copy of A and ends with its tip at the tip of this copy of B.



It is important to note the direction of $A + B$, in this case from left to right.

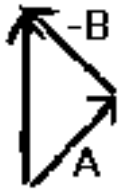
$-B$ is the vector that looks just like B but with tip and tail switched:

$B =$  $-B =$ 

A positive constant k times a vector A is a new vector pointing in the **same direction as** A but with length stretched (if $k > 1$) or shrunk (if $k < 1$) by the factor k. Negative multiples of A are said to have **direction opposite to** A.

$2A =$  and $(1/2)A =$ 

$A - B$ is defined to be $A + (-B)$. So $A - B$ is the vector on the left of the picture:

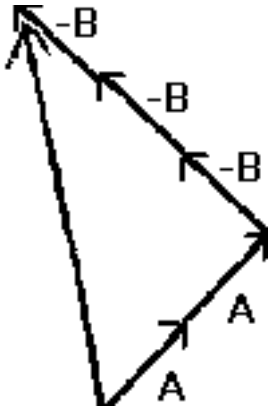


The process of adding two vectors is called **vector addition**. Multiplying a vector by a constant is called **scalar multiplication**.

The vector with zero magnitude is hard to represent as an arrow: it is called the **zero vector**, denoted 0. It doesn't really have a direction—or perhaps it has any direction. You pick. Context distinguishes it from the number 0.

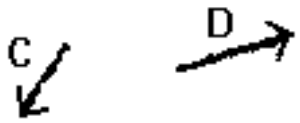
1.1. ***Exercise***. *You should satisfy yourself that:*

$$A + B = B + A,\ A - A = 0 \text{ and that } 2A = A + A.$$



On the far left is a picture of the vector sum $2A - 3B$. A vector such as this, formed as a sum of multiples of vectors, is often referred to as a **resultant vector**. It is also called a **linear combination** of the vectors involved, in this case A and B.

1.2. ***Exercise***. *Draw a picture of* $3C - 2D$ *and* $C + \frac{1}{2}D$ *where* C *and* D *are given by:*



There are a number of things in the world with which you are no doubt familiar that are commonly represented as vectors, and you should think a bit about the meaning of "magnitude" and "direction" in each specific case.

- **Displacement**—a representation of a movement from a starting place to an ending place, with emphasis on the completed movement rather than how it occurred or the specific starting spot.
- **Velocity**—a description of motion, whose magnitude is the **speed** and whose direction "points the way." The velocity would be the displacement vector over one time unit, if the motion continued unchanged for the whole time unit.
- **Forces**—these describe "pushes" by one thing against another. A force is the cause of acceleration. If you see changes in the motion of something, it is because there is a force acting on that thing. No such changes require that the resultant of all forces have zero magnitude.
- A representation of a uniform **Wind** or **Current**—in the air or water. This example is tied to velocity. It can be interpreted as the velocity of a dust particle swept up and carried along by an unvarying wind or current.

Why should vectors describe faithfully these categories of real-world experiences? I don't know. It is a puzzle. They just do. It is ONLY through experience, conjecture bolstered by many experiments, that we (i.e. physicists, engineers, you, me) decide that vectors are a reasonable tool to try to describe something in the world. Mathematicians can tell you how vectors behave. Only you can decide if vectors mimic well some aspect of the world.

There is, obviously, overlap and relationships among the items listed above. Each will be useful, alone and in combinations.

There are a couple of points I would like to make before getting down to business.

First, each instance of a vector in the world actually occurs at some specific spot, and whenever an arrow is drawn it is drawn somewhere specifically. When we think of something in the world as a vector, we take the point of view that any specific representative refers not only to itself but to all others with the same magnitude and direction too. When you refer to 7/3 you are often making a statement about 21/9 at the same time even without specifically mentioning that second fraction.

Second, a given push (a force) is a real thing that exists however we decide to describe it. The wind is just whatever it is and doesn't need us to tell it that it is 30 miles per hour from the North. A displacement across a room is a real

thing, in itself. But in physics and other classes we try to describe things, often using mathematics and numbers. This association always involves a huge pile of assumptions including, for example, a choice of a distance unit, a time unit, an "origin," directions for coordinate axes and methods for measuring lengths and angles and the passage of time and some way of gauging the magnitude of a "push" and on and on. There is also a conceptual framework, frequently generated by the esthetic sensibilities of the creators of the model, which help us think about the measurements. It is not always clear which among the conceptual underpinnings are necessary, or even if they are consistent. Our description depends not only on the real thing, but on all these choices involved in a representation too.

When we go through this process of assembling a model we must never forget that the map is not the territory. A nickname for a thing is not the thing itself. The universe **names itself**, and whatever shorthand we use to describe part of it leaves out almost everything. In applications we must always be looking "out the window" to make sure the world is still answering to our nickname for it. It is astounding how often, over the last couple of centuries, it comes when we call. We must be doing something right.

1.3. ***Exercise***. *This is an exercise designed to connect experiences people have had with the more abstract ideas we will be thinking about. It is best done in a group with the discussion guided by an instructor. Alternatively, it could be done by a single student accompanied by a good imagination. In each item below an activity will be described. After each activity there should be a discussion about what happened and possible alternative responses. The exercise will be repeated at the end of Section 4.*

A(i) The instructor picks a volunteer toward the back of the room. The instructor then moves about ten feet across the front of the room. The instructor then asks the volunteer to mimic that movement.
A(ii) The instructor moves ten feet in a different direction. Does this seem like "the same movement" as the first one?
A(iii) The instructor moves two feet in the same direction as the first movement. Does this seem like "the same movement" as the first one?

B(i) The instructor picks two volunteer who move toward an open area where everyone can see. The instructor gives each volunteer a heavy smooth ball, such as a billiard ball, and asks the first volunteer to roll it across the floor. The instructor asks the second volunteer to stand ready and, while the first ball is still rolling, mimic the motion of the first ball using his or her ball.
B(ii) The instructor rolls a ball with about the same speed but in a different direction. Does this seem like "the same motion" as the first one?
B(iii) The instructor rolls a ball in the same direction as the first motion but much faster. Does this seem like "the same motion" as the first one?

C(i) The instructor picks four volunteers, at least two of which claim to be physically durable. The instructor asks a volunteer to give "a medium sized push" to one of the durable volunteers. The instructor then asks another volunteer to mimic that push on a different durable volunteer.
C(ii) The instructor walks up to the original durable volunteer and gives a "medium

push" in a direction different from the first push. Does this seem like "the same push" as the first one?
C(iii) The instructor walks up to the original durable volunteer and gives a "very light push" in the same direction as the first push. Does this seem like "the same push" as the first one?

---

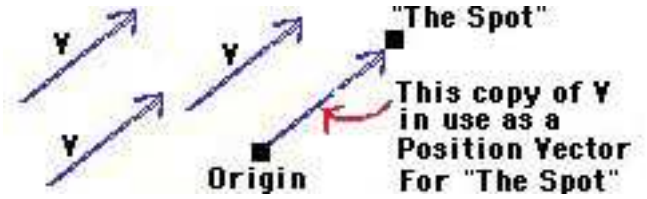
2. Position Vectors and Constant Velocity Motion: Part 1

In this section we will use vectors to describe specific locations and then the path and motion of a "particle" traveling in a straight line with constant speed.

Vectors which are used to describe specific locations are called **position vectors**. Since vectors do not **have** specific locations, we need to create a "convention" for how we should interpret a vector thought of in this way.

The only thing we need to create the interpretation beyond the ideas of the first section of this chapter, is to agree on a "center of the universe." This will be a known and agreed upon place, usually called the **origin**. All other locations will be described by the displacement vector needed to get to the location from this origin.

To reiterate: a vector which is to be used to describe a particular spot is called a **position vector**. To use a vector this way, you must first decide upon a "center" point called the origin. The copy of the vector with its tail at this origin is said to be in **standard position**. When in standard position, the nose will "point" to the spot in the plane we are trying to identify.



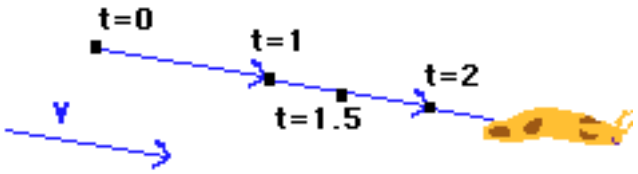
With this idea in hand, we will use position vectors to describe **constant velocity motion** of an object—that is, an object moving in such a way that displacements during any two equal time intervals are the same.

You can think of a moving object as a "moving point of light" that leaves a trace (a burn mark or a smoke trail perhaps) after it has passed over a point. If that doesn't suit your fancy, you can think of it as a "moving **slug**" that leaves a **slime** trail after it has passed over a point. We will label points on the track by the time or times when the light (or slug) passes over the point. That is what people mean when they say the motion is **parameterized**, and the time in this case is called the **parameter**.

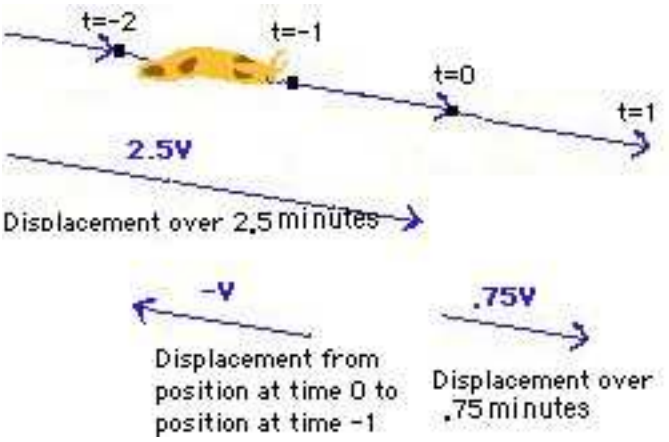
First, we remind the reader of the distinction between speed and velocity. Velocity is a vector that "points the way" of the motion. It is the displacement during one time unit. Its length is the speed.

We have thrown a new element into the mix here: to talk about velocity we must agree on an idea of **time** and a **means of measuring intervals of time.**

For specificity, let's think of our motion as describing the position of a constant velocity slug, with time measured in minutes from a moment we all agree is "time 0" and that the slug has constant velocity V. Each minute the slug moves the length of the velocity vector (the speed) in the direction of the velocity vector. If we find the copy of the velocity vector with tail at the slug starting place, the tip will be at the slug position after one minute of motion. If we only go for a fraction of a minute, or for many minutes, the slug position will be the tip of the appropriate multiple of the velocity vector when the tail of this vector is at the starting place. The displacement of the slug position over t minutes from **any** starting place will be the vector tV.



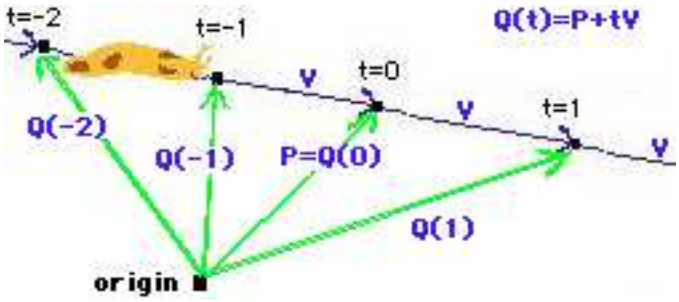
The starting time might have been chosen for convenience rather than the actual time the slug started moving. So negative times would merely refer to times before then. So to get to the location of the slug one minute before time 0 you would use a displacement vector $-V$ from its position at time 0.



Finally, we get to a position vector description of the journey of the slug. If an origin is chosen and the position vector of the slug at time zero is P with respect to this origin then the position vector of the slug at time t is given by

$$Q(t) = P + tV.$$

This is called a **parametric vector equation** for the position of the slug.



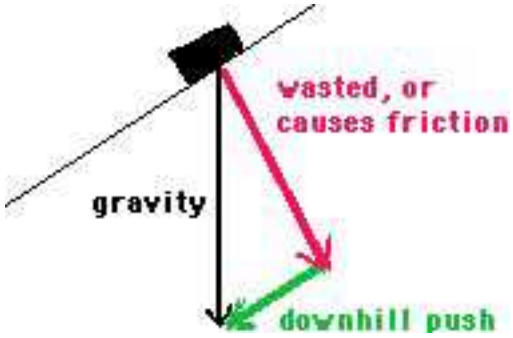
2.1. ***Exercise***. *In the picture of constant velocity slug motion you see below, the origin and slug positions at times* 4 *and* 8 *are identified. From this picture, create a vector* V *which represents the velocity of the motion. Identify the position vector* P *which points to the position of the slug at time* 0*. Finally, use the formula* $Q(t) = P + tV$ *to help you find the position on this picture at times* $-1, 1$ *and* 2*.*

3. Decomposition of Vectors: Part 1

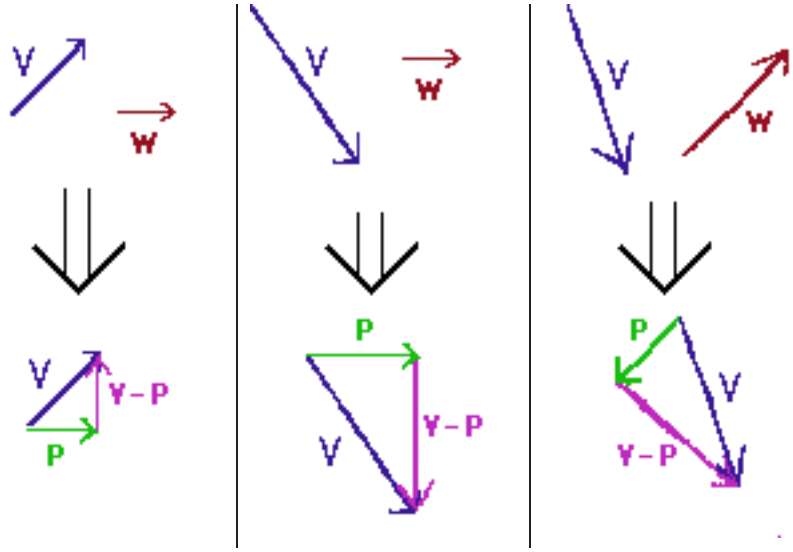
We know how to add vectors, combine them into a single resultant vector. A natural next step might be to see how we can break them into pieces in various ways. In this section we will think about how to break a vector into the sum of two others. One of these will be a multiple of a specified vector and the other perpendicular to that specified vector. This is a very important process in applications. The process is called **decomposition**. When we finally get to precise calculations, decompositions will be easy to find from an arithmetical standpoint.

In order to create the picture of a decomposition you need to know only one thing not shown in the first section of this chapter. You must have a concept of **perpendicularity**, and be able to tell when two arrows are perpendicular to each other by some method. In this section, the old standby "eyeball" method will suffice.

A common and very important usage of vector decomposition occurs when we consider force vectors. A classic example would be that of a box sliding down a slanted board. The most obvious force here is the weight of the box. But that force is directed straight down and the surface of the board prevents movement in that direction. The right way to handle this is to decompose the force caused by gravity into two perpendicular pieces: the part that is straight into the surface of the board (the source of friction) and the other pointing along the line of the board. It is only this last part which makes the box slide.



Whatever the source of the vectors involved may be, we will draw some pictures here to see "how to do it." We want to learn how to decompose a vector V into the sum of a vector P which is a multiple of some vector W and another vector $V - P$ which is perpendicular to W. We call the second vector $V - P$ because whenever $V = P + A$ it must be that $A = V - P$, so there is no point in introducing an independent name for the perpendicular part of the decomposition. Find below three different decompositions of this type in picture form.



Notice two things about the pictures above: First, in each case P and $V - P$ are perpendicular to each other. Second, P is a multiple of W.

In constructing the decomposition, the length of W is irrelevant. The only thing that is important about W is its direction.

To create the decomposition, draw a fresh picture of a V and W pair from one of the pictures above on a sheet of paper.

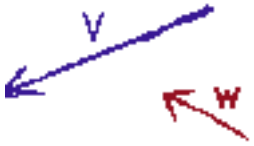
V should be somewhere in the middle and W off to the side for reference. Draw a dotted line through the tail of the copy of V. The dotted line must go along the same direction as W. Extend this dotted line a good bit on either side of V, across the whole paper.

Next lay your pencil down on the paper. Put the eraser on this dotted line with the point on the same side as V is on. Make the shaft of the pencil perpendicular to this dotted line. This is where you need to know about perpendicularity.

Slide the pencil up or down the dotted line, keeping it perpendicular to the dotted line, till the pencil tip points at the tip of V or the shaft crosses the tip of V. Stop. This gives the decomposition.

3.1. ***Exercise***.

(i) Draw a picture showing the decomposition as described above for the indicated V and W on the left.



(ii) Draw a picture showing the decomposition as described above for the indicated V and W on the left.

---

4. Vectors in the Plane

We have spent considerable time thinking about arrows and drawing pictures and introducing vocabulary. Those pictures will guide you in the later work, and help you to understand what the calculations are telling you. But if we are to produce exact answers rather than qualitative approximations we must do things differently. Our first precise description of vectors will be as arrows in the ordinary XY plane, **freighted with all of its preliminary choices of axes at right angles to each other, origin, choice of units and so on.**

To describe each arrow we must identify a tail followed by a tip. In the XY plane this requires a number of pieces of information—the coordinates of each and also which is tail and which is tip. There is, evidently, some redundancy here and we can cut this down by choosing that representative of a vector which has its tail at the origin: the **standard position** version of the vector. With this convention the vector can be completely described by referring to its tip alone.

The following arrows all represent the SAME vector (draw pictures and convince yourself of this!)

- tail at $(0,0)$ and tip at $(2,2)$
- tail at $(3,-7)$ and tip at $(5,-5)$
- tail at $(-2,-2)$ and tip at $(0,0)$

The vector corresponding to ALL these arrows, or any one of them, will be denoted $\langle 2,2\rangle$. The difference between $(2,2)$ and $\langle 2,2\rangle$ is the following: $(2,2)$ is the location of a spot in the plane; $\langle 2,2\rangle$ is a vector which, when represented as an arrow with its tail at the origin, points at $(2,2)$.

An additional benefit of the standard position description of a vector is that it is ready to use as a position vector whenever we want. $\langle 2,2\rangle$ is the position vector for the point located at $(2,2)$.

You should satisfy yourself by drawing pictures that multiplication by constants and vector addition can be handled using this standard representation according to the following pattern: If $A = \langle 2,2\rangle$ and $B = \langle 2,-2\rangle$ then

$$2A - 3B = 2\langle 2,2\rangle - 3\langle 2,-2\rangle = \langle 4,4\rangle - \langle 6,-6\rangle = \langle -2,10\rangle .$$

To multiply vectors by a constant, multiply the coordinates of the standard representation by the constant. To add or subtract vectors, add or subtract corresponding coordinates.

Usually we will denote vectors by capital Latin letters, such as V, W, A or B and the coordinates of the tip of a vector when in standard position will be represented as lower case letters with subscripts. So we might write $V = \langle v_1, v_2\rangle$. The individual coordinates of the tip in standard position are sometimes called the **components** of the vector.

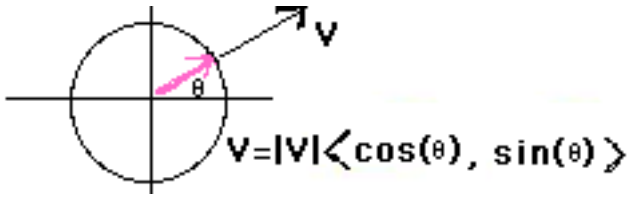
The **magnitude** of a vector $V = \langle v_1, v_2\rangle$ is denoted $|V|$ and is defined to be the length of any arrow representing V.

By the Pythagorean distance formula, this length is: $|V| = \sqrt{v_1^2 + v_2^2}$.

4.1. ***Exercise***. *Verify directly that, unless $V = 0$, the vector $\frac{V}{|V|} = \left\langle \frac{v_1}{|V|}, \frac{v_2}{|V|}\right\rangle$ has length 1. Why does it point in the same direction as V?*

Note: if θ is the counterclockwise angle from the positive X axis to V we have

$$V = |V|\left\langle \frac{v_1}{|V|}, \frac{v_2}{|V|}\right\rangle = |V|\langle cos(\theta), sin(\theta)\rangle .$$

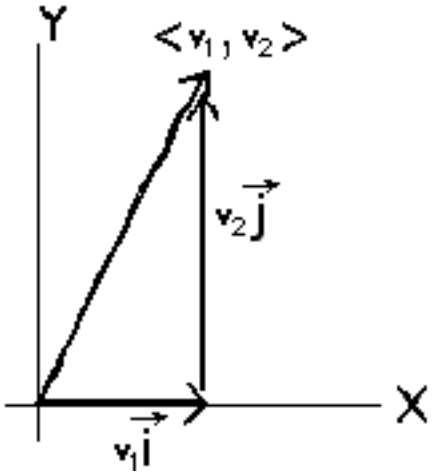


The angle θ is related to $arctan\left(\frac{v_2}{v_1}\right)$.

This is a very important representation for vectors in the plane. It separates cleanly the two things that make up a vector (direction and magnitude) in a way that the XY coordinates of the tip do not. The direction is indicated by an arrow $\langle cos(\theta), \sin(\theta)\rangle$ of unit length—these are called **unit vectors**—with its nose resting on the **unit circle** and "pointing the way." Unit vectors are also sometimes called **direction vectors**. The magnitude $|V|$ is the "stretch or shrink factor" which modifies the length of the direction vector.

So we have two ways of representing vectors in the plane. The first gives the XY coordinates of the tip when the vector is in standard position and is better for doing most kinds of vector algebra. The second is more intuitive and emphasizes the two defining properties of a vector. The ability to translate from one form to the other is a key skill.

There is an alternative notation for the XY representation of a vector that you will see from time to time.



We let $\vec{i}$ stand for the unit vector in the positive X direction, and let $\vec{j}$ stand for the unit vector in the positive Y direction.

$$\vec{i} = \langle 1, 0\rangle \text{ and } \vec{j} = \langle 0, 1\rangle .$$

Do not confuse this vector $\vec{i}$ with the complex number i. With this convention, any vector can be written as a vector sum involving $\vec{i}$ and $\vec{j}$.

$$\begin{aligned} V &= \langle v_1, v_2\rangle \\ &= v_1 \langle 1, 0\rangle + v_2 \langle 0, 1\rangle \\ &= v_1\vec{i} + v_2\vec{j}. \end{aligned}$$

There is nothing terribly significant about this alternative notation (other than that you will see it often) but it does introduce the idea of breaking a vector up into a sum of two perpendicular vectors.

4.2. ***Exercise***. *You should verify (and be able to translate from one form to the other yourself if needed) the following equations:*

$$\langle 2\ , 2\rangle \ = \ \sqrt{8}\ \left\langle \frac{\sqrt{2}}{2},\ \frac{\sqrt{2}}{2} \right\rangle \ = \ \sqrt{8}\ \left\langle cos\left(\frac{\pi}{4}\right),\ sin\left(\frac{\pi}{4}\right)\right\rangle .$$

$$\langle -2\ , -2\rangle = \ \sqrt{8}\ \left\langle \frac{-\sqrt{2}}{2},\ \frac{-\sqrt{2}}{2} \right\rangle \ = \ \sqrt{8}\ \left\langle cos\left(\frac{5\pi}{4}\right),\ sin\left(\frac{5\pi}{4}\right)\right\rangle .$$

$$\langle 0\ , -7\rangle = \ 7\ \langle 0,\ -1\rangle \ = \ 7\ \left\langle cos\left(\frac{3\pi}{2}\right),\ sin\left(\frac{3\pi}{2}\right)\right\rangle .$$

$$\begin{aligned}\langle -2\ , 3\rangle =& \sqrt{13}\ \langle\ cos\left(\pi + arctan\left(-1.5\right)\right), sin\left(\pi + arctan\left(-1.5\right)\right)\ \rangle \\ &\approx 3.61\ \langle\ cos\left(2.16\right), sin\left(2.16\right)\ \rangle \approx\ 3.61\ \langle -.556, .832\rangle .\end{aligned}$$

4.3. ***Exercise***. *Repeat Exercise 1.3. Then discuss and compare the three different ways of thinking about vectors we have seen:*

- *vectors as "experiences:" pushes on your shoulder, actual completed movement of your body from one spot to a second spot and as a description of a motion which you witness as it happens*
- *vectors as arrows (where arrows are "the same" if they have the same direction and length)*
- *vectors given by a coordinate pair such as* $\langle 3, 8\rangle$.

---

5. Dot Products

We will now define a way of multiplying two vectors called **dot product**.

If $V = \langle v_1, v_2\rangle$ and $W = \langle w_1, w_2\rangle$ we define $V \cdot W = v_1w_1 + v_2w_2$.

This product takes two vectors and produces a number obtained by simple arithmetic involving the coordinates of each.

5.1. ***Exercise***. *(i) Show that* $V \cdot W = W \cdot V$.

(ii) Show that $(7V) \cdot W = V \cdot (7W) = 7(V \cdot W)$.

(iii) Satisfy yourself that, in general, constants can be moved around in a dot product like the 7 *above.*

(iv) Show that $(V + P) \cdot W = V \cdot W + P \cdot W$ *and* $W \cdot (V + P) = W \cdot V + W \cdot P$.

(v) $V \cdot V = |V|^2$.

Properties (iii) and (iv) of the exercise together make the dot product an example of a **tensor**. Since there are two vectors involved, it is called a 2-tensor. Property (i) in this context is called **symmetry**. The dot product is an example of a symmetric 2-tensor. If you stay in this business awhile you will run into tensors of various kinds.

Dot products are incredibly useful.

If $V = |V|\,\langle cos(\alpha),\ sin(\alpha)\rangle$ and $W = |W|\,\langle cos(\beta),\ sin(\beta)\rangle$

then $V \cdot W = |V||W|\,\langle cos(\alpha),\ sin(\alpha)\rangle \cdot \langle cos(\beta),\ sin(\beta)\rangle$

$= |V||W|(cos(\alpha)cos(\beta) + sin(\alpha)sin(\beta)) = |V||W|cos(\alpha - \beta)$.

This means that you can use elementary arithmetic to get information about the angle $\alpha - \beta$ between two vectors!

We have just show that if θ is the **angle between vectors** V and W then

$$V \cdot W \ = |V||W|cos(\theta).$$

For example, the angle θ between $\langle 3, 7\rangle$ and $\langle 5, -3\rangle$ satisfies

$\langle 3, 7\rangle \cdot \langle 5, -3\rangle = 15 - 21 = \sqrt{9+49}\ \sqrt{25+9}\ cos(\theta)$

and so $\quad \frac{-6}{\sqrt{9+49}\ \sqrt{25+9}} \ = \ cos(\theta) \quad$ which means $\quad \theta \ \approx \ 98^o.$

5.2. ***Exercise***. *(i) Verify the angle found above by drawing a careful picture and doing trigonometry.*

Also, satisfy yourself that the following statements are true:

(ii) $V \cdot W > 0$ *implies that* V *and* W *are less than* 90^o *apart.*

(iii) $V \cdot W = 0$ *implies that* V *and* W *are* ***perpendicular****—that is, are at right angles to each other. Vectors that are perpendicular to something (a line, a plane, another vector) are often called* ***normal*** *or* ***orthogonal*** *to that other object. This vocabulary is most useful when neither* V *nor* W *is* 0.

(iv) $V \cdot W < 0$ *implies that* V *and* W *are more than* 90^o *apart.*

(v) $V \cdot W = |V||W|$ *implies that* V *and* W *point in the same direction.*

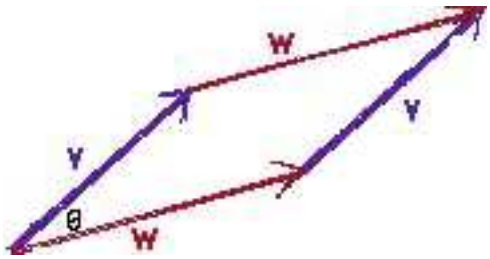
(vi) $V \cdot W = -|V||W|$ *implies that* V *and* W *point in the opposite direction.*

(vii) $\langle 5, 3\rangle$ *is normal to* $\langle -3, 5\rangle$. *Find a vector normal to* $\langle a, b\rangle$.

5.3. ***Exercise***. *If* $V = \langle v_1, v_2\rangle$ *and* $W = \langle w_1, w_2\rangle$ *we define* $\det(\mathbf{V}, \mathbf{W})$ *to be the number* $v_1w_2 - v_2w_1$. *The notation comes from the fact that, for those among you who know about* ***determinants****, it is the determinant, usually denoted* $\det A$, *of the matrix* A *where*

$$A \ = \ \begin{pmatrix} v_1 & v_2 \\ w_1 & w_2 \end{pmatrix}.$$

Here is an application for this number:



V and W can be used to form a parallelogram. Show that the area of this parallelogram is

$$|W||V|sin(\theta) = |W||V|\sqrt{1 - cos^2(\theta)}.$$

Now square this last number and verify that it is $(v_1w_2 - v_2w_1)^2$. Conclude that the area of the parallelogram is $|\det(V, W)|$.

---

6. Problems in the Plane: Displacements, Forces and Velocity

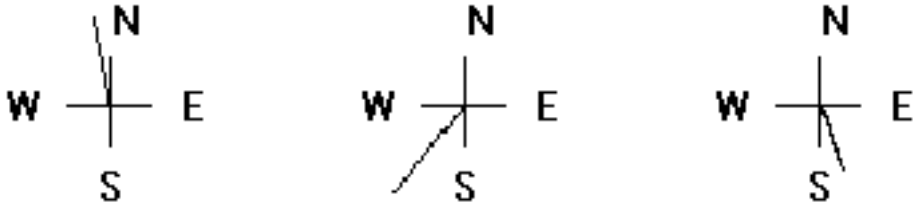
The reader should work through all the statements made in this section.

Problem 1: Displacements

Bearings are commonly used in navigation problems to indicate direction. A bearing looks like this:

$$N10^oW \qquad or \qquad S38^oW \qquad or \qquad S18^oE.$$

The meaning of a bearing is as follows: You first point your nose straight north or south, whichever is indicated by the first letter. Then you rotate the given angle toward the compass heading suggested by the second letter, either east or west. That is the direction.



The problem is as follows:

Suppose a cross country runner jogs 5 miles $N10^oW$ and 3 miles $S38^oW$ and then 4 miles $S18^oE$. How far is the runner from home, and what bearing should the runner take to get there?

The Solution:

The resultant displacement vector is:

$$\begin{aligned} &5\,\langle cos(100^o), sin(100^o)\rangle \\ +&3\,\langle cos(-128^o), sin(-128^o)\rangle \\ +&4\,\langle cos(-72^o), sin(-72^o)\rangle \end{aligned}$$

This vector is, approximately, $\langle -1.479, -1.244\rangle$, or about

$$1.93\,\langle cos(220^o), sin(220^o)\rangle\,.$$

It is about 1.93 miles to home and the runner should head 180^o away from the angle of the resultant vector. The bearing will be $N50^oE$.

Problem 2: Forces

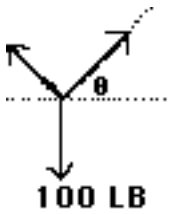
Three people are pushing on different sides of a huge ball but the ball is not moving. The pushes of first and second of these people are represented by force

vectors $\langle 30, 20\rangle$ and $\langle 15, -10\rangle$, respectively. Which force vector represents the push of the third?

The Solution:

Since the ball is not moving, the sum of the three forces must add to the zero vector. So if F is the force of the third person, $F + \langle 30, 20\rangle + \langle 15, -10\rangle = \langle 0, 0\rangle$. So $F = -\langle 30, 20\rangle - \langle 15, -10\rangle = \langle -45, 10\rangle$.

Problem 3: Forces



We suppose that there is a 100 pound weight hooked (not moving but free to slide) over a length of rope. The rope is at a given angle (on both sides) from the horizontal as indicated. The rope is very light in comparison to 100 pounds, so the weight tightens it to (nearly) a straight line. The **tension** is a measure of how hard the rope is pulling, and is the magnitude of a force vector whose direction lies along the rope. What is the tension, T, on this rope?

The Solution:

Since nothing is moving, all the forces must counteract each other at each point along the rope and, in particular, at the point where the weight is attached. The resultant of the three forces there, indicated by arrows, must be zero. The sum is

$$\langle 0, 0\rangle = 100\,\langle 0, -1\rangle + T\,\langle cos(\theta),\ sin(\theta)\rangle + T\,\langle cos(\pi - \theta),\ sin(\pi - \theta)\rangle\,.$$

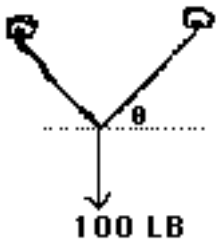
After fiddling around with some algebra, this gives

$$\langle 0, 0\rangle = \langle 0, -100 + 2T\ sin(\theta)\rangle \quad \text{and so} \quad T = \frac{50}{sin(\theta)}.$$

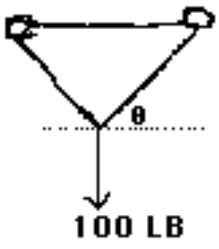
Note: the minimum tension occurs when the angle is 90^o. When the angle gets small enough, the weight WILL break the rope.

Problem 4: Forces

Let's embellish the last problem a bit and think about it from the standpoint of a bolt anchored to a wall and to which the rope is attached.



In one situation, the rope is tied off at each anchor point. In the other situation, the rope slides through rings at the same height and is allowed to run freely in a big loop. The issue here is to find the forces acting on the anchor bolt in each case.



The Solution:

Experience shows that a bolt like this will refuse to move until it breaks or pulls out of the wall. So it musters up whatever force is required to keep it from moving (to counteract the other forces on it) until that catastrophic event. Unbalanced forces always generate a change in motion. No change in motion means that the

resultant of all forces is the zero vector in each case. In the diagram on the left, the situation is clear. The force which the upper right bolt must counteract is

$$T\left\langle -cos(\theta), -sin(\theta)\right\rangle \quad \text{which has magnitude} \quad T = \frac{50}{sin(\theta)}.$$

But in the second case, there are TWO forces tugging on the upper right bolt, one along each segment of the rope leading away from that bolt. The resultant force to which it must respond is

$$T\left\langle -cos(\theta), -sin(\theta)\right\rangle + T\left\langle -1, 0\right\rangle = T\left\langle -1 - cos(\theta), -sin(\theta)\right\rangle.$$

This vector has magnitude

$$T\sqrt{1 + 2cos(\theta) + cos^2(\theta) + sin^2(\theta)} = T\sqrt{2}\sqrt{1 + cos(\theta)}.$$

By merely looping the rope around the pair of bolts rather than tying it off at each one you increase the magnitude of the force on the bolt by at least 40% and up to 100%. The worst case also happens for smaller angles, exacerbating an already bad situation.

If you made the apparatus using real bolts and a rope, friction each time the rope changes direction at an anchor would reduce this effect somewhat. Calculating friction can be quite complicated, involving the diameters of rope and bolt shaft, angle of contact and so on. But friction must be considered to make sense of more complicated situations or to make accurate predictions. But that is for a physics class!

Problem 5: Combining Velocities

Suppose that a plane is pointing $N10^oW$ and its throttle and altitude are such that if it were flying on a windless day it would be moving at 400 kilometers per hour. However there is a wind, coming from $N40^oW$ at 50 KPH. How fast is the plane moving with respect to the ground and with what bearing?

The Solution:

$$\begin{aligned} &400\left\langle cos(100^o), sin(100^o)\right\rangle - 50\left\langle cos(130^o), sin(130^o)\right\rangle \\ &\approx \left\langle -37.3,\ 355.6\right\rangle \approx 357.6\left\langle cos(96^o), sin(96^o)\right\rangle. \end{aligned}$$

So the plane is flying at a speed of about 358 KPH at a bearing of $N6^oW$.

---

7. Position Vectors and Constant Velocity Motion: Part 2

In this section we will take another look at the material of Section 2, this time with coordinates. For specificity, we will measure the parameter t in seconds.

Suppose we have an object moving in the plane with velocity that never changes and starting at some place at time $t = 0$, such as $(1, 2)$. In this context, $\langle 1, 2\rangle$ in standard position **points at** the spot and we will say that the object **is at** $P = \langle 1, 2\rangle$. Let us say that the velocity vector is found to be $V = \langle -3, 4\rangle$. The velocity vector is the displacement vector after one second of motion, so you could get these two pieces of information by watching the motion for one second. We will say that the velocity vector **lies in** the line of the motion. The vocabulary means

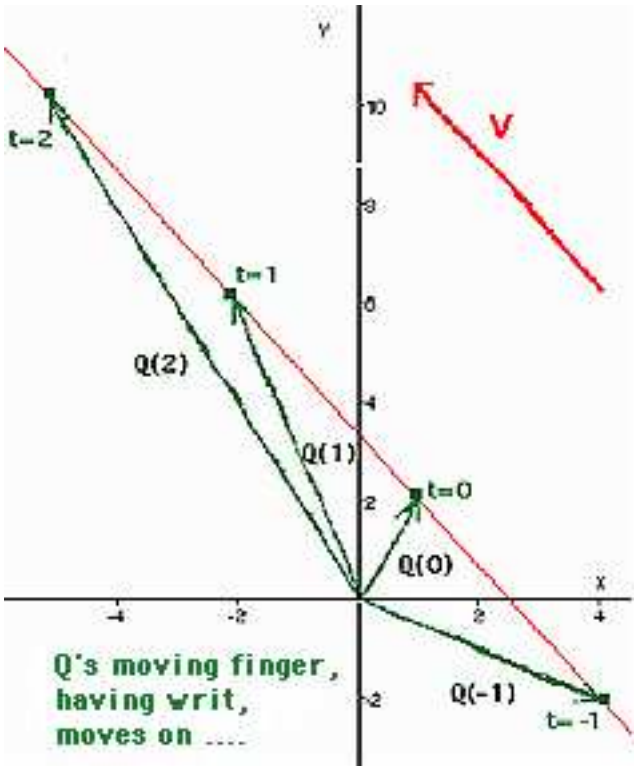
that if you choose the copy of V with its tail on the line of the motion, the tip (and so the whole arrow) is in the line too.

$$\text{Note that} \quad V \;=\; \langle -3, 4\rangle = 5\left\langle \frac{-3}{5}, \frac{4}{5}\right\rangle \approx 5\,\langle cos(126.9^o),\ sin(126.9^o)\rangle\,.$$

After 1 second the object will have moved 5 meters in the direction of V and so will be at $P+V$. After 2 seconds it will have moved another 5 meters in the direction $\langle -3, 4\rangle$ and so will be at $P+V+V = P+2V$.

At time $t=-1$, that is one second before it was at P, it must have been at $P-V$ (if it is to arrive at P one second later.) In general, for any time t, the position at time t, which we will denote Q or $Q(t)$, will be given by

$$Q(t) = P + tV.$$



As you will recall from Section 2, this is called a **parametric vector equation** for the motion. The tip of the arrow $Q(t) = P+tV$ sweeps out the motion as time passes. You might think of this as like a "slope-intercept form" for the linear motion. P is the location at time 0, and V carries the direction and speed information about the motion.

Note that:

$$Q(t) = P + tV = \langle 1-3t, 2+4t\rangle \text{ so } X = 1-3t \text{ and } Y = 2+4t,$$

where X and Y represent the X and Y coordinates of the tip of Q as functions of time. We can **eliminate** t from this whole business as follows:

$$t = \frac{1-X}{3} \qquad \text{so} \qquad Y = 2 + 4\left(\frac{1-X}{3}\right) \qquad \text{or} \qquad Y = \frac{-4}{3}\,X + \frac{10}{3}.$$

Another (easier) method of eliminating t is to use the fact that, for a generic point $Q = \langle X, Y\rangle$ on the line, $Q - P = tV$: that is, $Q - P$ is a multiple of the velocity vector V. So if N is any vector perpendicular to V then $(Q - P) \cdot N = 0$. This is called the **normal form** for the line in the XY plane..

In our situation we can pick $N = \langle 4, 3\rangle$ so $(\langle X, Y\rangle - \langle 1, 2\rangle) \cdot \langle 4, 3\rangle = 0$ which yields $(X - 1)4 + (Y - 2)3 = 0$ or $4X + 3Y = 10$.

In any case, we can see the motion is along a line in the plane, and we have a formula for this **geometrical track** of the parametric motion. Note how the slope is related to the components of the velocity vector $\langle -3, 4\rangle$. This is not chance.

There is something important missing when we eliminate the parameter. In this form we lose all knowledge of **when** we are anywhere on this line, which in applications is often the whole point! Knowing the geometrical track can be an aid, however, in drawing the graph or for other reasons.

In applications you will not always be given the velocity and position at time 0, just as in the old line exercises from elementary algebra you were not always given the slope and Y axis intercept. Still, we should be able to cook up an equation for the motion given two independent facts about it.

In the problems below we assume constant velocity motion in the plane. If you get stuck - DRAW PICTURES FOR VARIOUS t VALUES!

Problem 1: A Parametric Vector Equation—Where Will the Object Be?

Write a parametric vector equation for the position of a moving object located at $\langle 3, 5\rangle$ at time 0 and $\langle 4, 7\rangle$ at time 1. Where will it be at $t = 2$?

Movement during 1 second: $\langle 4, 7\rangle - \langle 3, 5\rangle = \langle 1, 2\rangle$ = velocity.

$Q(t) = \langle 3, 5\rangle + t\,\langle 1, 2\rangle$. So it will be at $Q(2) = \langle 3, 5\rangle + 2\,\langle 1, 2\rangle = \langle 5, 9\rangle$ at time 2. You can also get this by adding a copy of the velocity vector to the position at time 1.

Problem 2: A Parametric Vector Equation Plus Elimination

Write a parametric vector equation for the position of a moving object with velocity vector $\langle 1, -5\rangle$ and which is located at $\langle 5, 9\rangle$ at time 4. What is the XY formula for the geometrical track upon which the motion takes place?

At time 0 it was at $\langle 5, 9\rangle - 4\,\langle 1, -5\rangle = \langle 1, 29\rangle$.

So $Q(t) = \langle 1, 29\rangle + t\,\langle 1, -5\rangle = \langle 1 + t, 29 - 5t\rangle$. So $X = 1 + t$ and $Y = 29 - 5t$. So $Y = 29 - 5(X - 1)$ or, in slope-intercept form, $Y = -5X + 34$.

Problem 3: A Parametric Vector Equation Plus a "Wall"

Write a parametric vector equation for the position of a moving object located at $\langle -3, 0\rangle$ at time 5 and $\langle 12, 6\rangle$ at time 8. When will the object pass through a barrier set up on the line $Y = -X + 100$?

Movement during the 3 seconds from $t = 5$ to $t = 8$: $\langle 12, 6\rangle - \langle -3, 0\rangle = \langle 15, 6\rangle$. So the movement in 1 second is one third as much: $(1/3)\langle 15, 6\rangle = \langle 5, 2\rangle$ = velocity. At time 0 it was at $\langle 12, 6\rangle - 8\langle 5, 2\rangle = \langle -28, -10\rangle$.

So $Q(t) = \langle -28, -10\rangle + t\langle 5, 2\rangle$. This will be at the barrier at $Y = -X + 100$ when $-10 + 2t = -(-28 + 5t) + 100$. That means $t = 138/7$ seconds.

7.1. ***Exercise***. *(i) Write a parametric vector equation for the position of a moving object located at* $\langle 0, 5\rangle$ *at time* 0 *and with velocity vector* $\langle -1, 7\rangle$. *Where will it be at time* 10*? Where was it at time* -5*? When will it hit a wall set up on the line* $Y = X + 100$*?*

(ii) Write a parametric vector equation for the position of a moving object located at $\langle -2, 3\rangle$ *at time* 0 *and* $\langle -4, 1\rangle$ *at time* 1. *Draw a picture of this motion.*

(iii) Write a parametric vector equation for the position of a moving object with velocity vector $\langle -1, 8\rangle$ *and which is located at* $\langle 1, -9\rangle$ *at time* 6. *Draw a picture of this motion. What is the* XY *formula for the geometrical track upon which the motion takes place?*

(iv) Write a parametric vector equation for the position of a moving object located at $\langle -2, 9\rangle$ *at time* 5 *and* $\langle 12, 16\rangle$ *at time* 12. *When will the object hit a wall set up on the line* $Y = -X + 100$*?*

---

8. Decomposition of Vectors: Part 2

In this section we revisit the ideas of Section 3 only this time we use coordinates and can obtain more precise information. You will recall that the goal was to decompose a vector into the sum of two other vectors one of which was along a given direction while the other was perpendicular to that direction.

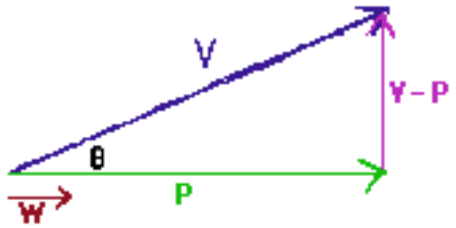
Being able to draw the picture is important and will allow you to estimate the decomposition fairly accurately in the plane, so you might want to go back and take another look at Section 3 at this point.

The fact we need for our calculation is:

$V \cdot W = |V||W|cos(\theta)$ where θ is the angle between nonzero vectors V and W.

We will use the fact that $\frac{V \cdot W}{|W|} = |V|cos(\theta)$ in the calculation below.

Notice in the following picture that, with W as shown, the length of P is $|V|cos(\theta)$.



So $P = |V|cos(\theta)\frac{W}{|W|} = \left(\frac{V\cdot W}{|W|}\right)\frac{W}{|W|} = \left(\frac{V\cdot W}{W\cdot W}\right)W.$

The vector $V - P$ should be perpendicular to W if the picture is correct.

Let's see:

$$\begin{aligned} W\cdot(V-P) &= W\cdot V - W\cdot P = W\cdot V - W\cdot\left(\frac{V\cdot W}{W\cdot W}\right)W \\ &= W\cdot V - \left(\frac{V\cdot W}{W\cdot W}\right)W\cdot W \\ &= W\cdot V - W\cdot V = 0. \end{aligned}$$

So $V - P$ and W are perpendicular as advertised.

The complete decomposition calculation is in three steps:

- Calculate $P = \left(\frac{V\cdot W}{W\cdot W}\right)W.$
- Then calculate $V - P.$
- Then verify that $P\cdot(V-P) = 0.$ (This catches most arithmetic "issues.")

Problem: A Decomposition

Decompose $\langle 4,5\rangle$ into the sum of a vector which is a multiple of $\langle 1,2\rangle$ and another which is perpendicular to $\langle 1,2\rangle$.

$$\begin{aligned} P &= \left(\frac{\langle 4,5\rangle\cdot\langle 1,2\rangle}{\langle 1,2\rangle\cdot\langle 1,2\rangle}\right)\langle 1,2\rangle \\ &= \left(\frac{4+10}{1+4}\right)\langle 1,2\rangle = \left\langle\frac{14}{5},\frac{28}{5}\right\rangle. \\ V-P &= \left\langle\frac{20}{5},\frac{25}{5}\right\rangle - \left\langle\frac{14}{5},\frac{28}{5}\right\rangle \\ &= \left\langle\frac{6}{5},\frac{-3}{5}\right\rangle. \end{aligned}$$

Verify:

$$\begin{aligned} P\cdot(V-P) &= \left\langle\frac{14}{5},\frac{28}{5}\right\rangle\cdot\left\langle\frac{6}{5},\frac{-3}{5}\right\rangle \\ &= \frac{84}{25} - \frac{84}{25} = 0. \end{aligned}$$

8.1. ***Exercise***. *(i) Draw a picture to verify that the decomposition in the problem, given algebraically, is "about right."*

(ii) Decompose $\langle -4,1\rangle$ into the sum of a vector which is a multiple of $\langle -1,8\rangle$ and another which is perpendicular to $\langle -1,8\rangle$. Using a picture, verify that your decomposition is "about right."

(iii) Decompose $\langle 0,1\rangle$ into the sum of a vector which is a multiple of $\langle -1,1\rangle$ and another which is perpendicular to $\langle -1,1\rangle$.

We will make two final points here:

First, the vector P found above is also called the **vector projection of V in the direction of W** and often denoted $Proj_W(V)$. The specificity provided by this notation is needed when there are decompositions of different vectors or along various directions within one problem. It is apparent from the definition that $Proj_W(V) = Proj_U(V)$ when U is a nonzero multiple of W.

Second, sometimes the decomposition vectors P and $V - P$ themselves are not needed but only their magnitudes. The number $\dfrac{V \cdot W}{|W|}$ is called the **scalar projection of V in the direction of W**. It is a positive number if the angle θ between V and W is less than 90^o. Its absolute value is the magnitude of P.

The various magnitudes are are related by:

$$|P| = |V||cos(\theta)| = \frac{|V \cdot W|}{|W|} \quad \text{and} \quad |V - P| = |V||sin(\theta)|.$$

Also note that:

$$|V|^2 = |P|^2 + |V - P|^2.$$

Using this last equation, you can get the magnitude of the third from the magnitudes of any two of the terms involved.

9. Problems in the Plane: Work, the Inclined Plane and a Robot Arm

In this section we are going to use dot products to perform some calculations whose meaning will be explored in much greater depth in physics and engineering classes.

Work

The concept of "work" in physics is not the usual idea of work that pops to mind when we say "I worked all day weeding the garden." In the colloquial sense, work means "I accomplished something today." or "I expended a lot of effort today." Neither of these fuzzy notions corresponds to the idea of work as defined in physics, which is part of the energy "accounting system."

Work done by a constant force, in physics, involves the magnitude of the part of the force that lies in the direction of displacement times the magnitude of that displacement. The speed of the motion is not relevant: only the displacement.

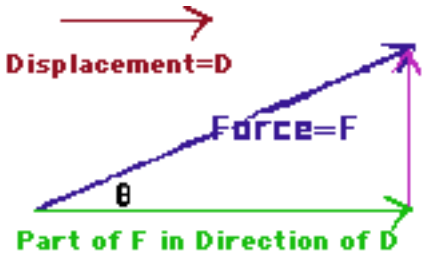
So if I am standing in a room holding a 200 pound sack of flour, beads of sweat popping out on my forehead—but not moving—then I am doing no work. If there is no movement there can be no work.

Even if I move, so long as I move horizontally and not up or down, no work has been done by or against gravity. At least part the force must lie in the direction of motion.

When I have found the component of the force along the line of the displacement, I calculate its length and multiply by the length of the displacement vector. If the angle between the force and the displacement is less than 90^o this number is

the work. If the angle between the force and the displacement is more than 90^o I multiply this number by -1. This number is the work.

When the work is positive the displacement took place aided by the force. When negative, the force was hindering the displacement.



From a calculation standpoint, if F is the force vector and D is the displacement vector we have, from the picture above,

$$\text{Work } = |D||F|cos(\theta) \ = |D||F| \left(\frac{D \cdot F}{|D||F|} \right) \ = D \cdot F.$$

So to calculate the work, dot the displacement vector against the force vector.

Why should this odd number be useful in physics? Good question. The story is pretty long and is the object of much thought in physics or engineering classes. However now, if anyone should ask, you can calculate it.

We should make several points here about the picture shown above. The force vector shown is less than 90^o away from the displacement vector, so the dot product is positive. If the force acts to "slow down" the displacement instead of "helping it along" the angle exceeds 90^o and the dot product will be negative. Second, this calculation is for constant force and straight line motion situations. Other situations are considered after you have had calculus. Finally, notice that the part of the force perpendicular to the motion is discarded—it serves no purpose in the work calculation.

Problem 1: Work

Suppose I am walking due east and a very hard wind is pushing on me 40 LB from $N20^oE$. How much work is done if I move 200 feet?

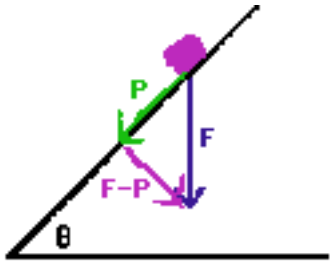
The Solution:

$D = (200 \ feet) \langle 1, 0 \rangle$ and $F = (40 \ pounds) \langle -cos(70^o), -sin(70^o) \rangle$.

So the work is $D \cdot F = -8000 \, cos(70^o) \ foot \ pounds$ or roughly 2736 $foot \ pounds$.

The Inclined Plane

If an object is sitting on an **inclined plane** the force of gravity is pulling straight down. But "straight down" is not a direction that the object can move. Any push directly into the face of the plane is countered by a corresponding push by the molecules of the surface to prevent the object from sinking in. Because of this, it is only the part of the force of gravity pointing down **along the hill** that can influence the motion of the object. The rest is "wasted" against the **constraint**—the hard surface of the hill.



It is the vector P (shown above) that acts to accelerate the object down the hill and you learn how to calculate the effect of the magnitude of P on the motion in basic physics classes. Sometimes in these classes the part $F - P$ of F that is perpendicular to the hill is discarded. We say the hill is "frictionless." Usually we say this because we don't want to think about **friction**, which is very complicated to understand physically, involving molecular forces and tiny rugosities binding on each other. But in real inclined planes, the push directly into the surface creates friction. $F - P$ causes an "effective" force up the hill and this will completely counteract the downhill force if the downhill force is too small. In many cases, the magnitude of the maximum "up the hill" force that could be generated by friction is $k|F - P|$ where k is a positive constant called the **coefficient of static friction**. k depends on how the surfaces are prepared, the material of which they are composed and a myriad of other complicated factors, and is measured for a specific situation, not calculated.

Problem 2: The Inclined Plain

Find P and $F - P$ in the picture above if the object weighs 100 pounds and the plane is inclined 37^o from horizontal.

The Solution:

To calculate the decomposition we can use ANY vector that points along the hill. The easiest one is $\langle cos(37^o), sin(37^o)\rangle$. So

$$\begin{aligned} P &= \frac{\langle cos(37^o),\ sin(37^o)\rangle \cdot \langle 0, -100\rangle}{\langle cos(37^o),\ sin(37^o)\rangle \cdot \langle cos(37^o), sin(37^o)\rangle}\ \langle cos(37^o), sin(37^o)\rangle \\ &= -\ 100\ sin(37^o)\,\langle cos(37^o), sin(37^o)\rangle\ \approx\ 60.18\,\langle cos(37^o), sin(37^o)\rangle \\ &\approx\ \langle -48.06\ , -36.22\rangle \\ F - P &=\ \langle 0, -100\rangle - 100 sin(37^o)\,\langle cos(37^o), sin(37^o)\rangle \\ &\approx\ \langle 0, -100\rangle - \langle -48.06, -36.22\rangle = \langle 48.06, -63.78\rangle \\ &\approx 79.86\,\langle cos(-53^o), sin(-53^o)\rangle \end{aligned}$$

If all you needed was the magnitude of these two vectors, that is easier:

$$|P| = |F| sin(\theta) \qquad \text{and} \qquad |F - P| = |F| cos(\theta).$$

In our case that means $|P|$ is about 60.18 pounds and $|F - P|$ is approximately 79.86 pounds.

Problem 3: The Inclined Plain

Consider the situation from Problem 2. Suppose the coefficient of static friction is $k = .65$. Will the object slide down the hill?

The Solution:

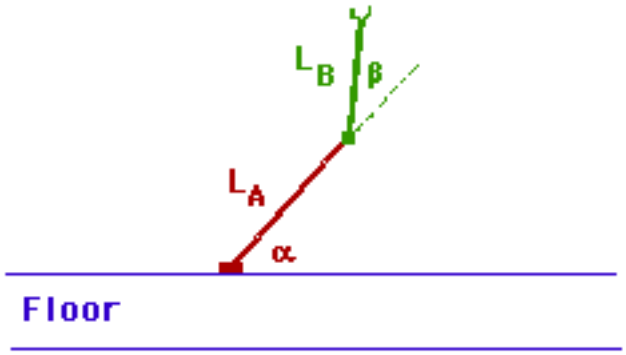
The maximum force that could be generated by friction has magnitude $(.65)(79.86) \approx 51.9$ pounds up the hill.

This is less than the magnitude of P so there will be a net force down the hill.

If there is a net force there is always a change in the motion, so the object will move.

A Tool on a Robot Arm

If you are trying to **control** the activity of a tool at the end of a robot arm, the first step is to have a means of sensing the location of the tool. One way of doing this is to have a device that measures the angle of each joint relative to the previous arm segment or an anchor point. The picture below indicates a robot arm confined to move in a vertical plane with two arm segments, together with the angles α and β which can be measured by the sensors at each joint: one sensor at the shoulder measures the angle between the first arm segment and the floor and one at the elbow which measures the angle between the upper arm and the forearm. These measurements are called **feedback**. The robot arm segments have fixed lengths, L_A and L_B respectively. Let us also suppose that the mechanism of the joint allows you to control the angles, but that the angles must be positive and cannot exceed 180^o.



If a specific angle feedback arrives at the controller computer, where is the tool? Using vectors, the tool is at the tip of the resultant vector

$$L_A \langle cos(\alpha), sin(\alpha)\rangle + L_B \langle cos(\alpha+\beta), sin(\alpha+\beta)\rangle .$$

9.1. ***Exercise***. *If the shoulder-to-elbow part of the arm has length* 2 *and the elbow-to-tool part has length* 1*:*

(i) Describe the part of the plane which is accessible to the tool.

(ii) Will there be more than one way to get to any accessible point?

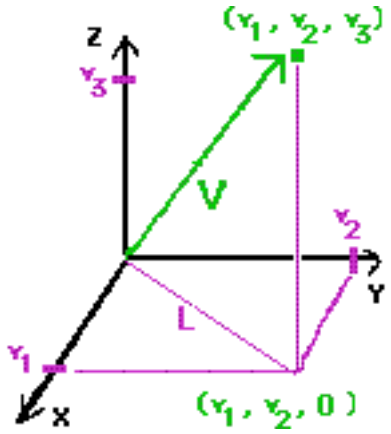
*(iii) * Suppose that you want to get to a specific point with the tool. How should you control the arm to get there? (I am asking for α and β. hint: Represent the point as a $\langle cos(\gamma), sin(\gamma)\rangle$ where $a > 0$ and $0 \leq \gamma \leq 180^o$.)*

---

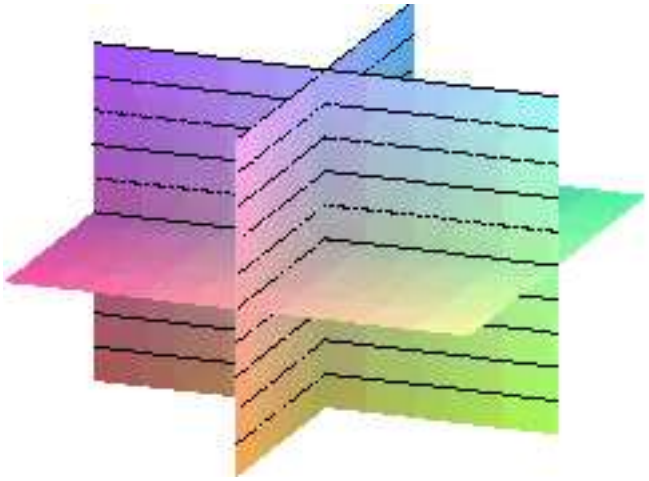
10. Vectors in Three Dimensions

Locations in the plane require two numbers (X and Y coordinates, for instance) to describe. Sometimes people say that the plane has **two dimensions**, or is **2D** because of that. Locations in space require three, commonly called X, Y and Z. Space is said to have **three dimensions**, or to be **3D**. When we describe locations by the X and Y coordinates in $2D$ or X, Y and Z coordinates in $3D$ we are said to be using **rectangular coordinates.** Vectors in space require three components too: the X, Y and Z components of the tip of the arrow when the tail is at the origin. When we make this identification for a particular application of vectors we assume concepts of origin, perpendicularity and distance to be agreed upon and reflected in the identification.

Vectors $V = \langle v_1, v_2, v_3 \rangle$ still have a length, found by using Pythagoras twice in the picture found below—first to find the length of L. You will find that the **magnitude** of a vector V, denoted $|V|$, is $\sqrt{v_1^2 + v_2^2 + v_3^2}$.



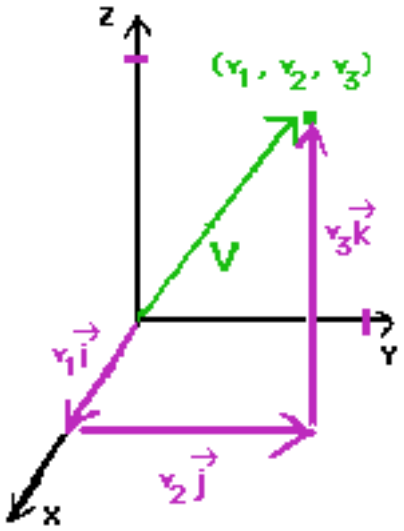
There are three special planes, called the **coordinate planes**, to which we refer in order to get "oriented" in a $3D$ picture. These are called the XY, YZ and XZ planes, and correspond to all the points with Z, X and Y coordinates, respectively, equal to 0. They divide space into eight **octants**.



There are three special vectors $\vec{i}$, $\vec{j}$ and $\vec{k}$ (not just $\vec{i}$ and $\vec{j}$ anymore) pointing along the three coordinate axes:

$$\vec{i} = \langle 1, 0, 0 \rangle , \qquad \vec{j} = \langle 0, 1, 0 \rangle \qquad \text{and} \qquad \vec{k} = \langle 0, 0, 1 \rangle .$$

This second definition for $\vec{i}$ and $\vec{j}$ invites confusion. However, with use you will come to consider that a "feature" rather than a "bug."



Any vector $V = \langle v_1, v_2, v_3 \rangle$ in space can be written also as

$$V = v_1\vec{i} + v_2\vec{j} + v_3\vec{k}$$

and both notations for vectors in space are in common use. One advantage of this notation is facilitate the transition from $2D$ to $3D$ which is frequently useful in applications. A vector $V = \langle v_1, v_2 \rangle$ in $2D$ is different from the vector $\langle v_1, v_2, 0 \rangle$ in space, but sums and multiples of vectors in the XY plane in space all correspond to the same operations on the related vectors in $2D$. You have merely changed your description of the arrows, not the arrows themselves. By using the notation $V = v_1\vec{i} + v_2\vec{j}$ you can make the move from $2D$ to the XY plane in $3D$ by merely redefining which $\vec{i}$ and $\vec{j}$ you mean in a formula. Mathematicians don't like this kind of thing very much, but Engineers seem to like it a lot.

When V is any nonzero $3D$ vector it is still true that $\frac{V}{|V|}$ is a unit vector. $\frac{V}{|V|}$ is called the **direction vector for** V.

So any nonzero vector V can be written as $V = |V|\frac{V}{|V|}$. The vector part has its nose on the **unit sphere** which consists of all points (X, Y) with $X^2+Y^2+Z^2 = 1$. This part indicates direction of V, while $|V|$ carries the magnitude information about V .

We define the **dot product** of two vectors $V = \langle v_1, v_2, v_3 \rangle$ and $W = \langle w_1, w_2, w_3 \rangle$ to be

$$V \cdot W = v_1w_1 + v_2w_2 + v_3w_3.$$

It is still true but a bit harder to show that

$$V \cdot W \ = |V||W|cos(\theta)$$

where θ is the angle between V and W as measured in a plane containing both vectors. There is a discussion of this in the next section if you are interested.

If you accept that formula as fact, then there is a simple meaning for the entries in

$$\frac{V}{|V|} = \left\langle \frac{v_1}{|V|}, \frac{v_2}{|V|}, \frac{v_3}{|V|} \right\rangle.$$

They are $\langle cos(\theta_1), cos(\theta_2), cos(\theta_3)\rangle$ where the angles θ_1, θ_2 and θ_3 are the angles between V and the coordinate axes! To show this, dot $\frac{V}{|V|}$ against $\vec{i}$, $\vec{j}$ and $\vec{k}$ one at a time. The numbers $cos(\theta_1)$, $cos(\theta_2)$ and $cos(\theta_3)$ are called **direction cosines** for the vector.

It is just as easy in $3D$ as in $2D$ to come up with **vectors perpendicular to a given one:** simply select them to have dot product 0 with the first. For example $\langle 3, 6, -1\rangle$ is normal to both $\langle -6, 3, 0\rangle$ and $\langle 0, 1, 6\rangle$.

The formulas for decomposition (and the proof that the decomposition has the expected properties) and the formulas involved in parameterized constant velocity motion are unchanged in three dimensions!

To **decompose** V into the sum of a vector (call it P) which is a multiple of W and another perpendicular to W we proceed as before:

Let $P \ = \left(\frac{V \cdot W}{W \cdot W}\right) W$. Then $V \ = \ P + (V - P)$ is the decomposition.

Dotting P against $V - P$ to yield 0 provides the check for your arithmetic.

To describe **constant velocity motion** the formula is still:

$$Q(t) = P + tV$$

where P is the position at time $t = 0$ and V is the velocity. The length of V is the speed.

Eliminating the parameter can be done in $3D$ too, although this has less utility than in $2D$. Essentially, you describe the geometrical track as the solution of a system of two linear equations in the three variables, X, Y and Z.

A parametric equation in $3D$:

$$Q(t) = \langle X, Y, Z\rangle = P + tV$$

yields the 3 equations:

$$X = p_1 + tv_1$$

$$Y = p_2 + tv_2$$

$$Z = p_3 + tv_3$$

Solving for t in one of these equations and **eliminating this parameter** from the other two gives the system of two linear equations in three unknowns: a line in space.

If you can find two vectors N_1 and N_2 which are not multiples of each other and which are both normal to V you can get formulas for a system of two equations (whose solution is the line) with less effort.

$$(Q-P)\cdot N_1 = 0$$
$$(Q-P)\cdot N_2 = 0.$$

This system is called the **normal form** for the line.

Problem 1: A Decomposition

Decompose $\langle 4,5,1\rangle$ into the sum of a vector which is a multiple of $\langle 1,2,3\rangle$ and another vector perpendicular to $\langle 1,2,3\rangle$.

The Solution:

$$\begin{aligned} P &= \left(\frac{\langle 4,5,1\rangle\cdot\langle 1,2,3\rangle}{\langle 1,2,3\rangle\cdot\langle 1,2,3\rangle}\right)\langle 1,2,3\rangle \\ &= \left(\frac{4+10+3}{1+4+9}\right)\langle 1,2,3\rangle = \left\langle \frac{17}{14},\frac{34}{14},\frac{51}{14}\right\rangle. \\ V-P &= \left\langle \frac{56}{14},\frac{70}{14},\frac{14}{14}\right\rangle - \left\langle \frac{17}{14},\frac{34}{14},\frac{51}{14}\right\rangle = \left\langle \frac{39}{14},\frac{36}{14},\frac{-37}{14}\right\rangle. \end{aligned}$$

Verification:

$$P\cdot(V-P) = \left\langle \frac{17}{14},\frac{34}{14},\frac{51}{14}\right\rangle \cdot \left\langle \frac{39}{14},\frac{36}{14},\frac{-37}{14}\right\rangle = \frac{663}{196}+\frac{1224}{196}-\frac{1887}{196} = 0.$$

Problem 2: Parametric Constant Velocity Motion

(i) Write a parametric vector equation for the position of an object moving with constant velocity and with position vector $\langle -3,0,6\rangle$ at time 5 and $\langle 12,6,18\rangle$ at time 8.

(ii) When will it run into the plane $X+Y+Z=100$?

(iii) Eliminate the parameter to give a system of equations for this line.

The Solution:

The movement in the 3 seconds from $t=5$ to $t=8$ is:

$$\langle 12,6,18\rangle - \langle -3,0,6\rangle = \langle 15,6,12\rangle.$$

The movement each second is one third of this:

$$(1/3)\langle 15,6,12\rangle = \langle 5,2,4\rangle.$$

The position at time 0 is:

$$\langle -3,0,6\rangle - 5\langle 5,2,4\rangle = \langle -28,-10,-14\rangle.$$

So the equation for parametric motion can be written in three ways:

$$Q(t) = \langle -28,-10,-14\rangle + t\langle 5,2,4\rangle$$

or

$$Q(t) = \langle -28+5t,-10+2t,-14+4t\rangle$$

or

$$X = -28 + 5t,\ Y = -10 + 2t \text{ and } Z = -14 + 4t$$

whichever you prefer.

It remains to solve the second part of the question: When will it hit "The Wall?"

$$X + Y + Z = (-28 + 5t) + (-10 + 2t) + (-14 + 4t) = 100$$

which implies $t = 152/11$ seconds is the time when it "hits the wall."

Finally, we eliminate t: The vectors $\langle -2, 5, 0\rangle$ and $\langle 0, 4, -2\rangle$ are both perpendicular to the velocity vector $\langle 5, 2, 4\rangle$. So the equations

$$(\langle X, Y, Z\rangle - \langle -28, -10, -14\rangle) \cdot \langle -2, 5, 0\rangle = 0$$
$$(\langle X, Y, Z\rangle - \langle -28, -10, -14\rangle) \cdot \langle 0, 4, -2\rangle = 0$$

give a non-parametric system for this line, which is the solution to the third part of the problem. These equations reduce to the system:

$$-2X + 5Y = 6 \text{ and } 4Y - 2Z = -12.$$

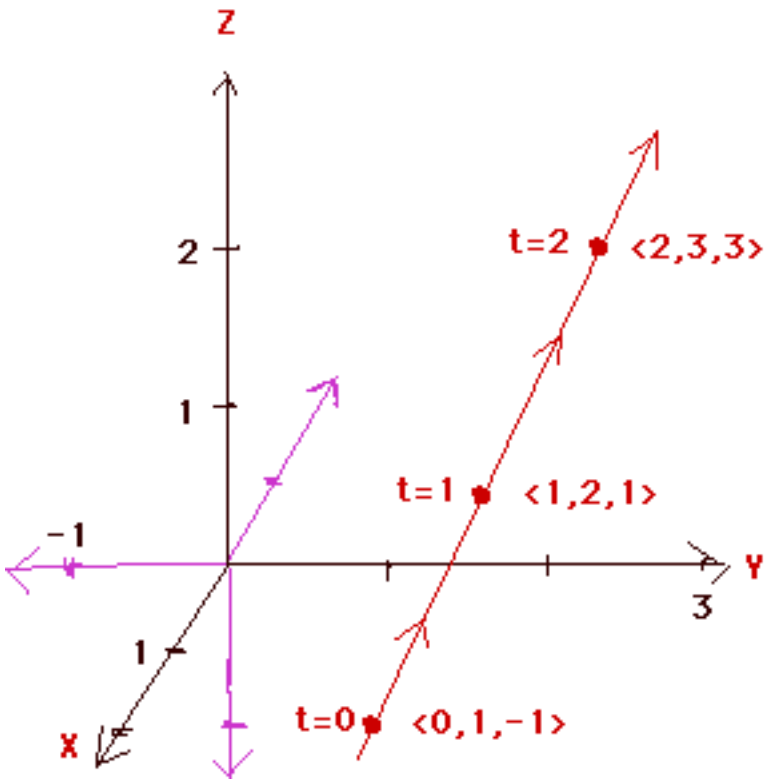
Problem 3: Draw a Picture of Parametric Constant Velocity Motion in $3D$.

Draw a picture to describe the constant velocity motion that has position vector $\langle 0, 1, -1\rangle$ at time $t = 0$ and $\langle 1, 2, 1\rangle$ at time $t = 1$.

The Solution:

We simply plot several points in a representation of a $3D$ coordinate system and connect the dots in the obvious way. There are a couple of issues to be emphasized. First, you should pick a scale on the axes so that the points you are plotting are well separated. You might want to adjust the "point of view" from which you look at the axes when you draw your picture to make this easier. Second, you should label your points with times to indicate the direction of motion. Third, don't expect too much. Unless you can rotate the picture around and see your picture from various angles the fact that you are attempting to represent three dimensions on a two dimensional piece of paper means you will lose a lot of information. Computer algebra systems are a HUGE benefit when graphing $3D$ pictures for that reason.

In our problem, the velocity vector is $\langle 1, 1, 2\rangle$ so as time passes we add on increasing multiples of that vector.



10.1. ***Exercise***. *(i) Decompose* $\langle -4, 1, 6\rangle$ *into the sum of a vector which is a multiple of* $\langle -1, 8, 1\rangle$ *and another perpendicular to* $\langle -1, 8, 1\rangle$.

(ii) Decompose $\langle 0, 1, -1\rangle$ *into the sum of a vector which is a multiple of* $\langle -1, 1, 2\rangle$ *and another perpendicular to* $\langle -1, 1, 2\rangle$.

(iii) Write a parametric vector equation for the position of a moving object located at $\langle 0, 5, 1\rangle$ *at time* 0 *and with velocity vector* $\langle -1, 7, 7\rangle$. *Where will it be at time* 10*? Where was it at time* -5*? Draw a picture of this motion.*

(iv) The vectors $\langle 5, 3, 0\rangle$ *and* $\langle 0, 7, -5\rangle$ *are both perpendicular to* $\langle -3, 5, 7\rangle$. *Following that pattern, find two vectors perpendicular to* $\langle a, b, c\rangle$.

(v) Write a parametric vector equation for the position of a moving object whose position vector is $\langle -2, 9, 1\rangle$ *at time* 5 *and* $\langle 12, 16, 8\rangle$ *at time* 12. *Draw a picture of this motion. When will it "hit a roof" at* $Z = 120$*? Finally, eliminate the parameter to form a system of two equations for this line.*

10.2. ***Exercise***. * *(i) We suppose* $Q(t) = P + tV$ *to be a parametric vector equation with geometrical track which we denote* L. *Show that* $\overline{Q} = \overline{P} + t\overline{V}$ *has the*

same geometrical track as Q, where

$$\overline{P} = P - \frac{P \cdot V}{V \cdot V}\ V \quad \textit{and} \quad \overline{V} = \frac{V}{|V|}\ .$$

Note also that $\overline{P} \cdot \overline{V} = 0$ and $\overline{V} \cdot \overline{V} = 1$.

(ii) The point on L nearest to the origin has position vector $\overline{Q}$ at which $\overline{Q} \cdot \overline{Q}$ is a minimum. But $\overline{Q} \cdot \overline{Q} = \overline{P} \cdot \overline{P} + t^2$ which obviously has a unique minimum for $t = 0$. So the near spot has position vector

$$\overline{Q}(0) = \overline{P} = P - \frac{P \cdot V}{V \cdot V}\ V.$$

This is the part of P perpendicular to V and is the one and only position vector of a point on L which is perpendicular to the line.

(iii) Show that the point on L closest to the point with position vector A has position vector

$$B = P - \frac{(P - A) \cdot V}{V \cdot V}\ V.$$

B is the only position vector of a point on L for which $B - A$ is perpendicular to the line. Draw pictures to convince yourself of why this should be true.

---

11. Angles in Higher Dimensions

We are going to do some calculations here to make it seem reasonable to use the dot product to determine the angles between vectors in higher dimensions and not just the 2 dimensional XY plane.

We will presume V and W are two vectors in some higher dimensional setting, such as 3 dimensional space. We presume that neither is a multiple of the other. The case where one is a multiple of the other is left as an exercise. We presume that distances in this setting are defined by an extension of the old Pythagorean distance formula. So the distance between the points $(a_1, a_2, \dots)$ and $(b_1, b_2, \dots)$ is given by

$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \dots}\ .$$

The three dots signify that you should keep on going in the same pattern till you run out of coordinates.

So the dot product of the vectors V and W is $V \cdot W = v_1 w_1 + v_2 w_2 + \dots$ and the magnitude of a vector such as V is

$$|V| \ = \ \sqrt{v_1^2 + v_2^2 + \dots} \ = \ \sqrt{V \cdot V}\ .$$

Our goal is to show that with these assumptions it makes sense to write

$$V \cdot W \ = |V||W| cos\,(\theta)$$

where θ is the angle between V and W.

Let's define the three vectors:

$$\begin{aligned} P &= \left(\frac{V \cdot W}{W \cdot W}\right) W \quad \text{and} \\ Q &= \frac{V - P}{|V - P|} \quad \text{and} \\ T &= \frac{W}{|W|} \end{aligned}$$

It is an interesting exercise to show that $Q \cdot T = 0$.

It is obvious that $T \cdot T = 1$ and $Q \cdot Q = 1$.

We now define a function F that takes ordered pairs in the plane to points in the space where the vectors V and W live.

We define $F(r, s) = (rq_1 + st_1, rq_2 + st_2, \dots)$, where the q_i and t_i come from the coordinates of Q and T respectively.

The following messy calculation is the key result.

The square of the distance between two points $F(a, b)$ and $F(r, s)$ is

$$\begin{aligned} &(aq_1 + bt_1 - rq_1 - st_1)^2 + (aq_2 + bt_2 - rq_2 - st_2)^2 + \dots \\ &= ([a - r]\, q_1 + [b - s]\, t_1)^2 + ([a - r]\, q_2 + [b - s]\, t_2)^2 + \dots \\ &= (a - r)^2 \left(q_1^2 + q_2^2 + \dots\right) + 2\,(a - r)\,(b - s)\,(q_1 t_1 + q_2 t_2 + \dots) \\ &\qquad + (b - s)^2 \left(t_1^2 + t_2^2 + \dots\right) \\ &= (a - r)^2 + (b - s)^2 \\ &\text{(because } Q \cdot T = 0,\ Q \cdot Q = 1 \text{ and } T \cdot T = 1.) \end{aligned}$$

This is just the square of the distance between (a, b) and (r, s)! This means that the distances between all pairs of points in the XY plane are unchanged when transported by F.

So any triangle in the XY plane, determined by three specific vertices in the XY plane, is taken by F to a triangle whose vertices are the same distance apart from each other as those of the original triangle. So the two triangles have the same interior angles.

Let $A = \langle |V - P|, V \cdot W/|W| \rangle$ and $B = \langle 0, |W| \rangle$.

The tips (when in standard position) of A and B are taken by F to the standard-position tips of V and W respectively (check that out!)

So the angle between A and B is the same as the angle between V and W.

So $A \cdot B = |A||B| cos(\theta)$ where θ is the angle between V and W.

The result now follows from a little algebra:

$$\begin{aligned} V \cdot W &= A \cdot B = |A||B| cos\,(\theta) \\ &= \sqrt{(V-P)^2 + \frac{(V \cdot W)^2}{W \cdot W}}\ |W|\ cos\,(\theta) \\ &= \sqrt{V \cdot V - 2V \cdot W \frac{V \cdot W}{W \cdot W} + \frac{(\ V \cdot W)^2}{(\ W \cdot W\ )^2}\ W \cdot W + \frac{(\ V \cdot W)^2}{W \cdot W}}\ |W|\ cos\,(\theta) \\ &= \sqrt{V \cdot V}\ |W|\ cos\,(\theta) \quad = \quad |V|\ |W|\ cos\,(\theta)\,. \end{aligned}$$

11.1. ***Exercise***. *(i) Under what conditions, exactly, will the equation*

$$|V+W|^2 = |V|^2 + |W|^2$$

be true? (hint: $|V+W|^2 = (V+W) \cdot (V+W)$*.)*

(ii) If the angle between V *and* W *is* θ*, what is the angle between* $-V$ *and* W*?*

*(iii) * Show that* $|\ |V|\ -\ |W|\ | \leq |V+W| \leq |V| + |W|$*. This is called the* ***Triangle Inequality****. What does this inequality have to do with triangles? When, exactly, will you have equality on one side or the other?*

11.2. ***Exercise***. *Show that* F *takes any point on a line segment between* (a,b) *and* (r,s) *to a point on the line segment between* $F(a,b)$ *and* $F(r,s)$*.*

---

12. The Cross Product

The **cross product** is another way of multiplying vectors, like the dot product. In this case, however, the "answer" is another vector rather than a number. We define this product for vectors in $3D$ only—if you have $2D$ vectors involved you must think of them as $3D$ vectors with zero third component to use the cross product.

If $V = \langle v_1, v_2, v_3 \rangle$ and $W = \langle w_1, w_2, w_3 \rangle$ we define

$$V \times W = (v_2w_3 - v_3w_2)\vec{i} + (v_3w_1 - v_1w_3)\vec{j} + (v_1w_2 - v_2w_1)\vec{k}.$$

Notice that the first entry in the product does NOT involve the first entry of either vector. The second entry in the product does NOT involve the second entry of either vector. The third entry in the product does NOT involve the third entry of either vector. Also, the positive term in the first and third entry in the product has the V and W entries "in order" while the middle term has them backwards.

If V and W are in the XY plane note that $V \times W$ is a multiple of $\vec{k}$.

Another interesting and useful product is called the **triple scalar product**. It is defined for triples of $3D$ vectors P, V and W by:

$$\det(\mathbf{P}, \mathbf{V}, \mathbf{W}) = P \cdot (\ V \times W\ )\,.$$

The triple scalar product is a number and can be calculated by dotting the first vector against the cross product of the last two vectors. The det notation comes from the fact that, for those among you who know about **determinants**, it is the determinant, usually denoted $\det A$, of the matrix A where

$$A = \begin{pmatrix} p_1 & p_2 & p_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{pmatrix}$$

12.1. ***Exercise***. *(i) Show that* $V \times W = -\,W \times V$.

(ii) Show that $(kV) \times W = k\,(V \times W) = V \times (kW)$ *for any real number* k.

(iii) Show that $(\,P + V\,) \times W = P \times W + V \times W$ *and* $W \times (\,P + V\,) = W \times P + W \times V$.

(iv) Show that $V \times W$ *is perpendicular to both* V *and* W.

(v) Show that if you switch any two of the vectors in a triple product, the sign of its value switches. For example $\det(P, V, W) = -det\,(V, P, W)$.

*(vi) * Suppose* $V \times W \neq 0$ *and* N *is a vector with* $N \cdot V = 0$ *and* $N \cdot W = 0$. *Show that* N *must be a multiple of* $V \times W$.

This last fact is an algebraically messy exercise involving coordinates, but it is quite important. Here is a hint.

Write out the components of $C = V \times W$. *Now consider the equations* $N \cdot V = 0$ *and* $N \cdot W = 0$. *They give you two equations involving* n_1, n_2 *and* n_3. *Use the "elimination method" three times on these two equations to eliminate* n_1, n_2 *and* n_3, *one at a time. Recognize the messy coefficients as* c_1, c_2 *and* c_3. *This will yield three equations:*

$$c_2 n_3 = c_3 n_2 \quad \text{and} \quad c_3 n_1 = c_1 n_3 \quad \text{and} \quad c_1 n_2 = c_2 n_1.$$

Since $C \neq 0$ *at least one of the components of* C *is nonzero. If* $c_i \neq 0$ *then* $N = \frac{n_i}{c_i} C$.

The facts $(i) - (iii)$ shown in the exercise above combined with the properties of dot products make the triple scalar product another example of a **tensor**. Since there are three vectors involved it is called a 3-tensor. The property from part (v) in this context is called **antisymmetry**. The triple scalar product is an example of an antisymmetric 3-tensor.

12.2. ***Exercise***. *Recall Exercise 5.3. If* $V = \langle v_1, v_2 \rangle$ *and* $W = \langle w_1, w_2 \rangle$ *we defined* $\det(V, W)$ *to be the number* $v_1 w_2 - v_2 w_1$. *Show that for any 2D vectors* V, P *and* W *and any constant* a

$$\det(aV + P, W) = a\ \det(V, W) + \det(P, W) \quad \text{and} \quad \det(V, W) = -\det(W, V).$$

These properties imply that $\det(V, W)$, *defined for pairs of 2D vectors, is an antisymmetric 2-tensor.*

You can show by doing a lot of multiplying and gathering of like terms that

$$(V \times W) \cdot (V \times W) + (V \cdot W)(V \cdot W) = (V \cdot V)(W \cdot W) .$$

We observe that the second and third terms above can rewritten as:

$$(V \times W) \cdot (V \times W) \ + |V|^2 |W|^2 \ cos^2(\theta) = |V|^2 |W|^2 \ ,$$

where θ is the angle between V and W . It follows that:

$$|V \times W| = |V||W| \ sin(\theta) .$$

This is reminiscent of the similar fact about dot products and is just as useful, though in different contexts.

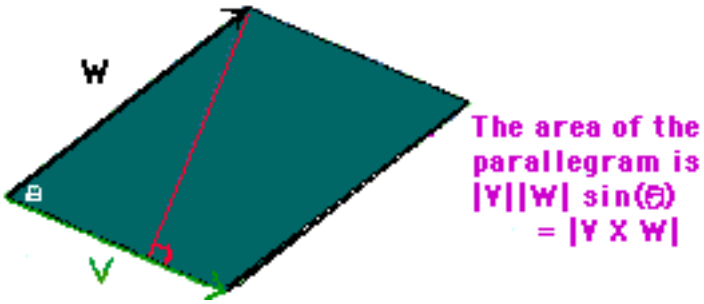
Here we will examine four related geometrical uses for cross product.

Area of a Parallelogram

First we find the **area of a parallelogram**. By dividing the picture below into two triangles we see that the area is just twice "one half the base times the height" which is

$$|V \times W| = |V||W| \ sin(\theta) .$$

Cross product can be used to find the area of this region.



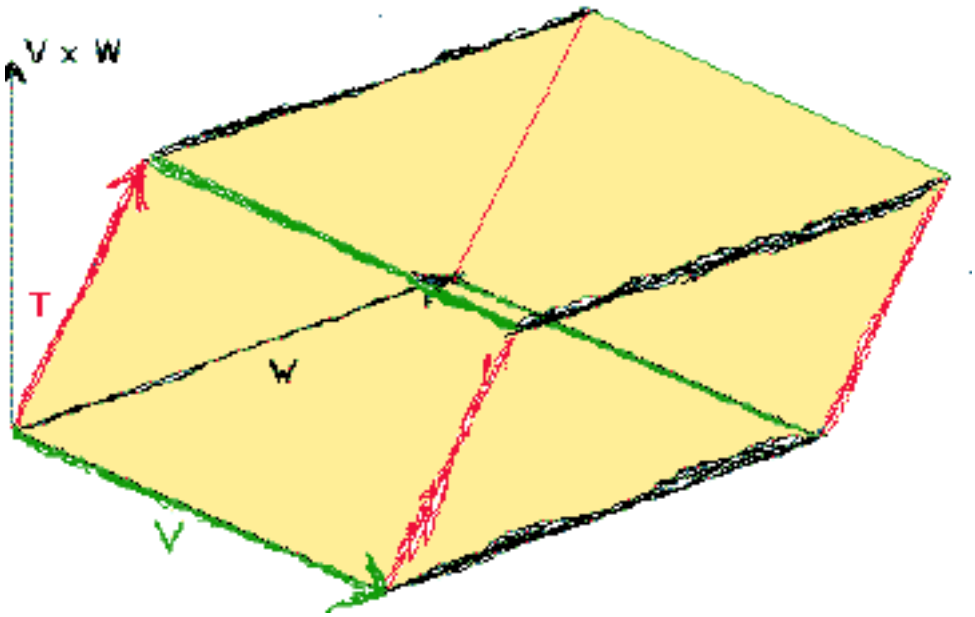
12.3. ***Exercise***. *In the calculation above for the area of the parallelogram, we drew a picture with the angle between V and W less than 90^o. Is the formula still valid if the angle exceeds 90^o?*

Volume of a Parallelepiped

Second, we consider the **parallelepiped** below. In general, parallelepipeds are $3D$ objects bounded by three pairs of parallel parallelogram faces.

Our parallelepiped is a (possibly) bent box determined by the three vectors T, V and W. You can think of it as a deck of parallelogram cards that has been pushed so that the edge of the stack runs at an angle—along the vector T.

If you pushed the stack straight again, the shape of the cards is unchanged and so is the volume of the stack.



If we want the volume of the stack we need to know the height of the stack. Then the volume will be the height times the area of a card. We just found that the area of a card is $|V \times W|$.

The height of the stack is the length of the part of T that is perpendicular to the cards. The vector $\frac{V \times W}{|V \times W|}$ is a unit vector perpendicular to the cards.

So the height of the stack is $\left| T \cdot \frac{V \times W}{|V \times W|} \right|$.

Finally, we have the **volume of the parallelepiped** as

$$\left| T \cdot \frac{V \times W}{|V \times W|} \right| |V \times W| = | \, T \cdot (\, V \times W \,) \, | = | \det (T, V, W) \, |$$

This gives us an interpretation for the triple scalar product as the volume of the shape generated by the three vectors. This turns out to be more useful than you might at first imagine.

The discussion from above might (or might not) constitute a compelling argument for you as to why we call this number the volume of a parallelepiped. In any case, that is *the definition* of volume. It is our dog. We can call it **Fido** if we want to.

At this point we digress to contemplate the **meaning** of procedures such as this. We all have an intuitive concept of the word "volume" as having to do with the extent of a "fat" object, and in the last paragraphs we aligned that concept for

a certain type of fat object with a number. This number is calculated by vector operations on edges. It is (at least) twice removed from our intuition about the extent of any real physical object.

In the first place, we model the edges of some real object with which we have experience by vectors. Then we perform operations to obtain a number.

At risk of boring you with an obvious point, this number is not the "extent" of any object in the world. It is just a number. Its value may be relevant in comparison to other numbers calculated with other vectors with their own imaginary link to still *other* real objects. It is useful to a scientist or engineer only to the extent it gives explanatory or predictive power over events in the world. Does the number correspond to intuition built on other cases already understood? Does it invite extension to cases not yet understood? Is it simple, or simpler than competing alternatives? If the answers are affirmative, the scientist will carry on, thinking of these numbers as a measure of "fat extent." If not, the model is discarded, at least in some circumstances. Every now and then some model violates a long-held, even cherished, intuition that must, nevertheless, be abandoned based on the shocking success of this model with its peculiar predictions. The creators of the model may win Nobel Prizes.

Mathematicians, on the other hand, might be intrigued by the structure of a model itself, and might not be bothered excessively if it didn't match intuition obtained from "looking out the window" at something real. Surprising or downright weird behavior is not rare in these models. As a means of assigning numbers to vectors, for example, our volume idea from above stands by itself. Trying to puzzle out the details of this structure and the relationships among this and other structures—such as "flat extent" and "straight extent"—could occupy a mathematician for a very long time.

You might regard this as a "division of labor" or even "symbiosis." The mathematicians stock the shelves with well understood and beautiful models, packaged and ready-to-go. They seek to understand the patterns they see as they create these ideas. They build related models, by analogy and lifetimes of effort and inspiration, hunting for unifying ideas and a "big picture" that will simplify needless complexity. The scientists root around on the shelves looking for models that match what they see out the window guided by their esthetic sensibilities and previously acquired intuitions about the world. Experimenters create ever more subtle ways of stealing a glimpse at some previously hidden part of the world. Every now and then some scientist makes a great leap of intuition and builds a brand new model, of a type never thought of by anyone, to try to understand what has been seen. And in this way they inspire and repay their mathematical brothers and sisters.

12.4. ***Exercise***. *A* ***tetrahedron*** *is a four-sided three dimensional object with equilateral triangular faces.*

Show that the vectors

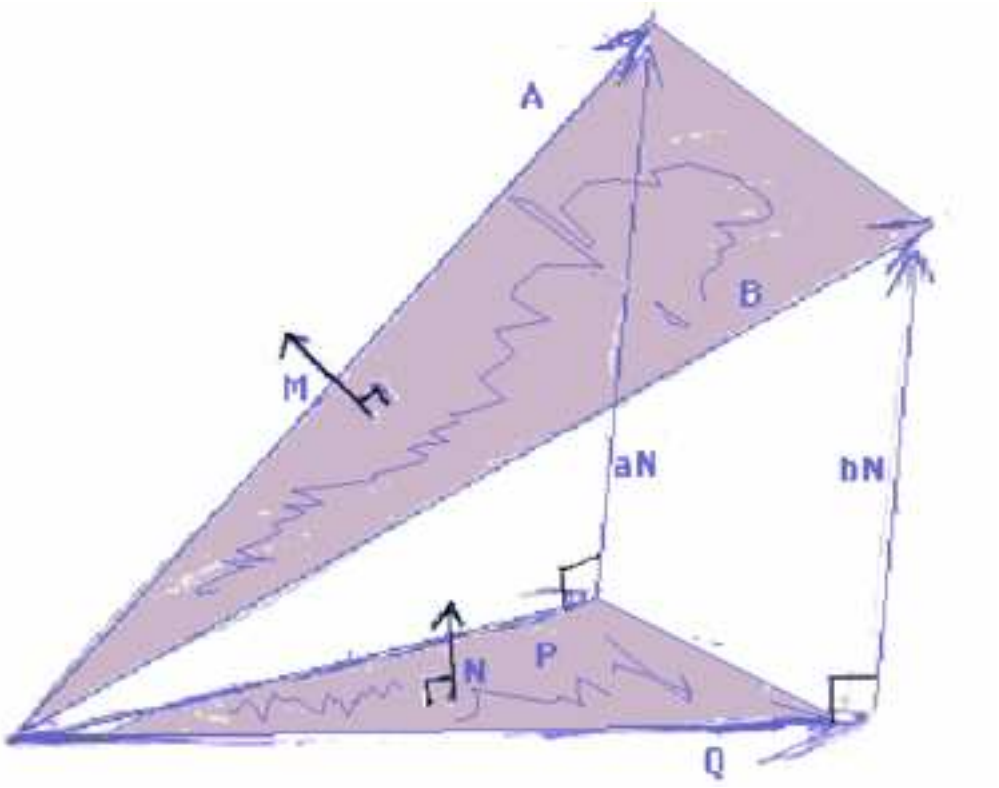
$$\langle 1,0,0\rangle \ , \ \left\langle \frac{1}{2},\frac{\sqrt{3}}{2},0\right\rangle \ , \ \left\langle \frac{1}{2},\frac{\sqrt{3}}{6},\frac{\sqrt{6}}{3}\right\rangle$$

lie on three edges of unit length of a tetrahedron. Conclude that a tetrahedron of edge length d has volume $\frac{d^3\sqrt{2}}{4}$.

Area of a Shadow

As a third example we consider the area of a tilted parallelogram and how that area is related to the **area of a shadow** of this parallelogram on a plane.

Let us presume that P, Q, A and B are vectors and $P \times Q = N \neq 0$ and $A = P + aN$ and $B = Q + bN$ for certain numbers a and b. Let $M = A \times B$. The picture below describes the situation. A and B form a tilted parallelogram with area $|M|$ while P and Q form the shadow parallelogram with smaller area $|N|$.



The **angle between two flat surfaces in 3D** is defined to be the angle between their normal vectors. In our case, the angle between the two parallelograms (or the triangles in the picture) is the angle between M and N. If θ is this angle

we have

$$\begin{aligned} |M||N|cos\,(\theta) &= M \cdot N = (A \times B) \cdot N \\ &= [(P + aN) \times (Q + bN)] \cdot N \\ &= (P \times Q) \cdot N + (P \times bN) \cdot N + (aN \times Q) \cdot N + (aN \times bN) \cdot N \\ &= N \cdot N \quad \text{because the last three terms are 0} \\ &= |N|^2. \end{aligned}$$

We have just shown that

$$|M|cos\,(\theta) = |N| \quad \text{or} \quad |A \times B|cos\,(\theta) = |P \times Q|$$

where θ is the angle between vectors M and N, which are normal to the tilted and shadow parallelograms. In words this is:

(The Area of a Tilted Parallelogram) cos(θ) = Area of the Shadow.

A Consequence of the Pythagorean Theorem

There is an interesting consequence of the Pythagorean Theorem applied to this last calculation. Recall that any vector, M included, can be written as

$$M = |M| \langle cos(\theta_1), cos(\theta_2), cos(\theta_3) \rangle$$

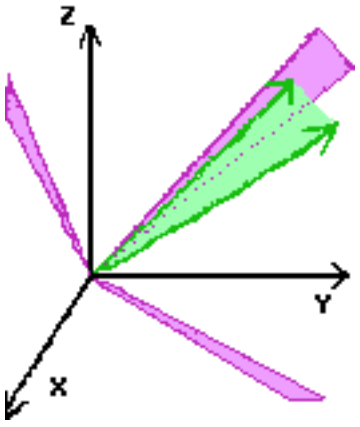
where the angles are those between M and the coordinate vectors $\vec{i}$, $\vec{j}$ and $\vec{k}$. These coordinate vectors are each normal vectors to one of the coordinate planes.

The equation

$$|M|^2 = |M|^2 cos^2(\theta_1) + |M|^2 cos^2(\theta_2) + |M|^2 cos^2(\theta_3)$$

is nothing more than the Pythagorean Theorem in $3D$ applied to M and hardly a surprise. But we have just shown that the area of the shadow of the tilted parallelogram onto the relevant coordinate plane is $|M|cos(\theta_i)$. So the last equation has the following intriguing interpretation.

The Square of the Area of the Tilted Parallelogram
is the Sum of the Squares of the Areas
of the Shadow Parallelograms on the Coordinate Planes.



CHAPTER II

Graphing: A Few Surfaces and Vector Functions
May 27, 2005

13. Surfaces in Three Dimensions and Representations of Planes

Our goal in this section is to think about various ways of representing planes and other surfaces in three dimensions. The pattern should be very reminiscent of the process you went through when you learned how to graph lines in beginning algebra.

If you were taught in the usual progression:

- You first learned about locating points in the plane (a piece of paper) and about creating coordinates of points by choosing axes and a scale and so on.
- Then you examined the relationship between a given first degree equation in two variables and a collection of points plotted on a graph. You came to believe that the solutions to first degree equations would form straight line graphs and ruminated upon what made you think a picture was a straight line. You learned about slope.
- Finally you learned how to go from "geometry" (that is, a graph) to an equation. You used one of several forms for your equation depending on the information readily at hand, such as the slope-intercept form or the point-slope form.

The interplay between the picture (good for thinking about, visual intuition can be used, can be a visually compelling map of things in the world) and the equation (good for exact calculations) was fruitful.

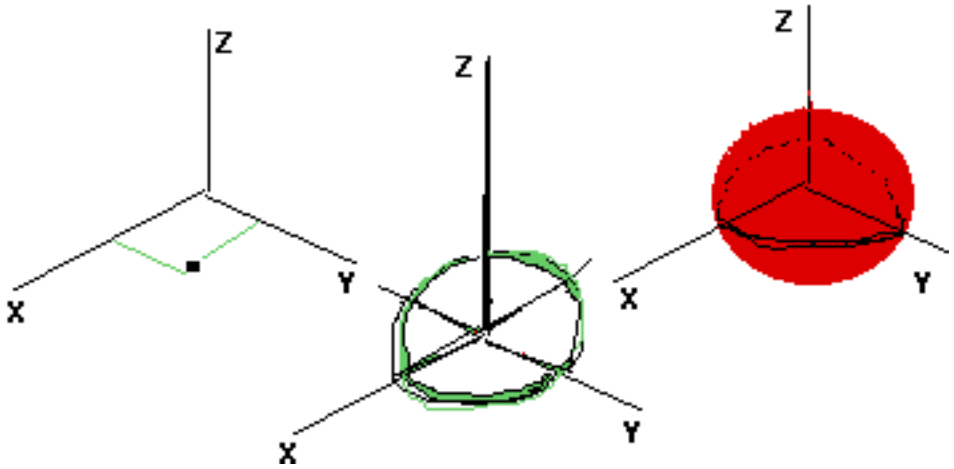
That is exactly how we shall proceed here in three dimensions with our added tools of vectors and/or parameters.

We can graph an equation in three variables, such as $X - 2Y + 3Z = 5$, in three dimensions by plotting lots of points: pick an X and Y value, work out Z from the equation, put a dot on your graph with those coordinates, repeat until tired. Eventually you will have plotted so many points that they will be hard to distinguish individually and your eye will perceive the points, all together, as a shape of some kind. The process of plotting all the solutions may be tedious and the utility of a picture of a $3D$ object on a piece of paper is somewhat compromised by problems with "perspective."

An equation like $(X-1)^2 + (Y-1)^2 + Z^2 = 0$ has a single solution: just one point.

An equation like $X^2 + Y^2 - 1 = \sqrt{-Z^2}$ has solutions consisting of the unit circle in the XY plane.

Others, like $X^2 + Y^2 + Z^2 = 1$, form an extended shape of the type we think of as a **surface**, in this case the unit sphere in $3D$.



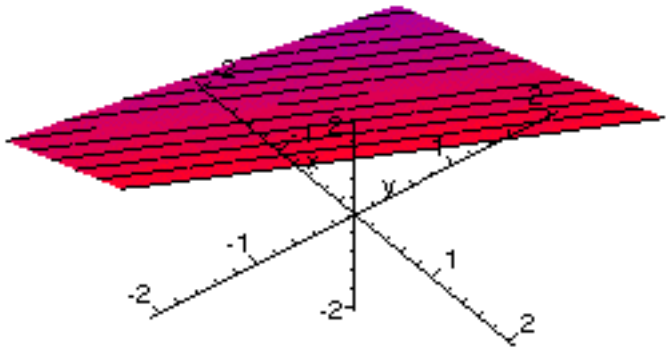
We will see that certain types of equations in three variables generate surfaces and some of these reproduce our thinking about flatness while others do not. We then explore what "flatness" means in terms of vector operations. We will call the flat surfaces **planes.**

As the last step, we have in mind planes as geometrical creatures and wish to represent them as equations in different ways, depending on the information that is readily available about the plane.

This will allow us to do exact calculations.

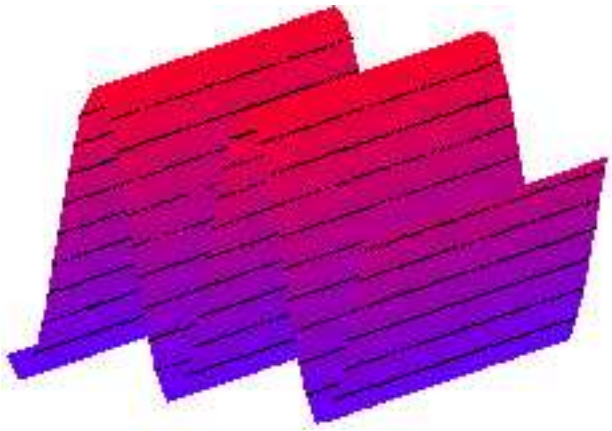
You may recall being told that the graph of all the (X, Y, Z) triples that satisfy a first degree (also called linear) equation involving three variables, such as $X - 2Y + 3Z = 5$, is a plane.

It is not completely obvious that the graph will be "flat." If you examine what happens when you cut through the graph at constant height—that is, pick a particular Z—you will find that the graph at that height is the line $X - 2Y = 5 - 3Z$.

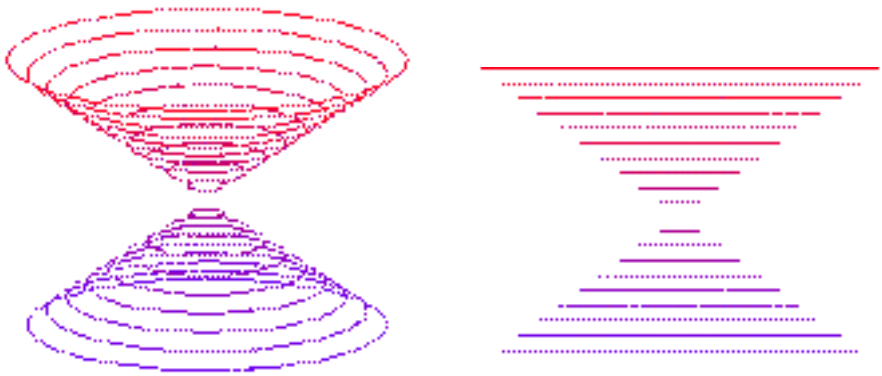


These lines are all of the same slope but cross the YZ plane (when X is 0) at various places, depending on Z. If you plot several of these on a careful graph it certainly seems that they are lining up to form a flat surface. But you can glue lines

together in many ways to form surfaces. Such surfaces are called **developable**. For example a wavy graph below is composed of parallel lines.

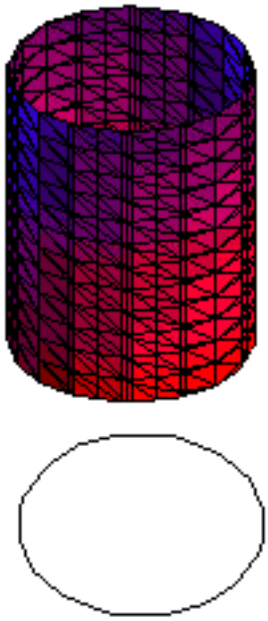


If you have access to **Maple** or **Mathematica** or some other computer math utility you can easily plot these graphs and rotate them around and look at them from various angles, illuminated as if from colored lights positioned at various places, or colored according to height. You can get a real feel for quite complicated surfaces. The "learning curve" to get up and running is very short, the cost is modest and they run on virtually every desktop system. If you want to learn how to use a computer algebra utility, now would be the time to begin. I have included instructions about how to create some of the graphics you see in the text in endnotes referenced by superscript here and there as they occur.



A **cone** is given by the equation $X^2 + Y^2 = Z^2$.

In the picture[1] above, at each constant Z you are looking at a circle of radius $|Z|$. This cone is a collection of lines crossing at the origin. You can see them by looking at the picture from a direction perpendicular to the Z axis. The crossed lines are the edges of the stacked circles.

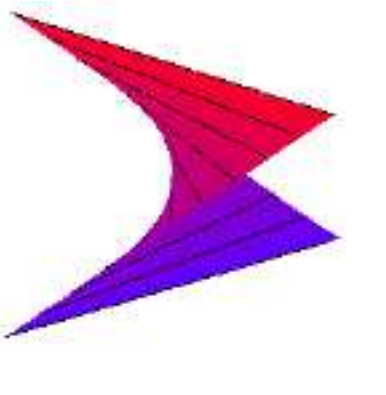


A **cylinder** is given[2] by an equation such as

$$X^2 + Y^2 = 1.$$

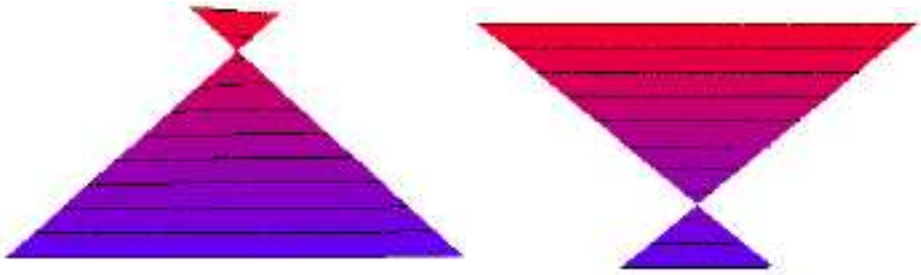
Since Z is not mentioned, it is not restricted.

This is a bunch of vertical lines arranged in a circle. By looking down the axis from above you can see the "line segment ends" form a circle.



Still another example is provided by $Y = XZ$. This is called a **helicoid**. At each constant Z slice, the graph[3] is a line with a different slope.

For constant X and for constant Z this equation generates lines. But the graphs on constant Y slices are hyperbolas.

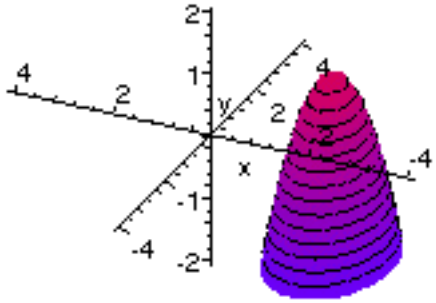


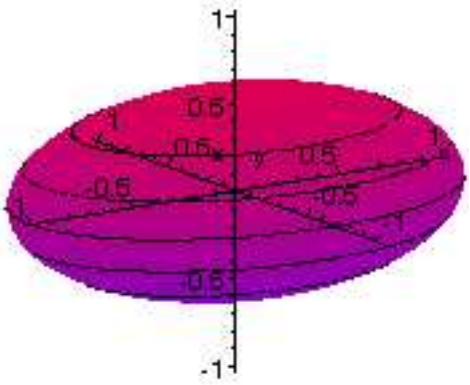
The two pictures above are seen from a vantage point looking straight along the line at different heights along the Z axis.

Below we find[4] an "off-center bump" surface.

The equation is $(X-1)^2 + 2(Y+2)^2 = -Z + 1$.

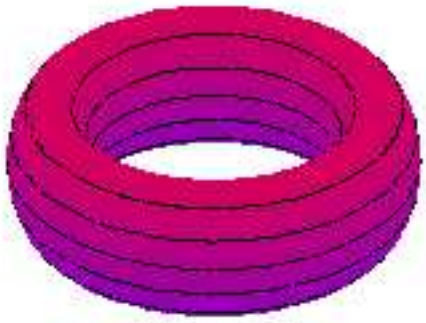
Horizontal cuts are ellipses. Cuts parallel to the XZ or YZ planes are parabolas. It is called a **paraboloid**.





The graph[5] you see above is called an **ellipsoid**. Taking a narrow slice through the figure with any plane parallel to a coordinate plane yields an ellipse. The formula is

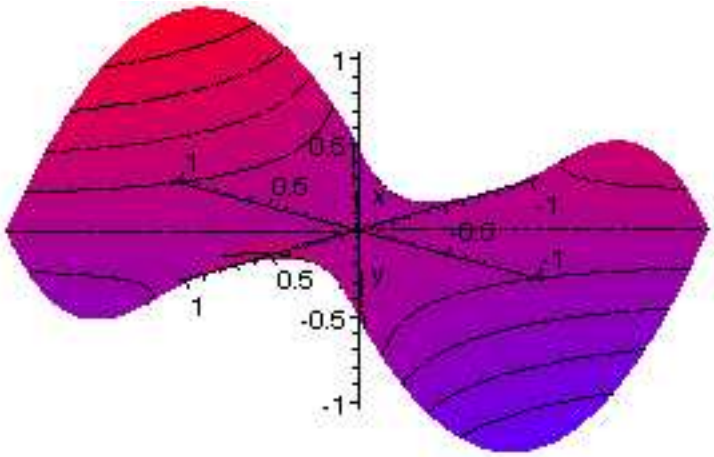
$$X^2 + 2Y^2 + 3Z^2 = 1.$$



The figure[6] above is called a **torus**. Its formula is

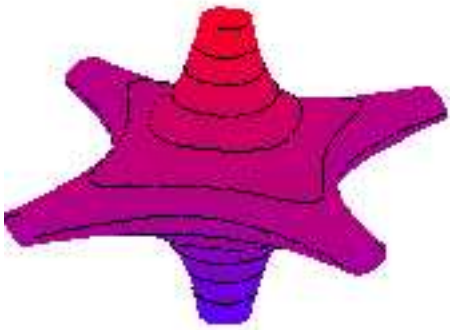
$$\left(X^2 + Y^2 + Z^2 - 4.25\right)^2 = 16\left(.25 - Z^2\right).$$

If you slice through with any plane containing the Z axis you get two circles.



This one[7] is called the **Monkey Saddle**, for reasons that will become clear if you imagine the disposition of a monkey's tail should it decide to ride a horse. Its formula is

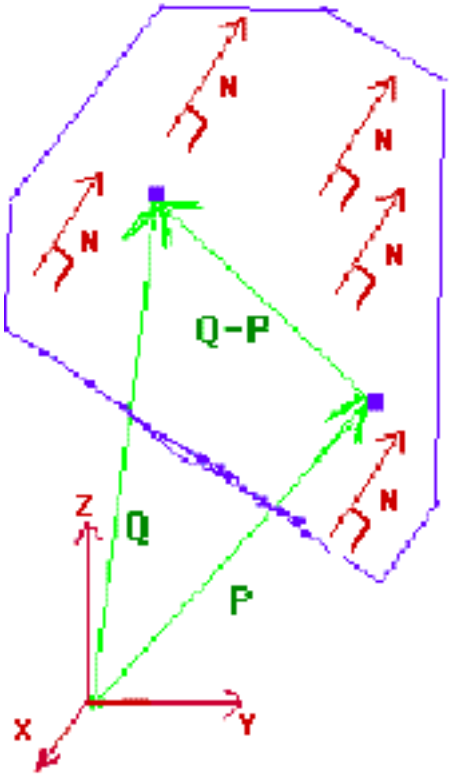
$$X^3 - XY^2 = Z.$$



I don't think this one[8] has a name. It makes the coordinate axes, particularly near the origin, fat. Its formula is

$$X^2Y^2 + X^2Z^2 + Z^2Y^2 = 1.$$

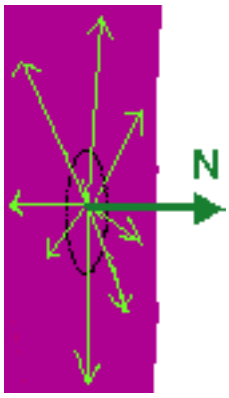
You will notice that all of these non-flat examples correspond to second degree or higher—not first degree—equations.



You should do enough graphing (of these and one or two more first degree equations) so that you become convinced that first degree equations in three variables yield flat surfaces, while others do not. Several, at least, should be done by hand, from more than one perspective. Don't shortcut this process. The intuition you develop here will be needed later. You will find that the calculations we learn about, when done right, are all amazingly easy. Figuring out *which* calculation to do is the hard part, and depends on this intuition.

Let's go back now to the graph of $X - 2Y + 3Z = 5$.

$(2, 0, 1)$ is a solution of this equation. Let $Q = \langle X, Y, Z\rangle$ be a position vector for a generic point (X, Y, Z). Let $P = \langle 2, 0, 1\rangle$, a position vector for a particular solution to the equation. Let $N = \langle 1, -2, 3\rangle$, obtained from the coefficients in the equation.



Then $P \cdot N = 5$. So $(Q - P) \cdot N = 0$ *exactly* when Q points at a solution to the equation. To reiterate, (X, Y, Z) is a solution if and only if $Q-P$ is perpendicular to N.

The collection of all vectors in standard position perpendicular to a specified N captures the essence of "flatness" in space. We have just shown that the collection of all $Q - P$ is just this kind of surface.

So the collection of all $Q - P + P = Q$, which has exactly the same shape but is shifted by P, is "flat" too.

The vector N determines the angle of the plane and plays the same role as the slope for lines from beginning algebra.

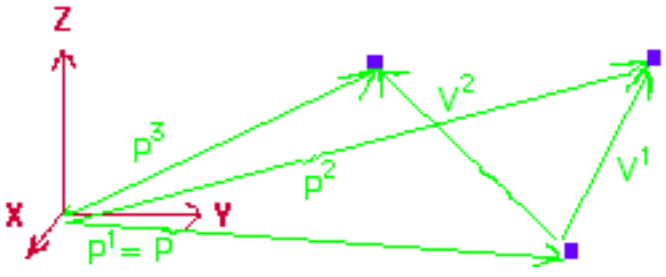
Generalizing, if $AX + BY + CZ = D$ is a first degree equation in three variables and P is a vector that "points at" a particular solution (that is to say, $Ap_1 + Bp_2 + Cp_3 = D$) then an equation for a generic point Q that "points at" a solution is

$$(Q - P) \cdot N = 0 \qquad \textbf{The Normal Form for a Plane}$$

where $N = \langle A, B, C\rangle$ is normal to the surface consisting of all solutions. This formula is analogous to the point-slope formula for lines from beginning algebra.

Here is a table to guide you through the basic possibilities to generate equations for planes in $3D$:

If you know . . .	**The vector formula is . . .**
The formula for your plane is $n_1X + n_2Y + n_3Z = D$.	Substitute numbers for two of the variables (such as 0) and work out the third, to generate a particular solution P. Let $N = \langle n_1, n_2, n_3 \rangle$. Let $Q = \langle X, Y, Z \rangle$ point at a generic spot in space. Q is a position vector for a point in the plane if and only if $(Q - P) \cdot N = 0$.
You have a particular point (p_1, p_2, p_3) that is on a plane. You also have two nonzero vectors V^1 and V^2 which lie in this plane. These two vectors cannot be multiples of each other.	Let $P = \langle p_1, p_2, p_3 \rangle$. Let N be the cross product of V^1 and V^2. N cannot be 0: the conditions on V^1 and V^2 guarantee and are implied by this. Q is a position vector for a point in the plane if and only if $(Q - P) \cdot N = 0$.
Position vectors P^1, P^2 and P^3 which point at different **noncollinear** points on your plane.	Let $V^1 = P^2 - P^1$ and $V^2 = P^3 - P^1$. These vectors must not be multiples of each other: noncollinearity guarantees and is implied by this. Let P be one of the three original vectors given. Let $N = V^1 \times V^2$. Q is a position vector for a point in the plane if and only if $(Q - P) \cdot N = 0$.



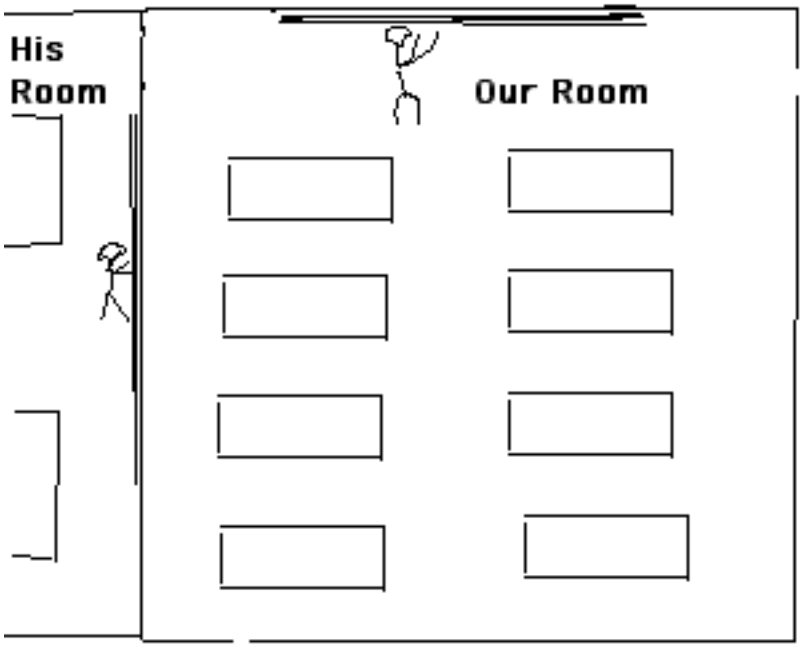
14. Parametric Planes and Translating Among Coordinate Systems

A Parable About Coordinate Systems

We are all in a classroom and I have just drawn a coordinate system on the front board with the X axis sticking out horizontally into the room and Y axis to the right and Z axis up from an origin at about waist level in the middle of the board. I am measuring units in feet along these axes.

I receive a cell phone call disrupting the class because I forgot to turn off my phone. In great embarrassment, I reach for the phone to turn it off but recognize the phone number from the phone calling me, displayed on the caller ID screen of my phone.

It is from my friend who I know is also teaching at the exact same time. Not only that, he is teaching in the class next door.



Unable to contain my curiosity, I violate all the rules of common sense and etiquette and answer the phone.

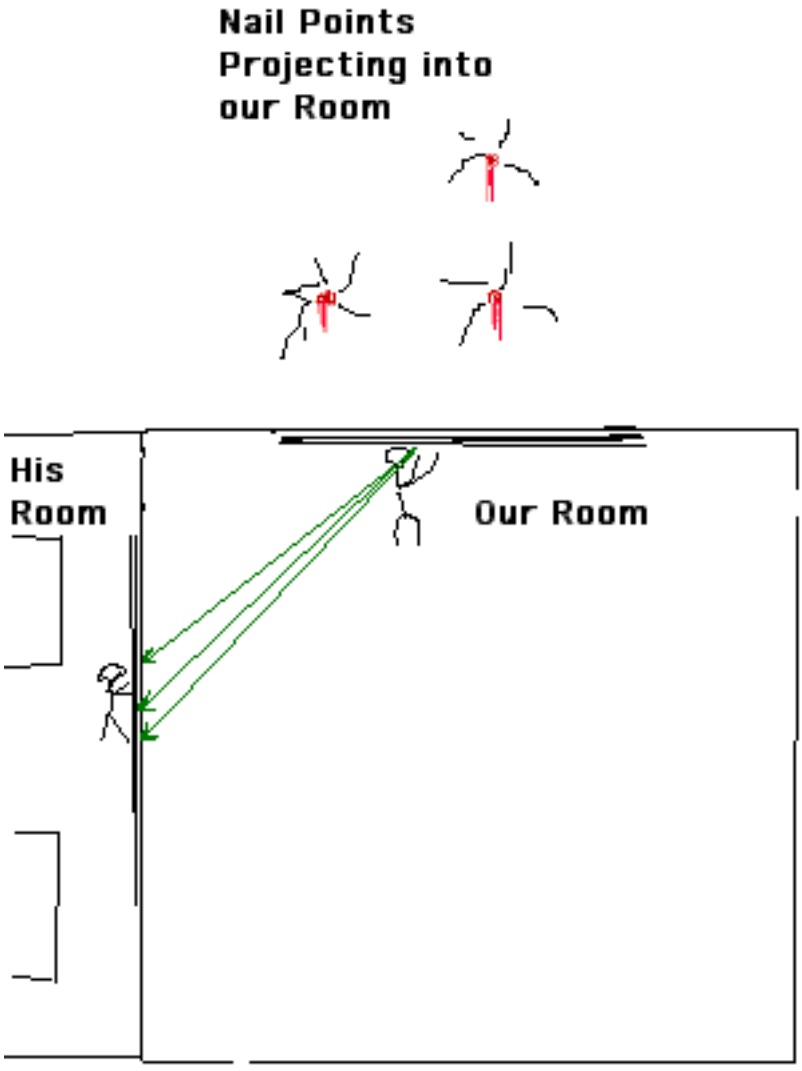
He tells me that he has just been drawing a two dimensional graph next door in his class. He tells me with enormous enthusiasm about this marvelous relationship he and his students have discovered and I scribble down his instructions.

"Finally," he says, "the sun came through the window and illuminated a certain corner of my picture and well, you just won't believe it! YOU HAVE TO SEE THIS! Gotta go!"

I am intensely curious about his picture, but from things he said it is clear that parts of the picture have to be oriented properly with respect to the outside world. I realize he has not told me enough to recreate what he has done for my class.

"Wait!" I cry, " Wait!! Where is your coordinate system! How can I reproduce your marvelous construction from the coordinates you have given me! These are not my coordinates, which are centered in the front of my classroom! How can I translate all the instructions you have told me involving your coordinate system?"

"I have NO TIME," he replies hastily. "You will just have to muddle along somehow by yourself. I see, however, a box of nails left by some workmen right here by the board."



At that moment I hear furious pounding coming from our common wall. The points of three nails stick out through the wall and he returns to the phone.

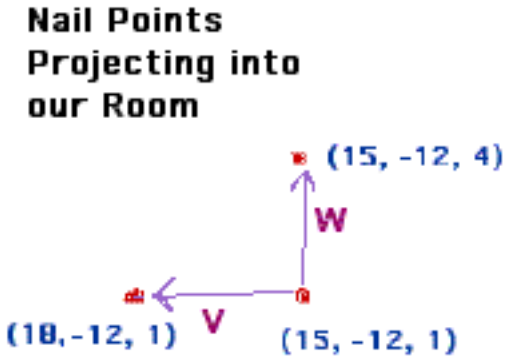
"There" he picks up the phone "I have to go RIGHT NOW. But these nails represent my origin and segments, one of my units long, along each of my two axes. He then hangs up."

"Great Scott," I shout! "My friend has given me just enough information to create a very efficient translator for all his intricate directions!"

Carefully, I measure the coordinates of each nail hole (the wall, evidently, is very thin.)

"AHA!" I reason. "His origin is very close to the place I call $(15, -12, 1)$. He has set his origin about a foot higher on the wall than I did."

Next I notice that his single units are each 3 of my units long. "I see, he must have been using a yardstick to measure distances, not a footstick," I opine. "Whenever he refers to a displacement of one unit in the direction of his X axis, he was talking about multiples of the vector $V = \langle 3, 0, 0\rangle$! And whenever he refers in his instructions to a displacement of one unit in the direction of his Y axis, he was talking about multiples of the vector $W = \langle 0, 0, 3\rangle$!"



"Since I can point to his origin using the vector $P = \langle 15, -12, 1\rangle$ I can point at a place he calls (s, t) with my vector $P + sV + tW$. This is how I translate from his coordinates to mine."

$$(s, t) \longleftrightarrow Q = P + sV + tW$$

This looks a lot like the parametric vector equation for a line! $Q = P+sV+tW$ is called the **parametric vector equation for a plane**, in this case his XY plane. s and t are the parameters. In this case they are the numbers he calls the X and Y coordinates. Now I know how to follow his directions! If he says "Draw a line between $(2, 6)$ and $(-1, 7)$" I should draw the line between the tips of my standard position vectors $P + 2V + 6W = \langle 21, -12, 19\rangle$ and $P - V + 7W = \langle 13, -12, 22\rangle$.

And that, dear readers, is the end of this story. The marvelous invention of my friend, and what it means for humankind, is for another time and place. All I can say about that is that I always, **yes ALWAYS**, from that day to this, remember to turn off my cell phone before class starts.

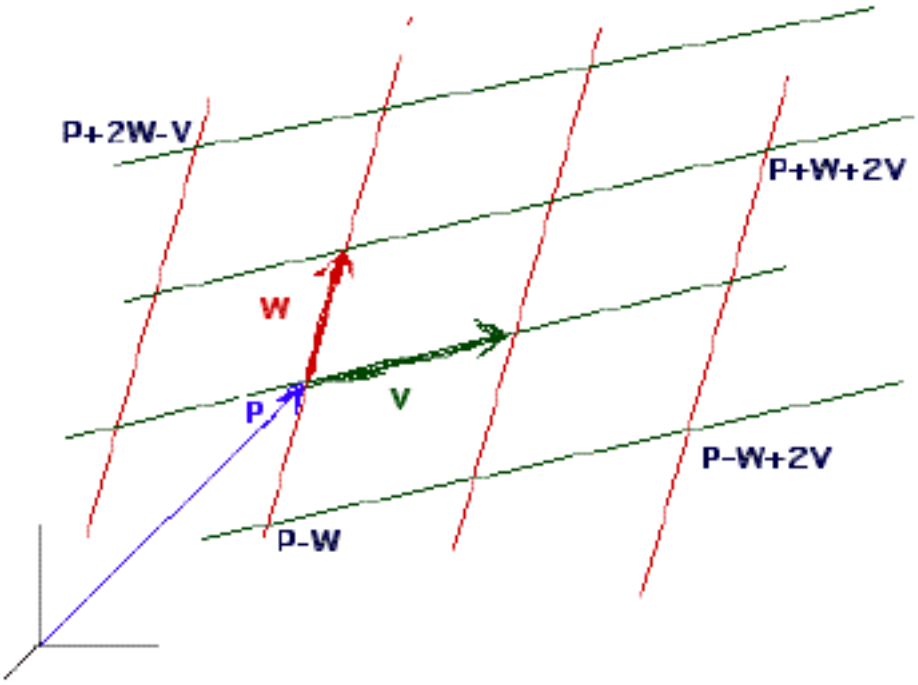
Leaving, sadly, our parable aside, (that was fun!) it is clear that there was nothing special about P, V and W. If we have any plane, we can find a point P and two noncollinear vectors V and W that lie in the plane.

We use P to get out to a spot on the plane and then add multiples of V and W to move around on the plane once we are there. They give us a coordinate grid, similar to the XY coordinate grid, but possibly bent a bit. The parameters s and t are coordinates out on the plane. They refer to multiples of V and multiples of W, respectively. They represent displacements from some generic center point, rather than similar displacements by multiples of $\vec{i}$ and $\vec{j}$ from the origin.

To actually obtain the coordinates for a generic point $Q = (X, Y, Z)$ on the plane, examine the scalar projections of $Q - P$ onto the V and the W directions.

$$s = \frac{(Q-P)\cdot V}{|V|} \qquad \text{and} \qquad t = \frac{(Q-P)\cdot W}{|W|}.$$

If you lived on the plane and were unaware of the third dimension, s and t (and your initial origin choice and axis directions V and W in your plane) would be all the coordinates you would need.



Taking this a step further, suppose the directions from my friend for the construction in our parable had included three dimensions, not just the two on his board.

What extra information would we need and how should we use it to form a "3D translator?"

With the work we have done already it is simple. We need to know the vector in our coordinates that corresponds to his third direction. In our example, that might be $\langle 0, -3, 0\rangle$, three feet straight out of his blackboard into his room. He might call this his Z axis. Let us call that vector U. When he refers to a position

(s,t,r) we want to be able to locate the place he refers to in OUR coordinates. If the number r refers to his idea of distance in the direction we think of as the vector U, and s and t refer to multiples of V and W respectively, then our translator is

$$(s,t,r) \longleftrightarrow Q = P + sV + tW + rU$$

We have done two very important things in this section. First, we have learned how to represent planes by using two parameters, just like the one parameter representation of lines. Second, we have learned to **translate among coordinate systems** with different origins and different "unit" coordinate vectors.

14.1. ***Exercise***. *With the U suggested above, find the standard position vector in our coordinates that points to a place my friend thinks of as* $(-3,6,2)$.

14.2. ***Exercise***. * *With the same information as above, create a translator to be used by my friend in the other room: that is, one that takes OUR coordinates and converts them to vectors in HIS world.*

14.3. ***Exercise***. * *We suppose* $Q(t) = P + sV + tW$ *with* $W \neq 0 \neq V$ *and with* V *not a multiple of* W. *So* Q *is a parametric vector equation of a plane. Let*

$$\widetilde{V} = V - \frac{V\cdot W}{W\cdot W}\,W \text{ and } \overline{V} = \frac{\widetilde{V}}{|\widetilde{V}|} \text{ and } \overline{W} = \frac{W}{|W|} \text{ and } \overline{P} = P - \overline{V}\cdot P\;\overline{V} - \overline{W}\cdot P\;\overline{W}.$$

Now define $\overline{Q}(a,b) = \overline{P} + a\overline{V} + b\overline{W}$.

(i) Show that

$$\overline{V}\cdot\overline{V} = \overline{W}\cdot\overline{W} = 1 \text{ and } \overline{P}\cdot\overline{V} = \overline{P}\cdot\overline{W} = \overline{W}\cdot\overline{V} = 0.$$

(ii) Show that the planes parameterized by Q *and* $\overline{Q}$ *both contain the three points* P *and* $P+V$ *and* $P+W$, *so* Q *and* $\overline{Q}$ *parameterize the same plane.*

(iii) The point on this plane nearest to the origin corresponds to a and b at which $\overline{Q}(a,b)\cdot\overline{Q}(a,b) = \overline{P}\cdot\overline{P} + a^2 + b^2$ *is as small as possible. This obviously happens when* $a = b = 0$. *So the near spot has position vector* $\overline{P} = P - \overline{V}\cdot P\;\overline{V} - \overline{W}\cdot P\;\overline{W}$.

Show that this is the one and only position vector of a point on the plane which is perpendicular to the plane.

(iv) Show that the position vector of the point on the plane closest to the point with position vector A *has position vector*

$$B = P - \overline{V}\cdot(P-A)\;\overline{V} - \overline{W}\cdot(P-A)\;\overline{W}.$$

B *is the one and only position vector of a point on the plane for which* $B - A$ *is perpendicular to the plane.*

---

15. Vector Functions

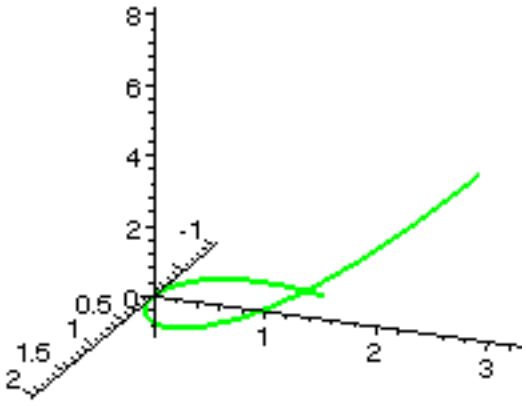
In this and later sections we will talk about **vector valued functions.** You have already seen some of these. The parametric vector equations we worked with in the first chapter are an example, but all of those represented straight line and constant speed motion. The ones we consider here can exhibit more interesting behavior.

We will presume that $Q(t)$ is a vector for each t in the interval (a, b). That is what is meant when we say that Q is a "vector valued function" or "vector function." This vocabulary distinguishes these functions from the **real functions** you have been working with since early algebra. In these notes $Q(t)$ will be in the plane for all t or in space for all t, although much of what we do is independent of the dimension.

A **curve** or **path** is the set of points traced out by the tip of a vector function such as Q thought of as a position vector as it runs through a subinterval of its parameter interval. A curve is a purely geometrical object. The same curve might come from various vector functions, but when you refer to a curve the implication is that there **is** at least one Q around somewhere to trace it out. When we considered constant velocity parametric motion we used the phrase "geometrical track" rather than "curve" or "path."

$$\text{So} \quad Q(t) = \langle X(t), Y(t) \rangle \quad \text{or} \quad Q(t) = \langle X(t), Y(t), Z(t) \rangle \, .$$

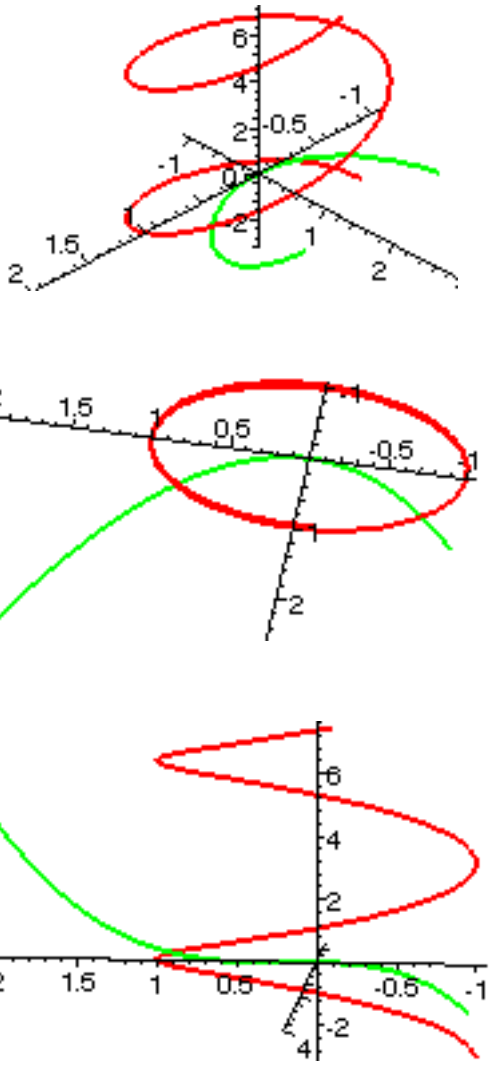
When it is convenient we will suppress t and write $Q = \langle X, Y \rangle$ or go back to subscripting the entries of Q and write $Q = \langle q_1, q_2, q_3 \rangle$.



Computer algebra systems, with their ability to draw graphs and rotate them to show the curve from various viewpoints are definitely the way to go when you wish to draw a picture of a curve, particularly in $3D$.

Here is a picture[9] of the curve $Q(t) = \langle t, t^2, t^3 \rangle$ on the interval $[-1, 2]$ from a typical perspective.

Here are three pictures[10] from different perspectives of the same curve plotted together with the **helix** $H(t) = \langle cos(t), sin(t), t \rangle$ with the parameter interval $[-3, 8]$. A helix is a regular spiral shape, in this case winding around the Z axis.

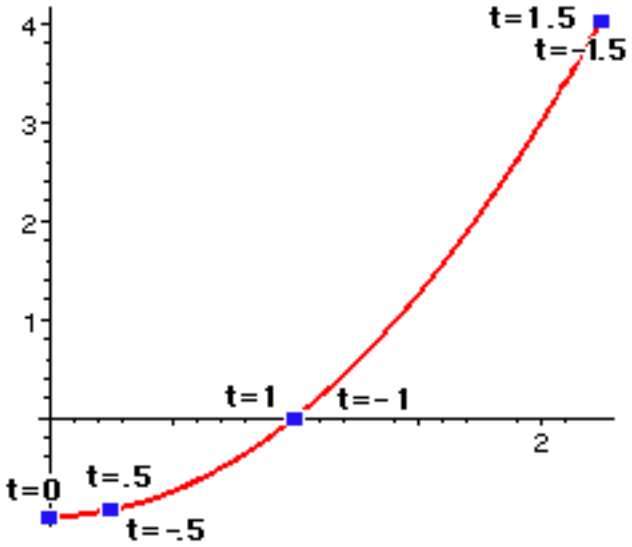


Any single one of the pictures you see above could correspond to many different curves. Even taken together, there is no way to guess from these pictures **when** a parameterization is at a place on the curve. The sole purpose for spending the time to draw any picture is so you can harness your visual intuition to better understand relationships among changing quantities. You have an astoundingly powerful processor in your head for this kind of thing, but it needs the right kind of information to work for you. There are a couple of ways to make pictures of this kind more useful for visualizing the motion, at least for parameterizations that don't behave too badly.

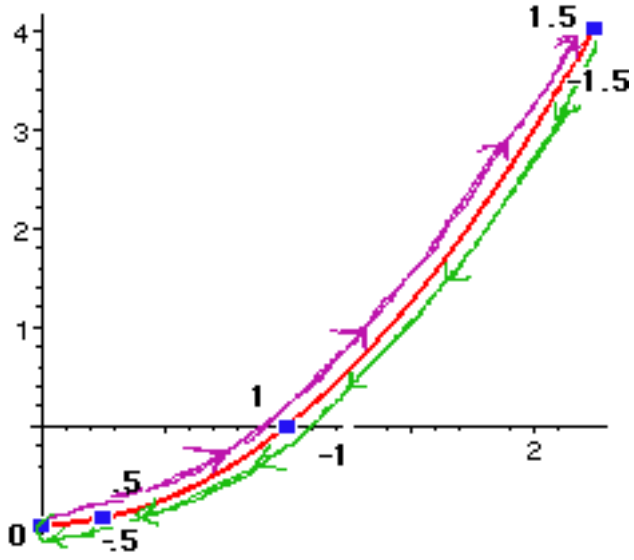
Let's consider the curve parameterized by $Q(t) = \langle\ t^2,\ t^4 - 1\ \rangle$ on the interval $-1.5 \le t \le 1.5$. All the points on the curve are on the parabola $Y = X^2 - 1$ (check this!) but knowing this does not help us much in understanding the motion.

Find below a table of X and Y values corresponding to regularly spaced times.

t	X	Y
-1.5	1.25	≈ 4.1
-1	1	0
-.5	.25	$\approx -.9$
0	0	-1
.5	.25	$\approx -.9$
1	1	0
1.5	1.25	≈ 4.1



In the picture to the right of the table you will see a piece of the parabola with points labelled with the times when the parameterization was at that point. At time -1.5 we are at the upper right. Half of a second later we have moved to $(1, 0)$. Half a second after that we are in the vicinity of $(.25, -.9)$. We seem to be slowing down because the displacement in this half second appears to be shorter. During the next half second this seems to be even more pronounced, a definite "slowdown" of the motion. After $t = 0$ the motion reverses itself and retraces its steps, speeding up as it moves up and right. Below you will find a picture which shows the features of this discussion.



When you draw a picture like that you should try to do the following: First, plot and label points at **evenly spaced times** so you can compare apparent speeds.

Second, use arrows attached to the curve to make plain the direction of the movement. If the curve retraces part of itself try to show that important feature by "doubling back" as was done in the picture. From a picture like that you can often estimate the speed, estimate the direction of the movement and get a sense of how the parameterization moves across the curve over time.

15.1. ***Exercise***. *Plot enough points so you understand the movement along the curves given by parameterizations*

$$Q(t) = \left\langle\ t^2,\ cos\left(\frac{t}{4}\right)\ \right\rangle \quad \textit{and} \quad P(t) = \langle\ 1-2t,\ e^t\ \rangle$$

on the interval $0 \le t \le 2$. *You should try to use the ideas from the last paragraphs to make your graph more useful.*

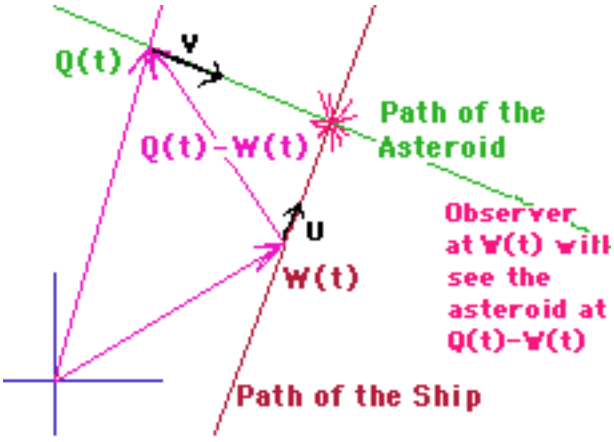
Relative Motion

Suppose we have two constant velocity objects moving with **synchronized** times (same units of time, same time zero) whose positions are given by

$$Q(t) = A + tV \quad \text{and} \quad W(t) = B + tU$$

Even if the two curves cross they need not **collide**. They must be at the crossing place **at the same time.** This will happen when (and if) $Q(t) - W(t)$ is ever zero. For visualization purposes, let's imagine that $W(t)$ represents the position of a spaceship, while $Q(t)$ is the position of an asteroid.

$Q(t) - W(t)$ represents the **relative position** of the asteroid as seen from the observer on the ship.



After 1 second the asteroid has moved by one copy of its velocity vector V while the ship has moved by one copy of U so the **relative velocity**, the change in relative position in one second, as witnessed by the observer on the ship, is $V - U$.

Setting $Q(t) - W(t) = 0$ we find that $A - B = -t(V - U)$ is the condition for a collision. If this is to happen after time 0, we must have $A - B$ a negative multiple of the relative velocity. This means that at time 0 the observer looked up and saw the asteroid coming straight at the ship.

15.2. ***Exercise***. *Show that if the observer looks up and sees the asteroid coming straight at the ship for* ***any*** *time, not just* $t = 0$*, then a collision will occur. On the other hand, if the observer looks up and* ***ever*** *sees the asteroid* ***not*** *pointing straight at the ship, then a collision will not occur. (These facts are due to the "constant velocity" nature of the motion.)*

If there is a collision, any damage to the ship will be caused by a large relative velocity, not the velocities of ship or asteroid alone. The **relative speed** of the impact, $|V - U|$ is what generates the damage. The more "head on" the collision, the more effective the velocities V and U will be in causing damage, while least damage will occur (for given individual speeds of ship and asteroid) if U and V point in the same direction. Anyone who has played on "bumper cars" at a fair learns this idea early.

The **relative angle** between the motion of ship and asteroid is defined to be the angle between their velocity vectors as perceived by an observer stationary with respect to an agreed-upon coordinate system. In our case the angle θ is defined by $V \cdot U = |V||U|cos(\theta)$. This is the angle of the collision as seen by a stationary witness looking down on the plane of the solar system and watching this action.

But what if the witness does not agree with our idea of "stationary?" We hypothesize the existence of an **Alien from Arcturus** passing above the plane of the solar system with constant velocity and sending a radio signal back home describing what "it" sees.

If we see the alien as moving with constant velocity S then to it our whole measurement structure is moving with velocity $-S$. When we see our asteroid displaced by V in one second it sees displacement $V - S$. When our ship moves by a vector U over one second, it sees a movement of $U - S$.

To the alien the relative angle of the motion of ship and asteroid is the angle between $V - S$ and $U - S$ which can easily be different from the angle between V and U.

However the alien will still see the relative velocity, the item that is important in calculating damage potential, to be $(V - S) - (U - S) = V - U$ just as before.

15.3. ***Exercise***. *Show that if* $U \neq V$ *and our alien from Arcturus is passing by with velocity* $\frac{U+V}{2}$ *(to us) then it will perceive the asteroid and ship to be moving in opposite directions with equal speed.*

If V *is not a multiple of* U *and* $S = \frac{U \cdot V}{V \cdot V} V$ *what is the relative angle of the motions of ship and asteroid as perceived by the alien?*

15.4. ***Exercise***. *Consider the three constant velocity parametric motions with times synchronized. Which pairs would collide? For those pairs that do, when do they collide? What is the impact speed? With what velocity should you move so that, to you, the impacts appear to be "head on" and the speeds of the two objects seem to be equal?*

$$\langle 2t \;\; 5t+5\rangle \quad \textit{and} \quad \langle 13-t \;\; -t+29\rangle \quad \textit{and} \quad \langle t+3 \;\; 2t+14\rangle\,.$$

Superposition of Linear and Circular Motion

Interesting $2D$ pictures are provided by combinations of linear and circular motion, and using vector functions you can handle them intuitively and with a minimum of fuss.

A vector function $\langle cos(t), sin(t)\rangle$ will trace out the unit circle in a counterclockwise direction when followed over a time interval of length 2π. So if ω is a number, $\langle cos(2\pi\omega t), sin(2\pi\omega t)\rangle$ will trace out the circle $|\omega|$ times when followed over one time unit. $|\omega|$ is called the **frequency** of the circular motion. If ω is positive, the motion will be counterclockwise. If it is negative the motion will be clockwise. If R is a number, the behavior of $R\,\langle cos(2\pi\omega t), sin(2\pi\omega t)\rangle$ is similar to this, except the circle will have radius $|R|$. Finally, if t_0 is a number, $R\,\langle cos(\,2\pi\omega(t-t_0)\,), sin(\,2\pi\omega(t-t_0)\,)\rangle$ is "behind" the last motion by time t_0 or, if you prefer, by phase angle $2\pi\omega t_0$.

This is pretty straightforward stuff, but what happens if you add two such motions together? From a vector standpoint you can think of it as an arrow moving in a circle with another arrow attached to its tip, swinging around in its own circle, somewhat like the robot arm from Section 9. If you were to see only the resultant motion of the second tip and not the vector parts the situation might look complicated and mysterious indeed. Throw in a constant velocity motion on top of the circular ones and it makes the motion even harder to recognize.

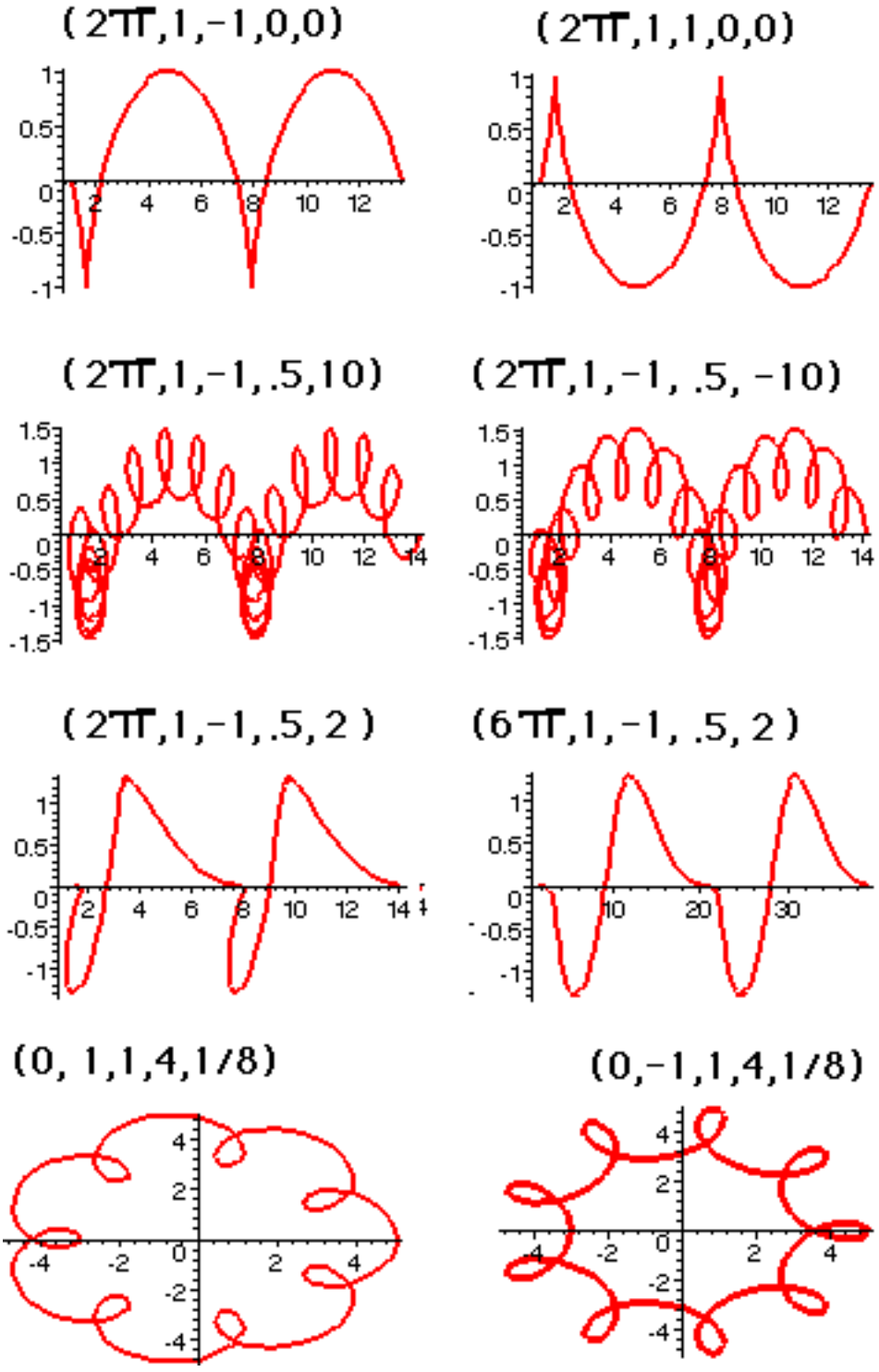
Here is a formula to create a parametric motion $Q(t)$ using, of course, a computer graphing utility.

$$\langle st, 0\rangle + R_1\,\langle cos(2\pi\omega_1 t), sin(2\pi\omega_1 t)\rangle + R_2\,\langle cos(\,2\pi\omega_2(t-t_0)\,), sin(\,2\pi\omega_2(t-t_0)\,)\rangle\,.$$

The number s is the speed of the moving "center" of the whirling assemblage of vectors, which is moving to the right for a positive s and at $\langle st, 0\rangle$ for each t. Added to that linear motion is a circular motion given by an arrow $R_1\,\langle cos(2\pi\omega_1 t), sin(2\pi\omega_1 t)\rangle$ extending from the moving center. Finally we add on the circular motion $R_2\,\langle cos(\,2\pi\omega(t-t_0)\,), sin(\,2\pi\omega(t-t_0)\,)\rangle$ with its own radius and frequency and, possibly, time shifted relative to the first.

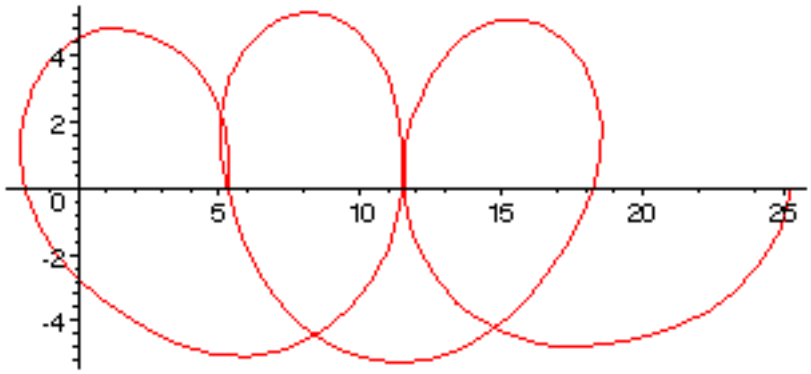
Find below a few pictures[11] created by using various combinations of these constants. These are called **cycloids** of various types.

On each picture you will find the list of constants: (s, R_1, ω_1, R_2, ω_2). Since we only look at a few examples we have chosen $t_0 = 0$ in each case. A more thorough exploration of these pictures would consider the case where the two component circular motions "start" out of phase.



An Example with Data

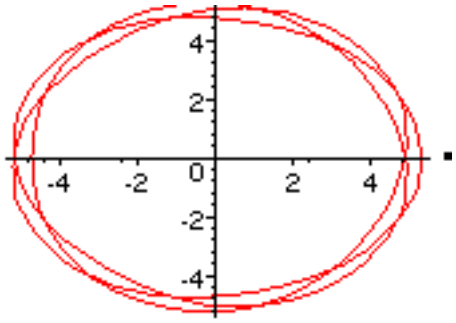
Let's look at the situation now from a different perspective. Consider the picture below.



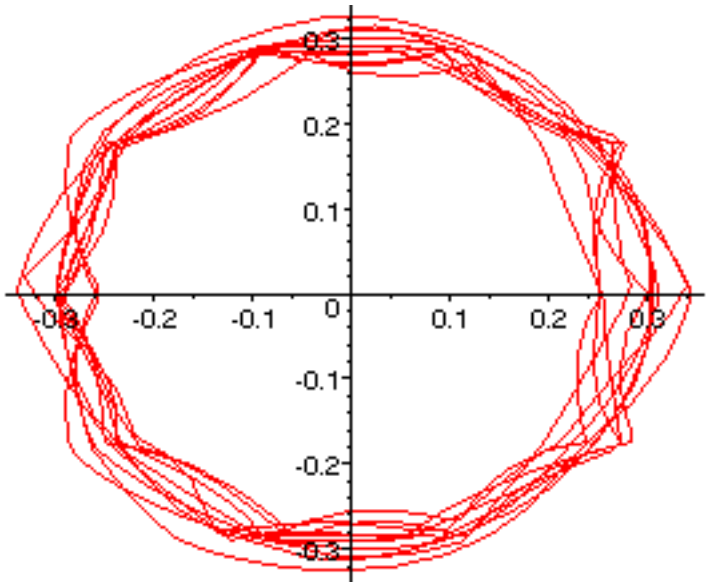
This graphic is coming to you not as the graph of any function but as a collection of data points: you may think of them as observations which you and a team of colleagues have collected of an object moving across the sky over time. Both time and position coordinates are included with the data in an endnote.[12] Observations were taken every few days for one year. Graphics utilities can be told to plot a large list of consecutive data points like that and connect them. The process of bridging the gap between known values is called **interpolation**, and when you bridge them with line segments it is called **linear interpolation**. That is what you see above.

Based on philosophical considerations or experience with cycloid-like graphs or introspection you and your colleagues have a theory: namely that you are witnessing a combination of linear and circular motions of the object.

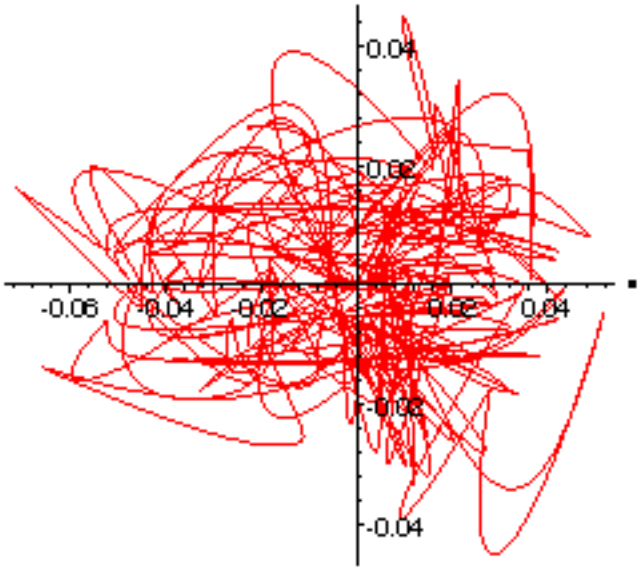
Looking at the points, it certainly seems to be some kind of wiggly repetitive movement trending, generally, to the right. Looking at the starting point $\langle 5, 0\rangle$ and ending point one time unit later at $\langle 25, 0\rangle$, the whole business seems to be moving right at 20 distance units per year. You should subtract $\langle 20t, 0\rangle$ from the observation function, graph[13] the difference, and see if the result is more recognizable. Here is what you get.



Aha! This looks a bit like three slightly distorted circles of radius 5. Once again you subtract, but this time you should subtract something close to $\pm 5\,\langle cos(\pm 6\pi t), sin(\pm 6\pi t)\rangle$. The circle $5\,\langle cos(6\pi t), sin(6\pi t)\rangle$ does the trick, and when you subtract it from the modified data graph[14] this is what you get:



The distortion is now roughly circular, of radius .3. If you look closely, you can count ten little circles. We are led to subtract $.3\,\langle cos(20\pi t), sin(20\pi t)\rangle$ as well as the first circle and the linear motion from the data graph[15] and end up with the following.



The formula for the motion is the sum of everything we subtracted off plus this little wiggly function.

Are these wiggles caused by inaccuracies in our ability to measure? Are they caused by non-optimal choices for the radii and frequencies of the subtracted circular motions, or the speed of the linear motion? Are they artifacts of the linear interpolation we used? Real things, after all, don't usually move on polygons.

Could it be that there are problems with the theory that we were witnessing circular motions superimposed on a linear motion? The task at this point is to address the data handling issues one after another. Then we identify the little deviations as within what you would expect from errors generated by the data gathering and handling scheme—or not. If they are, you are happy because your theory has been supported by this data. If not, you would report this and go looking for a better theory. Either way you write a paper and you and your coauthors all go out to dinner!

CHAPTER III

Vector Calculus of Curves May 27, 2005

16. Limits and Continuity for Vector Functions

In this and later sections we will extend to vector functions some of the facts about continuity, derivatives and integrals you probably learned in first year Calculus[16] for ordinary real functions.

We introduce a new notation for the real line, the plane and space. The real line will be denoted $\mathbb{R}$ or $\mathbb{R}^1$. Our standard representation of "the plane" will be denoted $\mathbb{R}^2$ while $\mathbb{R}^3$ denotes our representation of "the points in space."

We will suppose Q is a vector valued function on an interval (a, b). If B is a vector in the same world as Q we write, for c in the interval (a, b),

$$\lim_{t\to c} Q(t) = B \text{ precisely when } \lim_{t\to c} |Q(t) - B| = 0.$$

We say that the **limit** as t goes to c exists and equals B in that case. The notation $\lim_{t\to c} Q(t) = B$ is used both to assert the existence of the limit and to name the limit as the vector B.

16.1. ***Exercise***. *There are conditions that are equivalent to the existence of the limit as stated, and often easier to use. For example:*

(i) $\lim_{t\to c} q_i(t) = b_i$ *for all the coordinate functions* $q_i(t)$.

(ii) $\lim_{t\to c} (Q(t) - B) \cdot (Q(t) - B) = 0$.

Satisfy yourself that each of these conditions are equivalent to $\lim_{t\to c} Q(t) = B$.

16.2. ***Exercise***.

Show that if N *is a vector and* $\lim_{t\to c} Q(t) = B$ *then* $\lim_{t\to c} Q(t) \cdot N = B \cdot N$.

On the other hand, if $\lim_{t\to c} Q(t) \cdot N$ *exists for EVERY vector* N *then* $\lim_{t\to c} Q(t)$ *exists. (hint: What happens if* N *is a coordinate vector?)*

Q is called **continuous at** c when $\lim_{t\to c} Q(t)$ exists and equals $Q(c)$.

We say that Q is **continuous on a specific subinterval** of (a, b) if it is continuous at every point in the subinterval. If we assert, for example, that Q is continuous on $[c, d]$ we are saying that every point of $[c, d]$ is in (a, b) and that Q is continuous at every point in $[c, d]$.

We will say that the number or vector A is a **good approximation** to a number or vector B if $|A - B|$ is small. We specifically do not require $|A - B|$ to be 0 for a good approximation, only small. "Good" in this context depends entirely on what kinds of differences you care about. "Small" might be very small indeed. However if Q is continuous[17] at c then no matter how small your concept of "good" requires $|Q(t) - Q(c)|$ to be, you can guarantee that $Q(t)$ will be a good

approximation to $Q(c)$ merely by requiring t to be a "good enough" approximation to c.

---

17. Derivatives of Vector Functions

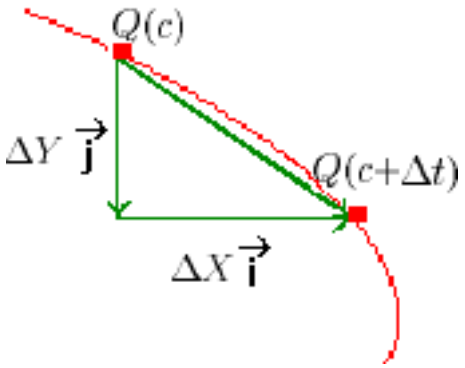
Q is called **differentiable** at c if and only if

$$\lim_{\Delta t \to 0} \frac{Q(c+\Delta t) - Q(c)}{\Delta t} \text{ exists.}$$

When the limit does exist, it is commonly denoted $\mathbf{Q'(c)}$. If specificity is needed, the notation $\mathbf{\frac{dQ}{dt}(c)}$ or $\mathbf{\frac{d}{dt}Q(c)}$ can be used. This limit is called the **derivative** of Q at c and the process of finding a derivative is called **differentiating** or **taking the derivative**.

Note that when $Q'(c)$ exists and if we use the handy notation ΔQ for $Q(c+\Delta t) - Q(c)$ we have

$$0 = \lim_{\Delta t \to 0} \left| \frac{Q(c+\Delta t) - Q(c)}{\Delta t} - Q'(c) \right|$$
$$= \lim_{\Delta t \to 0} \left| \frac{\Delta Q}{\Delta t} - Q'(c) \right| = \lim_{\Delta t \to 0} \left| \frac{\Delta Q - Q'(c)\Delta t}{\Delta t} \right|.$$



We can derive two pertinent facts from this equation. First, $\frac{\Delta Q}{\Delta t}$ will be a **good approximation** to $Q'(c)$ provided Δt is small enough. Second, $|\Delta Q - Q'(c)\Delta t|$ is small **even in comparison to** Δt when Δt is small.

A function is said to be **differentiable on a specific subinterval of** (a, b) if it is differentiable at every point in the subinterval.

The function defined on a subinterval of (a, b) by differentiating Q might itself be continuous or differentiable, which leaves open the possibility of higher derivatives. Q'', Q''' and so on are defined, when they exist, in the obvious way. This habit of adding on primes to indicate more derivatives has limits—in general $Q^{(n)}$ or $\frac{d^n Q}{dt^n}$ will denote the n–th derivative for a positive integer n.

If we assert that a function is differentiable on a closed interval $[c, d]$ we are implicitly assuming that the function is (or *could be*) defined on a larger open interval containing $[c, d]$ and that the function is differentiable on this larger open interval.

Q' is often called the **velocity** of Q, its magnitude $|Q'|$ is called the **speed**, and Q'' is called the **acceleration**. This vocabulary is most commonly used when the parameter represents time and the components of Q are all distances.

17.1. ***Exercise***. *Show that Q' exists exactly when all the coordinate functions for Q have derivatives. When they do, $Q' = \langle X', Y', \dots \rangle$. (The dots imply that you keep going till you run out of coordinates.)*

17.2. ***Exercise***. *Show that the following familiar results are still true. We assume that f and g are ordinary functions and Q and H are vector functions and all are differentiable and $g \neq 0$. For the last result we presume that t itself is a differentiable function of u. We let r and s be constants and let N be a constant vector. This exercise says that everything obvious that you would like to do with vectors is OK. (hint: All of these are done using components and assuming basic Calculus facts.)*

*(i) (**Constant Multiple Rule**)* $(N \cdot Q)' = N \cdot Q'$.

*(ii) (**Sum and Constant Multiple Rule**)* $(rQ + sH)' = r(Q') + s(H')$.

*(iii) (**The Product Rule**)* $(f\ Q)' = f'Q + f(Q')$.

*(iv) (**The Product Rule**)* $(Q \cdot H)' = (Q') \cdot H + Q \cdot (H')$.

*(v) (**The Product Rule**)* $(Q \times H)' = (Q') \times H + Q \times (H')$.

*(vi) (**The Quotient Rule**)* $\left(\dfrac{Q}{g}\right)' = \dfrac{g(Q') - g'Q}{g^2}$.

*(vii) (**The Chain Rule**: Precise Formulation)* $(Q \circ t)'(u) = Q'(t(u))t'(u)$.

*(viii) (**The Chain Rule**: Common Alternative Notation)* $\dfrac{dQ}{du} = \dfrac{dQ}{dt}\dfrac{dt}{du}$.

The equivalent of the **Mean Value Theorem**[18] does not hold for vector functions, though it does for each of the coordinate functions. Specifically, if Q is continuous on the closed interval $[r, s]$ and differentiable on (r, s) there need not exist $c \in (r, s)$ for which $Q'(c) = \dfrac{Q(s) - Q(r)}{s - r}$.

A counterexample is provided by the helix we saw earlier. If $s = r + 2\pi$ then $\dfrac{Q(s) - Q(r)}{s - r}$ is vertical (parallel to the Z axis) but $Q'(t)$ is never vertical.

Even when there is a c for which $Q'(c)$ points in the right direction, the speed could be wrong. The problem is that the values of c which "work" for the coordinate functions individually need not match.

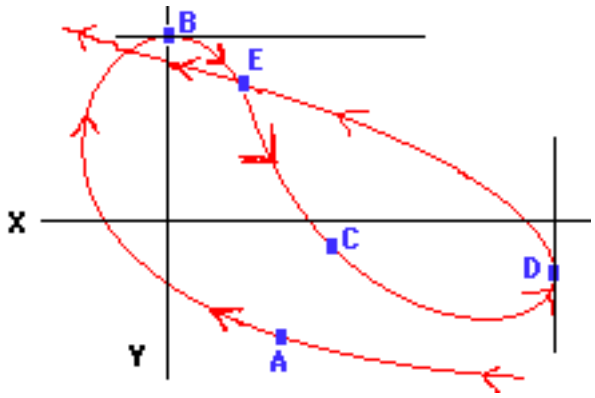
However, if N is any constant vector, then $N \cdot Q$ is a real function so there will be a $c_N \in (r, s)$ for which $N \cdot Q'(c_N) = N \cdot \dfrac{Q(s) - Q(r)}{s - r}$. Sometimes that is enough.

Here is an example. Suppose $Q' = 0$ on an interval. Then we can conclude that Q is constant on that interval: otherwise, one of the coordinates of Q would be different at two spots r and s on the interval. Suppose it is the X coordinate that is not constant and $X(r) \neq X(s)$. But then there must be c between r and s with $X'(c) = \dfrac{X(s) - X(r)}{s - r} \neq 0$ and so $Q'(c)$ is not the zero vector, contrary to our assumption. So $Q' = 0$ on an interval requires that all the coordinate functions are constant and so Q itself is constant. The N from the previous paragraph would, in our example, be $\vec{i}$.

Here is another example, that is actually pretty important for us. Recall that any function f is called **one-to-one** if the only way that $f(a) = f(b)$ is if $a = b$. Suppose a vector function Q is continuously differentiable and $Q'(c) \neq 0$. Then Q is one-to-one on some interval $[r, s]$ with c in (r, s). Justification: Continuity of Q' implies that at least one of coordinate functions of Q, say X, has derivative with constant sign on an interval $[r, s]$ containing c. So the mean value theorem applied to X implies that X is either strictly increasing or strictly decreasing on that interval: X is one-to-one. So Q is too. Once again we are looking at $Q \cdot N$ where $N = \vec{i}$.

17.3. ***Exercise***. *Two vector functions whose derivatives exist and agree on an interval differ by a constant vector on that interval.*

17.4. ***Exercise***. ** Show that if the $n + 1$st derivative $Q^{(n+1)} = 0$ for all t on an interval then there are constant vectors C_0, $C_1, \ldots, C_n$ for which $Q(t) = t^n C_n + \cdots + tC_1 + C_0$ on the interval.*

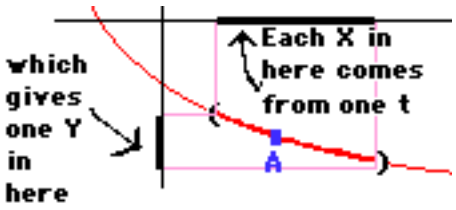


It is possible to read quite a lot of useful information about derivatives, and exercise your intuition about them at the same time, from an examination of a graph

such as the one above. We will assume that the derivative of a parameterization $Q(t) = \langle X(t), Y(t)\rangle$ for the $2D$ curve shown is continuous and never zero. The curve is traversed in the direction of the arrow.

Let's think about $\frac{dY}{dt}$, $\frac{dX}{dt}$ and $\frac{dY}{dX}$ at the time and place indicated as A. As t increases and the parameterization passes through A we are rising and moving to the left. Y is getting bigger and X is getting smaller. So $\frac{dY}{dt}$ is positive and $\frac{dX}{dt}$ is negative. If you imagine the slope of a line tangent to the curve at A it is clear that $\frac{dY}{dX}$, should be negative, consistent with the fact that $\frac{dY}{dX} = \frac{dY}{dt}/\frac{dX}{dt}$ at A by the chain rule.

The last statement requires a bit of explanation, since Y is not given explicitly as a function of X. We assume that the derivative $\frac{dX}{dt}$ is continuous and nonzero near A. So the function $X(t)$ is one-to-one[19] , at least in some small time interval $[c-\varepsilon, c+\varepsilon]$ corresponding to a little piece of the curve around $A = Q(c)$.



Also, since X is continuous, it actually takes on every value between $X(c-\varepsilon)$ and $X(c+\varepsilon)$.

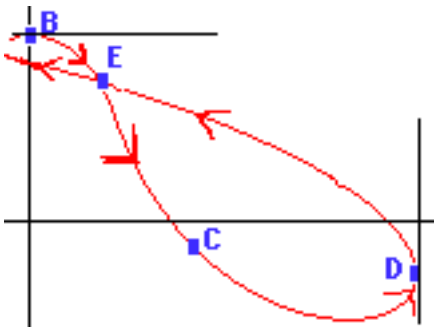
We can now make Y a function of X on the little X interval by defining $Y(X)$ to be that unique $Y(s)$ for which $X = X(s)$, where s is in $[c-\varepsilon, c+\varepsilon]$. The Y value is linked to X through the parameter.

Since $\frac{dY}{dt}$ and $\frac{dX}{dt}$ both exist at A they can be approximated by $\frac{\Delta Y}{\Delta t}$ and $\frac{\Delta X}{\Delta t}$. So if we look at a small change ΔX near A and the associated change ΔY near A we see that $\frac{\Delta Y}{\Delta X} = \frac{\Delta Y}{\Delta t}/\frac{\Delta X}{\Delta t}$ where Δt is the time change that produced both ΔY and ΔX. Since $\lim_{\Delta t\to 0} \frac{\Delta Y}{\Delta t}/\frac{\Delta X}{\Delta t} = \frac{dY}{dt}/\frac{dX}{dt}$ it must be that $\lim_{\Delta X\to 0} \frac{\Delta Y}{\Delta X}$ exists and is same number.[20]

Before going to an exercise let's go back to the big graphic and look at a different point, D. It seems that the curve is going straight up there. So $\frac{dY}{dt} > 0$, $\frac{dX}{dt} = 0$ and $\frac{dY}{dX}$ does not exist.

17.5. ***Exercise***.

Determine the signs (or 0 if that makes most sense) of $\frac{dY}{dt}$, $\frac{dX}{dt}$ *and* $\frac{dY}{dX}$ *at the points* B *and* C *in the diagram. What should you do about* E*?*



---

18. Integrals of Vector Functions

If Q is continuous on $[r, s]$ we define

$$\int_r^s Q(t)dt \quad \text{to be} \quad \left(\int_r^s X(t)dt\right)\vec{i} + \left(\int_r^s Y(t)dt\right)\vec{j} + \dots.$$

This vector is called the **integral** of Q with the specified limits.

If Q is **continuously differentiable** (that is, it is differentiable and its derivative is continuous) on $[r, s]$ then

$$Q(s) - Q(r) = \int_r^s Q'(t)dt.$$

So the usual methods of evaluating integrals work with vector integrands.

18.1. ***Exercise***. *Is it true that for any vector N and continuously differentiable Q,*

$$(Q(s) - Q(r)) \cdot N = \int_r^s Q'(t) \cdot N \; dt?$$

Suppose $r = t_0 < t_1 < \cdots < t_n = s$ is a **partition** of the interval $[r, s]$ and let $\Delta t_i = t_i - t_{i-1}$ for each relevant i.

When Q is continuous, $\sum_{i=1}^{n} Q(t_i)\Delta t_i$ will be an **approximation** to $\int_r^s Q(t)dt$.

The approximation[21] will be a good one provided that the largest of the Δt_i (this is called the **mesh** of the partition) is small enough.

18.2. ***Exercise***. *Suppose G, Q and H are continuously differentiable vector functions, u is an ordinary continuously differentiable function of t and $H(u(t))$ is defined on $[r, s]$. In the statements below, notation such as dQ is intended to be shorthand for $Q'(t)\; dt$.*

Using ordinary facts from Calculus, show that the following are true.

*(i) (**Integration by Parts**)* $H \cdot Q\Big|_r^s - \int_r^s H \cdot dQ = \int_r^s Q \cdot dH.$

*(ii) (**Integration by Parts**)* $H \times Q\Big|_r^s - \int_r^s H \times dQ = \int_r^s Q \times dH.$

*(iii) (**Integration by Parts**)* $u\; Q\Big|_r^s - \int_r^s u\; dQ = \int_r^s Q\; du.$

*(iv) (**Integration by Substitution**)* $\int_r^s H(u(t))\; u'(t)\; dt = \int_{u(r)}^{u(s)} H(u)\; du$

*This last formula is also called a **Change of Variables** formula.*

---

19. When is a Curve Confined to a Line or a Plane?

Our earlier ruminations about normal forms for lines in $2D$ or lines and planes in $3D$ provide the key to answering the section title question.

The reader should regard this section as one long exercise. In virtually every sentence there is something to check. This section contains interesting results: it would be useful to know if a curve representing something of interest is "really" one or two dimensional, rather than three dimensional. It would greatly simplify your visualization about what is happening with the curve. However the real point of placing this section here is that it gives the reader a chance to see if he or she understands the basic definitions of integrals and derivatives for vector functions.

If Q is a $2D$ vector function, and r is any particular parameter value and $P = Q(r)$ then Q will be confined to a line precisely when there is some nonzero vector N with $(Q - P) \cdot N = 0$. The choice of r is not relevant. The situation is the same (and for the same possible vectors N too) no matter what choice of r you pick.

If Q is a $3D$ vector function the situation is similar. Q will be confined to a plane precisely when there is some nonzero vector N with $(Q - P) \cdot N = 0$. Q will be confined to a line precisely when there are two nonzero vectors N_1 and N_2 with $(Q - P) \cdot N_1 = 0$ and $(Q - P) \cdot N_2 = 0$, where N_1 and N_2 are not multiples of each other. The line of confinement is the intersection of the two different planes. Once again, the specific r picked is not relevant.

So in both $2D$ and $3D$ we are led to consider Q for which $(Q - P) \cdot N = 0$ for some constant vector N and all values of the parameter t. Note, by the way, that if $(Q - P) \cdot N$ is constant for all values of the parameter then that constant must be zero.

In the following remarks we presume that Q is continuously differentiable.

A Necessary Condition For $(Q - P) \cdot N = 0$

For such a Q and if $(Q-P) \cdot N$ is always 0 we have $\dfrac{d}{dt}(Q-P) \cdot N = Q' \cdot N = 0$. In fact, $Q^{(n)} \cdot N = 0$ for any higher n–th derivatives which might exist as well. We have found a necessary condition for $(Q - P) \cdot N = 0$.

A Sufficient Condition For $(Q - P) \cdot N = 0$

We now change our point of view a bit and suppose we have a Q and N for which $Q' \cdot N$ is always 0. Then

$$0 = \int_r^t Q'(s) \cdot N \; ds = \int_r^t Q'(s) \; ds \cdot N = (Q(t) - Q(r)) \cdot N.$$

So $(Q-P) \cdot N$ is always 0. We have found a sufficient condition that guarantees that a constant vector N is always perpendicular to $Q - P$.

Finding an N or showing that none can exist for a particular Q is pretty easy in practice. Let's consider the $2D$ and $3D$ cases separately.

$2D$ Confined to a Line

In $2D$, pick any s for which $Q(s)-Q(r)$ or any $Q^{(n)}(s)$ is nonzero. (If this is not possible, then Q is constant and the curve is a single point, not a very interesting case.) If N is any nonzero vector perpendicular to this, then any possible vector normal to a line confining the curve must be a multiple of this N. Now examine $(Q-P)\cdot N$ or $Q'\cdot N$, whichever is handier. If either of these is always 0 you have found the line of confinement. If either is **ever** nonzero you have shown that the curve is not confined to any line.

$2D$ Confined to a Line: A Representation

If Q is confined to a line, it has an interesting representation. Let A be the unit vector $\dfrac{Q(s)-Q(r)}{|Q(s)-Q(r)|}$ where s has been chosen so that the difference $Q(s)-Q(r)$ is nonzero. (As before, if this is not possible then Q is the constant function, an uninteresting case.) Then $Q(t)=f(t)A+P$ where $f(t)=A\cdot(Q(t)-Q(r))$. That is because $Q-P$ can have no component in the direction of N, and A is perpendicular to N. There are three points to make here. The multiples of A form a line through the origin parallel to the one we want. Q has been represented as a multiple of a constant vector A plus the vector P, which gets you away from the line through the origin and out to the line of confinement. Second, $f(t)$ is at least as differentiable as Q, since it is made from sums of multiples of the coordinate functions of Q. Third, because A is taken to be a unit vector, the speed is $|f'(t)|$ for every t.

$3D$ Confined to a Line: A Representation

Now suppose that Q lives in $3D$. We will presume as before that Q is not constant. So there must be s for which $Q(s)-Q(r)\neq 0$. Let A be the unit vector $\dfrac{Q(s)-Q(r)}{|Q(s)-Q(r)|}$.

If $(Q(t)-Q(r))\times A=0$ or $Q'(t)\times A=0$ (whichever is easier to check) for all t then the angle between $Q-P$ or $Q'(t)$ and A is always 0.

In the first case, $Q(t)=f(t)A+P$ just as above, with the same interpretation: Q is confined to a line and $|f'(t)|$ is the speed of the motion for every t.

In the second case, $Q'(t)=g(t)A$ for continuous $g(t)=Q'(t)\cdot A$ so $Q(t)-Q(r)=\left(\int_r^t g(t)dt\right)A$ and, once again Q is confined to a line and has representation $Q(t)=f(t)A+P$ where $f(t)=\int_r^t g(t)dt$.

$3D$ Confined to a Plane

With A as above, we suppose there is at least one u for which $(Q(u)-Q(r))\times A\neq 0$ or $Q'(u)\times A\neq 0$, whichever is easier to check.

Let N be this nonzero cross product. If Q does live in a plane then the normal to this plane must be perpendicular to all of the $Q'(t)$ and also all differences $Q(t)-Q(r)$. Our vector N is perpendicular to two vectors which are not multiples of each other and which are perpendicular to any possible normal vector for the plane. So N must be a multiple of any possible normal to the plane: if the curve does lie in a plane, N is normal to that plane.

So $(Q(t)-Q(r))\cdot N = 0$ or $Q'(t)\cdot N = 0$ (whichever is easier to check) for all t or not. If not, then Q is not confined to any plane. If either (and hence both) are always 0 then Q is confined to a plane, and $(Q-P)\cdot N = 0$.

$3D$ Confined to a Plane: A Representation

When Q is confined to a plane there is a handy representation for Q. Given unit vector A as above let W be any vector in the plane which is not a multiple of A. W might be $Q(u)-Q(r)$ or $Q'(u)$, for example. Now let $V = W-(W\cdot A)A$ and define $B = \dfrac{V}{|V|}$. $A\times B$ is normal to the plane (and so is a nonzero multiple of N) and $A\cdot B = 0$. I claim that $Q(t) = f(t)A + g(t)B + P$, where $f(t) = Q(t)\cdot A$ and $g(t) = Q(t)\cdot B$. Also f and g are at least as differentiable as Q. The speed is $\sqrt{(f'(t))^2+(g'(t))^2}$.

19.1. ***Exercise***. *Go through the discussion above for the following vector functions and decide if they are confined to lines or planes. If they are, find the line or plane and also create a representation of the vector function as described above.*

$Q(t) = \langle 7+cos(t), t-3\rangle$

$J(t) = \langle 2t^2-4t, (t-1)^2\rangle$

$H(t) = \langle cos(t), sin(t), t\rangle$

$K(t) = \langle cos^2(t), sin^2(t), 3+3cos^2(t)\rangle$

$L(t) = \langle t^3-t^2+1, 3t^3+6t^2+5, 2t^3+3t^2+9\rangle$

19.2. ***Exercise***. *Suppose P is any vector and A and B are nonzero constant vectors which are not multiples of each other, but are not presumed to be unit vectors and are not presumed to be perpendicular to each other. Suppose f and g are differentiable functions.*

(i) $Q(t) = f(t)A + P$. *What is the speed of Q?*

(ii) $H(t) = f(t)A + g(t)B + P$. *What is the speed of H? Why is the formula given in the text above simpler?*

19.3. ***Exercise***. *Suppose $Q''\cdot N = 0$ for all t on an interval and a nonzero constant vector N. Then $Q(t) = tA + B + H(t)$ on that interval, for constant vectors A and B and a vector function H with $H\cdot N$ always 0. (hint: Make N a unit vector. Then $Q = (Q\cdot N)N + [Q-(Q\cdot N)N]$.)*

19.4. ***Exercise***. ** We have spent quite a bit of time thinking about $2D$ and $3D$ vectors, but what about $1D$? Suppose that the world consists of nothing but the real line. Let $\vec{i}$ in this world stand for the vector $\langle 1 \rangle$, which represents an arrow with tail at the origin and tip at the real number 1. Are there angles in this world? Dot products? A formula for constant velocity motion? Work? Derivatives and integrals of vector functions?*

$3D$ Confined to a Plane: Another Sufficient Condition

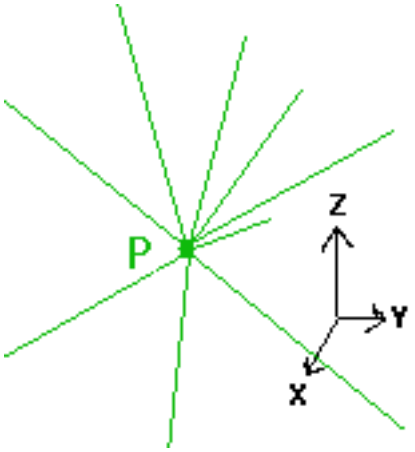
Suppose that $(Q(t) - P) \times Q''(t) = 0$ for all t. This situation happens a lot in applications: when the acceleration is due to a so-called **central force** acting on the moving point $Q(t)$ from the point P.

In that event $(Q(t) - P) \times Q'(t) = K$, a constant vector. Suppose $K \neq 0$. So $(Q(t) - P) \cdot K = 0$. This means that Q is confined to a plane.

19.5. ***Exercise***. *Suppose $(Q(t) - P) \times Q'(t) = 0$, the zero vector for all t in an open interval (a, b). This means that for each t either $Q(t) - P = 0$ or $Q'(t)$ is a multiple of $Q(t) - P$. This could happen, for example, in the case we did not consider above when $(Q(t) - P) \times Q''(t) = 0$ for all t.*

*(i) * Show that if $Q(t) - P$ is never 0 then the the curve for Q lies along a line segment emanating from P. (Hint: By assumption $Q'(t) = h(t)(Q(t) - P)$. Now let $Q(t) - P = |Q(t) - P| \frac{Q(t)-P}{|Q(t)-P|} = g(t)U(t)$ where U is a unit vector and g is never 0. Note $U \cdot U = 1$ so $\frac{d}{dt}(U \cdot U) = 2U \cdot U' = 0$. Now $h(t)g(t)U(t) = \frac{d}{dt}(g(t)U(t)) = g'(t)U(t) + g(t)U'(t)$. Dot the left and right terms against U' yielding $hgU \cdot U' = g'U \cdot U' + gU' \cdot U'$. This yields $0 = gU' \cdot U'$ so $U' = 0$.)*

*(ii) ** Show that if $Q(t) - P$ is ever 0 then the the curve for Q lies along a collection of line segments emanating from P. At most two of these segments contain an unbounded part of the curve.*



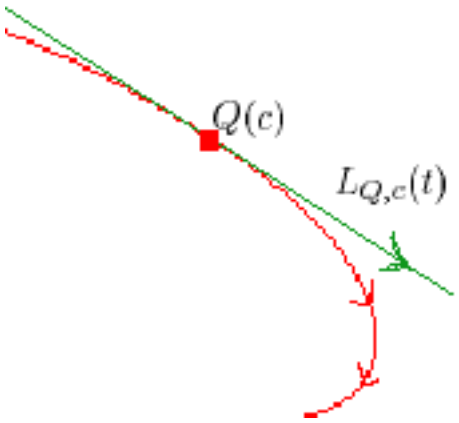
20. Tangent Lines

In this section all of our vector functions will be **continuously differentiable.** Recall that if Q is a vector function which is differentiable at c then $|\Delta Q - Q'(c)\Delta t|$ is small **even in comparison to** Δt when Δt is small, where ΔQ is shorthand for $Q(c+\Delta t) - Q(c)$.

Rephrasing this, we have:

$$L_{Q,c}(t) \;=\; Q'(c)(t-c) + Q(c) \overset{\text{small } |t-c|}{\approx} Q(t).$$

The left hand side is a linear motion with constant velocity vector $Q'(c)$ and shares the point $Q(c)$ with our vector function Q at $t = c$. Q will be near to $L_{Q,c}$ (in comparison to $|t-c|$) when $|t-c|$ is small, so $L_{Q,c}$ could be used to approximate Q for t near to c. $L_{Q,c}$ is called the **linearization** of Q at c.



Let's think about this another way. If Q represents the track of a moving particle and if Q' is not constant then there must be forces present to disturb the motion of the particle. Forces are the causes of a change in the motion, and a change in the motion means there is a net force acting on the object. These forces can be (and usually are) generated by something external, but you can think of the forces as coming from a rocket motor attached to our particle and pushing it along the curve of Q. Alternatively, we can imagine the curve of Q to be a rigid wire to which our particle is bound by wheels. The forces come from the effect of brakes or motors mounted on the moving particle. All the different parameterizations of this curve come from different starting times and different ways of putting on the "gas" or the "brakes."

$L_{Q,c}$ represents the motion of our particle after time $t = c$ if the rocket motor quit at $t = c$. Using the other analogy, $L_{Q,c}$ represents the motion of the particle if the "clamp" which binds the particle to the curve were released at $t = c$.

Problem: Hitting a Target Point

Suppose we have a moving particle clamped to a wire and scheduled to be at position $Q(t) = \langle 2t^2, -t+1\rangle$ at each time t. If the driver wishes to reach the point $\langle -6, 0\rangle$ when should she release the clamp? How long after release will she arrive at this spot?

The Solution:

When the clamp is released the particle will carry on with whatever velocity it had at the moment of release. So it will hit the target at $\langle -6, 0\rangle$ if we can find c for which $\langle -6, 0\rangle - Q(c)$ is a **positive** multiple of $Q'(c)$.

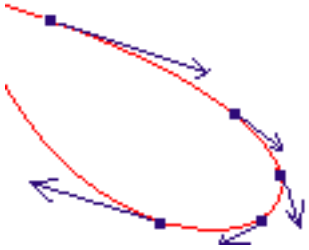
$$\langle -6, 0\rangle - \left\langle 2t^2, -t+1\right\rangle = \left\langle -6-2t^2, t-1\right\rangle = K\left\langle 4t, -1\right\rangle.$$

This yields the two equations:

$$-6 - 2t^2 = 4Kt \quad \text{and} \quad t - 1 = -K.$$

Solving for K in the second of these and substituting into the first yields $t = -1$ and $t = 3$. However $t = 3$ yields a negative K (the particle is moving exactly **away** from the target at that time) while $t = -1$ yields $K = 2$. So she should release the clamp at time $t = -1$. She will arrive at the target 2 seconds later.

A useful way of thinking of the changing velocity vector for Q is to visualize for each c the copy of $Q'(c)$ which has its tail at $Q(c)$. The line along $Q'(c)$ will always graze the curve at $Q(c)$ and show the direction a particle would fly away if all the forces were turned off as it passed that spot. Its length is the speed of the parameterization and gives a sense of how fast the curve is being traversed by Q at $Q(c)$ relative to its speed at other places on the curve.



The next idea we will consider here is the concept of the tangent line to a curve at a point P on a curve. The concept of the linearization $L_{Q,c}$ of Q is associated with the specific parameterization of the curve upon which Q lives. If $P = Q(c)$ and $Q'(c) \neq 0$, the linearization of Q at c gives us a parameterization of a line which grazes the curve at P and it makes sense to call this line the tangent line to the curve at P. However a given curve can have many different parameterizations. Q, for example, can be tweaked to produce lots of them. If r and s are any real numbers with $r \neq 0$ then $W(t) = Q(r(t-s))$, with t between $\frac{a-rs}{r}$ and $\frac{b-rs}{r}$, also traces out all the same points that Q does, and is as differentiable as Q. s is related to a **shift of center** and r to a **speed change** for the motion along the curve. Specifically, $W(\frac{c-rs}{r}) = P$ and $W'(t) = rQ'(r(t-s))$. So when $\widehat{c} = \frac{c-rs}{r}$ we have $W(\widehat{c}) = P$ and $W'(\widehat{c}) = rQ'(c)$. This means that the linearization $L_{W,\widehat{c}}$ traces out the same collection of points as does $L_{Q,c}$. Even more, we have shown that **any vector that lies in this line** is $W'(\widehat{c})$ for at least one differentiable parameterization of the curve, where $W(\widehat{c}) = P$.

20.1. ***Exercise***. *Prove the last sentence.*

20.2. ***Exercise***. *Suppose Q parameterizes a curve. We are interested in the section of the curve corresponding to $[\alpha, \beta]$, a part of the domain of Q. Suppose A is a position vector.*

*(i) * There is (at least one) point on the section of the curve which is nearest and (at least one) point farthest from A. (hint: Let $dist(t) = \sqrt{(Q(t) - A) \cdot (Q(t) - A)}$ for t in the interval $[\alpha, \beta]$. This function is continuous and so actually attains its maximum and minimum.)*

*(ii) ** The minimum distance is 0 only when $A = Q(\gamma)$ for some γ in $[\alpha, \beta]$. (hint: If the distance is 0 there is a sequence of times t_n in the interval $[\alpha, \beta]$ for which $dist(t_n)$ converges to 0. So there is a subsequence u_n of t_n which converges to some number v in $[\alpha, \beta]$. The continuity of Q requires that $\sqrt{(Q(v) - A) \cdot (Q(v) - A)} = 0$ which means $A = Q(v)$.)*

20.3. ***Exercise***. ** If $Q'(c) \neq 0$ then there is an interval $(c - \varepsilon, c + \varepsilon)$ around c upon which Q is one-to-one: that is to say, If r and s are in $(c - \varepsilon, c + \varepsilon)$ and $r \neq s$ then $Q(r) \neq Q(s)$. (hint: If $Q'(c) \neq 0$ then continuity of Q' implies that the derivative of at least one of the coordinate functions of Q is always positive or always negative on some interval around c.)*

20.4. ***Exercise***. *** Suppose W and Q are any two continuously differentiable parameterizations of the same curve. Suppose that $Q(c) = W(\widehat{c}) = P$ and both $Q'(c)$ and $W'(\widehat{c})$ are nonzero. Then there is a nonzero number r with $Q'(c) = r\ W'(\widehat{c})$. To prove this show the following:*

(i) In view of the discussion involving "shift of center" from above we may assume that $c = \widehat{c} = 0$.

(ii) Because of the last exercise we may assume that both Q and W are one-to-one on intervals around 0. Conclude that there are such intervals $I_t = [a_t, b_t]$ and $I_u = [a_u, b_u]$ with $a_t < 0 < b_t$ and $a_u < 0 < b_u$ and for each member t of I_t there is exactly one member u of I_u with $Q(t) = W(u)$ and also for each u in I_u there is exactly one t in the interval I_t with $Q(t) = W(u)$. So u can be regarded as a function of t on I_t: for each t in I_t let $u(t)$ be the unique member of I_u for which $Q(t) = W(u(t))$.

(iii) Show $lim_{t \to 0} u(t) = 0$. (hint: If $lim_{t \to 0} u(t) \neq 0$ there would be a sequence t_n converging to 0 but for which $u(t_n)$ converges to a nonzero number v in I_u. Continuity of W would require $W(u(t_n))$ to be near $W(v)$ and not P for large n, contradicting the fact that $W(u(t_n)) = Q(t_n)$ which converges to $Q(0) = P$.

(iv) From parts (i) and (ii) we find for any nonzero associated t and u that

$$\frac{Q(t) - P}{t} = \left(\frac{W(u) - P}{u}\right)\left(\frac{u}{t}\right).$$

(v) From (iii) the fractions involving Q and W converge to $Q'(0)$ and $W'(0)$ respectively as t goes to 0. At least one of the coordinate functions of $Q'(0)$ is

nonzero, and after thinking about that, conclude that $\lim_{t\to 0}\frac{u}{t}$ *exists and is nonzero. This number is the* r *we were looking for and is, in fact,* $\frac{du}{dt}(0)$.

Because of this last exercise we know the following: Suppose P is a point on a curve and there is **any** continuously differentiable parameterization Q of the curve with $Q(c) = P$ and $Q'(c) \neq 0$. Then a nonzero vector V lies in the geometrical track of $L_{Q,c}$ exactly when $V = W'(\widehat{c})$ for some continuously differentiable parameterization W of the curve with $W(\widehat{c}) = P$. This line (the geometrical track of any $L_{Q,c}$ where $Q(c) = P$ and $Q'(c) \neq 0$) is called the **tangent line** to the curve at P. A tangent line will exist except at places on the curve where the derivative of **every continuously differentiable parameterization is** 0. A place like that on a curve is called a **cusp**. We will see an example of a cusp in the next section.

---

21. A Cycloid and Bezier Curves

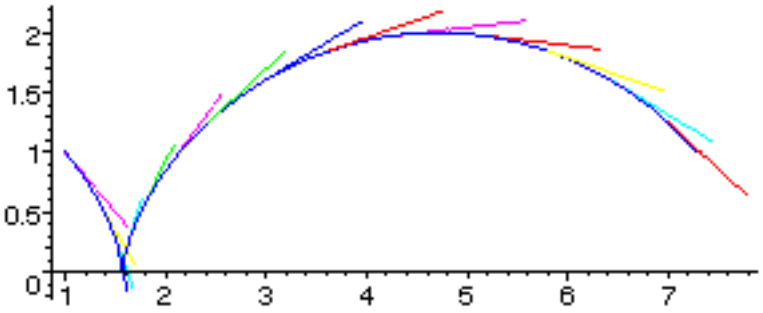
Let's take a look at a cycloid example with parameterization

$$Q(t) = \langle .2\pi t + cos(.2\pi t), 1 - sin(.2\pi t)\rangle .$$

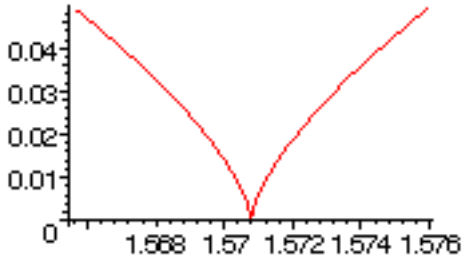
The derivative of Q is

$$Q'(t) = \langle .2\pi - .2\pi sin(.2\pi t), -.2\pi cos(.2\pi t)\rangle .$$

We attach a few velocity vectors at the place where they are relevant on the curve and plot the curve as t varies from 0 to 10 below.



Something interesting is going on near $x = \frac{\pi}{2}$ (this corresponds to $t = 2.5$) on this graph. If y is thought of as a function of x something bad happens here—a sharp point. It seems that the velocity vectors will switch direction instantly from straight down to straight up as they pass through the point $(\frac{\pi}{2}, 0)$. However nothing dramatic seems to be happening with formula for Q' around $t = 2.5$. What is going on?



The magnitude of $Q'(t)$ is found to be (after a calculation) $\sqrt{.8}\pi|cos(.2\pi t)|$. At $t = 2.5$ the velocity is the zero vector. So continuous $Q'(t)$ manages to switch instantly from pointing down to pointing up by passing through the zero vector—everything slows to a stop for an instant.

21.1. ***Exercise***. *** Could there be a parameterization for this cycloid with nonzero velocity at "the point?"*

The second topic in this section involves an interesting collection of functions which have applications in computer graphics and engineering design—these are called **Bezier curves**. They come into play when you want to smoothly patch together curves to create a shape without sharp corners, and when computational considerations are important to you. Computers are fast at adding and multiplying and that is how polynomials are calculated. We would like to have a vector function whose coordinate functions are polynomials with lowest possible degree with specified starting and ending positions and velocities.

The conditions on starting and ending positions and velocities specify four equations involving the coefficients of each coordinate polynomial. We will see that these four equations can be accommodated if we use a third (but not lower) degree polynomial in the parameter at each coordinate.

$$B(t) = \langle a_3t^3 + a_2t^2 + a_1t + a_0,\ b_3t^3 + b_2t^2 + b_1t + b_0,\ \dots \rangle$$

21.2. ***Exercise***. *(i) Show that there is one and only one third degree polynomial $P(t)$ with $P(0) = A$, $P(1) = B$, $P'(0) = C$ and $P'(1) = D$ for each choice of A, B, C and D. Letting $A = 1$, $B = 1$, $C = 1$ and $D = 1$ we note that a second degree polynomial won't in general do the job. (This one is not too hard.)*

*(ii) * More generally show that there is one and only one third degree polynomial with $P(0) = A$, $P(b) = B$, $P'(0) = C$ and $P'(b) = D$ for any nonzero b and any choice of A, B, C and D. (hint: From (i) we know that there is a third degree polynomial $Q(t)$ with $Q(0) = A$, $Q(1) = B$, $Q'(0) = bC$ and $Q'(1) = bD$. The polynomial $P(t) = Q(\frac{t}{b})$ is the one we want.)*

*(iii) * Finally, show that there is one and only one third degree polynomial with $P(a) = A$, $P(b) = B$, $P'(a) = C$ and $P'(b) = D$ for any a and b with $a \neq b$ and*

any choice of A, B, C and D. (hint: From (ii) we know that there is one and only one third degree polynomial $Q(t)$ with $Q(0) = A$, $Q(b-a) = B$, $Q'(0) = C$ and $Q'(b-a) = D$. The polynomial $P(t) = Q(t-a)$ is the one we want.)

In constructing our Bezier curves, let's focus first on the parameter interval $[0, 1]$ and then use our shifting and speed change ideas from above on the more general cases.

21.3. ***Exercise***. *Suppose A, B, V_A and V_B are specified vectors.*

(i) The function

$$Bezunit(t) = (1-t)^3 A + 3(1-t)^2 t(A + \frac{1}{3}V_A) + 3(1-t)t^2(B - \frac{1}{3}V_B) + t^3 B$$

has the following values:

$$Bezunit(0) = A \quad Bezunit(1) = B \quad Bezunit'(0) = V_A \quad Bezunit'(1) = V_B.$$

(ii) $1 = (1-t+t)^3 = (1-t)^3 + 3(1-t)^2 t + 3(1-t)t^2 + t^3$ and these four numbers are nonnegative for t between 0 and 1. So $Bezunit(t)$ is a weighted average of the four vectors A, $A + \frac{1}{3}V_A$, B and $B + \frac{1}{3}V_B$. Among other things, this tells us that each coordinate of $Bezunit(t)$ for t between 0 and 1 is intermediary between the biggest and smallest of the corresponding coordinates of these four vectors. $Bezunit(t)$ cannot get too big or too small.

(iii) $Bezunit(t)$ has velocity vector

$$Bezunit'(t) = (1-t)^2 V_A + 2(1-t)t(3B - 3A - V_A - V_B) + t^2 V_B.$$

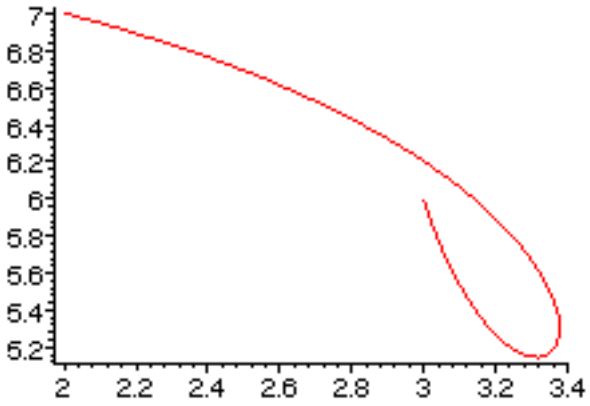
Once again, this vector is a weighted average, since $1 = (1-t+t)^2 = (1-t)^2 + 2(1-t)t + t^2$. We have some control over how big our speed can get in terms of the four initial vectors.

(iv) $Bezunit(t)$ has acceleration vector

$$Bezunit''(t) = (1-t)(6B - 6A - 2V_B) + t(6A - 6B + 2V_A + 4V_B).$$

This vector is a weighted average too. We have control over how big a rocket motor we must have on our moving point to produce this motion.

Here is a picture[22] of *Bezunit* when $A = \langle 2, 7 \rangle$, $B = \langle 3, 6 \rangle$, $V_A = \langle 6, -3 \rangle$ and $V_B = \langle -1, 6 \rangle$.



The next step is to do a shift and speed change, to allow for parameter intervals other than $[0,1]$. If we want the action to take place over the interval $[c,d]$ instead of $[0,1]$ we could look at the composite function $Bezunit\left(\frac{t-c}{d-c}\right)$. That will get the endpoints right. However there is a problem with the derivatives.

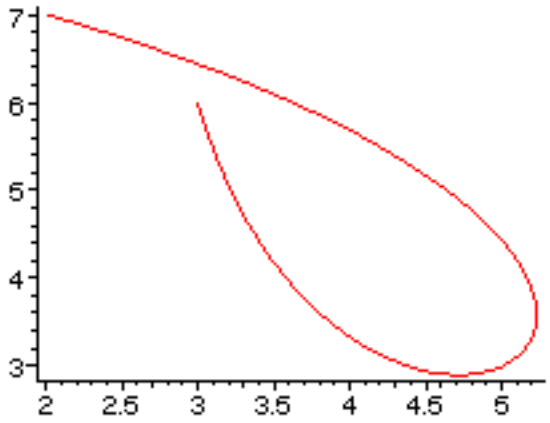
$$\frac{d}{dt}Bezunit\left(\frac{t-c}{d-c}\right)=\frac{1}{d-c}Bezunit'\left(\frac{t-c}{d-c}\right).$$

We have changed the length of the derivative by a factor $\frac{1}{d-c}$. Left unmodified, the derivatives at the endpoints will be off by that factor. We have to replace V_A and V_B in the original formula for $Bezunit$ by $(d-c)V_A$ and $(d-c)V_B$. After some algebra and cleanup this is what we get:

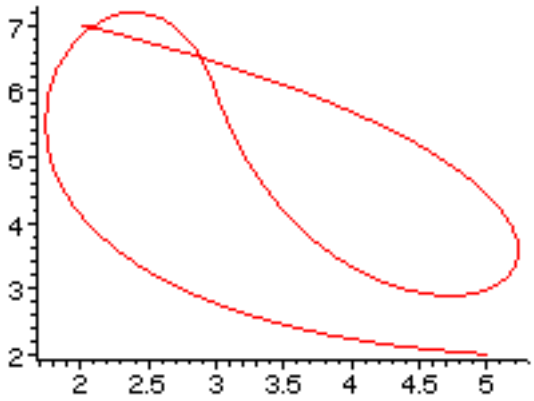
$$\begin{aligned}Bezspeed(t)=&\frac{(d-t)^3}{(d-c)^3}A+\frac{3(t-c)(d-t)^2}{(d-c)^3}[A+\frac{(d-c)}{3}V_A]\\&+\frac{3(t-c)^2(d-t)}{(d-c)^3}[B-\frac{(d-c)}{3}V_B]+\frac{(t-c)^3}{(d-c)^3}B.\end{aligned}$$

21.4. ***Exercise***. *Verify that* $Bezspeed(c)=A$, $Bezspeed(d)=B$, $Bezspeed'(c)=V_A$ *and* $Bezspeed'(d)=V_B$.

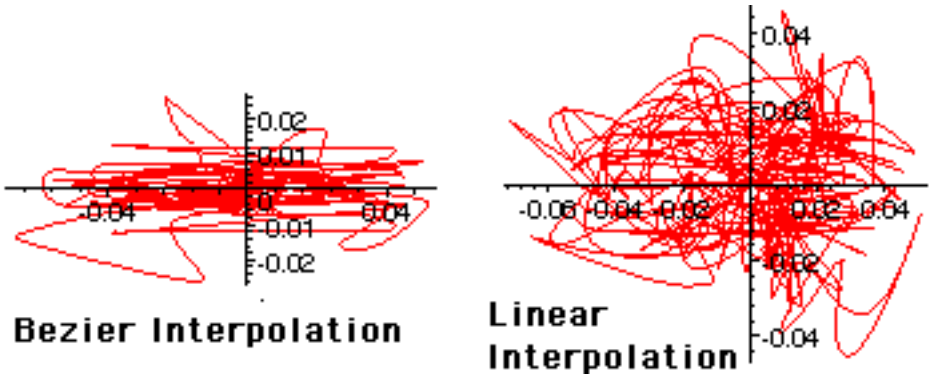
Here is a picture[23] of $Bezspeed$ when $A=\langle 2,7\rangle$, $B=\langle 3,6\rangle$, $V_A=\langle 6,-3\rangle$ and $V_B=\langle -1,6\rangle$, just as in $Bezunit$ above, but with time interval between 6 and 9 rather than between 0 and 1. Notice that the curve loops around more slowly. That is because the acceleration is smaller than before, but the time during which it acts is three times longer.



We can now patch a couple of these curves together, with matching positions and derivatives at the joining spot. Find below a picture[24] of $Bezpatch(t)$ with parameter interval 6 to 11. From 6 to 9 it is just as above. The starting position and velocity on the second piece, from time 9 to 11, matches the final position and velocity of the first piece. The final position is the point $C = \langle 5, 2\rangle$ with final velocity $V_C = \langle 8, -1\rangle$.



As a final illustration we can use the Bezier patching function to glue together the 101 data points from the example in Section 14. We found a final wiggly error function that we could not explain and wondered if our theory about how the points should be connected was wrong or if they represented data acquisition uncertainty or perhaps they were an artifact of the linear interpolation method we used to "connect the dots." We can patch these points together and interpolate using Bezier functions. The velocity we use to round out the corner at data point P_k at time T_k is the average velocity over the neighboring time intervals: $\frac{P_{k+1}-P_{k-1}}{T_{k+1}-T_{k-1}}$. When we do this for the middle 99 data points and subtract off the linear and both circular motions as before[25] we get the following.



You will notice a substantial diminution of the Y coordinate wiggle but not much obvious improvement in the X coordinate discrepancy from our "theory."

21.5. ***Exercise***. ** The source of the discrepancy seen above and why the fit got better on one axis and not another is an interesting puzzle. Why, in fact, is the size of the discrepancy what it is? To discover what is going on it is necessary to look at the data set itself (in the endnotes) and not just at the picture. It is possible to make reasonable conjectures about where I probably got this data. (Hint: I did not "make the data up" and I did not obtain the data from observations!) After examining the structure of the data, you can then conclude that what we see in the Bezier wiggles is in line with the best you could expect. Examining subcollections of the data could provide evidence for your conjecture.*

---

22. Line Integrals

In this section we are going to think of a curve as more than merely a collection of points strung together, but as a physical object—a wire perhaps—with properties such as weight, length and so forth. If you know about such things you could also imagine positive and negative charges arrayed along the wire. We might be interested in the total charge, length or mass on a piece of the curve.

We will presume that the piece, which we will denote $\mathfrak{C}$, of the curve which has caught our attention is **parameterized by a continuously differentiable vector function** Q on the interval $[c, d]$. We are going to use Calculus (derivatives and integrals) to produce a number which we will interpret as the length of the curve between the points $C = Q(c)$ and $D = Q(d)$ on the curve. We will call this number the arclength.

We presume that the curve does not cross itself by insisting that Q be one-to-one on $[c, d]$ or that Q be one-to-one on $(c, d]$ and $Q(a) = Q(b)$. If the latter condition holds we call $\mathfrak{C}$ a **loop**.

We will further presume that Q **does not slow down to zero velocity anywhere on** $[c, d]$.

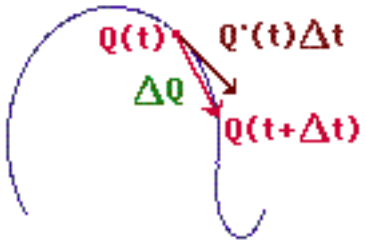
A parameterization of $\mathcal{C}$ with these three properties will be called a **good parameterization** of $\mathcal{C}$.

If $\mathcal{C}$ has a good parameterization, $\mathcal{C}$ is called a **good curve**.

If $\mathcal{C}$ has a good parameterization Q as above with $Q(a) = Q(b)$ then $\mathcal{C}$ is also called a **good loop**.

If a curve of interest contains a few cusps or crossing places we will have to break it up into pieces with the cusps or crossing places on the ends and deal with the pieces individually.

These assumptions are important. It is possible to cook up a curve (which lacks the differentiability property we presume) that zig-zags so wildly that it doesn't make sense to ascribe any length to it—even what seems to be the smallest piece has "infinite[26] length." The second and third conditions are more along the lines of conveniences. If we let Q slow down to zero velocity it could reverse direction and trace over a recently traversed part of the curve a second time "backwards" and complicate our calculation. The third condition gets at this problem from a larger perspective: we also don't want Q to loop around and retrace some distant part of the curve a second time.



Let's look at a small piece of good $\mathcal{C}$. If you examine the picture it seems that for small Δt the length along the curve from $Q(t)$ to $Q(t + \Delta t)$ should be nearly $|Q(t + \Delta t) - Q(t)| = |\Delta Q|$. But $|\Delta Q|$ itself is nearly $|Q'(t)\Delta t|$. This is a "double approximation." The curve length seems to be about the straight line length which is nearly $|Q'(t)\Delta t|$ for small Δt.

We break the curve up into many little pieces corresponding to partitioning $[c, d]$ by selecting times $c = t_0 < t_1 < \cdots < t_N = d$. If we let $\Delta t_i = t_i - t_{i-1}$ and $\Delta Q_i = Q(t_i) - Q(t_{i-1})$ for i between 1 and N then we would imagine that the length along the curve would be close to $\sum_{i=1}^{N} |\Delta Q_i|$ which itself would be nearly $\sum_{i=1}^{N} |Q'(t_i)|\Delta t_i$. This is a Riemann sum, and because Q has continuous derivative will be arbitrarily close to $\int_c^d |Q'(t)|dt$ if the mesh of the partition is small enough. It is this integral which is **defined** to be the **arclength along the curve from** $C = Q(c)$ **to** $D = Q(d)$.

There are two things we should be worried about here.

First, though the transition from the Riemann sum $\sum_{i=1}^{N} |Q'(t_i)|\Delta t_i$ to the integral $\int_c^d |Q'(t)|dt$ is solid, the transition from $\sum_{i=1}^{N} |\Delta Q_i|$ to $\sum_{i=1}^{N} |Q'(t_i)|\Delta t_i$ is problematic. In going from ΔQ_i to $|Q'(t_i)|\Delta t_i$ we will usually be making a small error on each segment. Since the number of segments increases without bound,

these small errors could amount to something. We deal with this little wrinkle[27] in the endnotes.

Second, and assuming the first problem to be dealt with, it seems that the arclength depends on Q. But if it means what we think it means it should depend only on the relevant part of the curve between the points C and D. Q was just a tool to calculate this number. If there were some other parameterization W of that part of the curve which satisfies our requirements on a time interval $[e, f]$ we would like to know that the arclength calculation with this new parameterization yields the same result.

Each time t in $[c, d]$ and each time u in $[e, f]$ corresponds to exactly one point on the curve. So u can be thought of as a function of t by letting $u(t)$ be that member of $[e, f]$ for which $W(u(t)) = Q(t)$. We need the result from the following exercise:

22.1. ***Exercise***. ** We assume that both W' and Q' are continuous and nonzero. We conclude (after some work!) from Exercise 20.4 (iii) that u is continuous. From this and Exercise 20.4 (iv) and (v) we conclude that $u'(t)$ exists. Now the equation $Q'(t) = W'(u(t))\ u'(t)$ and, once again, the fact that both Q' and W' are nonzero and continuous imply that $u'(t)$ is continuous and nonzero. We conclude that the "translation function" $u(t)$ from one time measuring scheme along the curve to another is continuously differentiable and, under our conditions, never slows to a halt.*

We can now conclude that arclength does not depend on the parameterization through an application of integration by substitution.

$$\int_{t=c}^{t=d} |Q'(t)|\ dt = \int_{t=c}^{t=d} |W'(u(t))|\ |u'(t)|\ dt = \int_{u=e}^{u=f} |W'(u)|\ du.$$

22.2. ***Exercise***. ** $|u'(t)|\ dt$ was replaced in the integral from above by du. Why was that correct?*

To illustrate the ideas from above we suppose Q is a parameterization of the type we have been discussing and that g is a real valued function defined for points on the curve. (Actually, in many cases g will be defined for every point, not just those on the curve, but that doesn't matter here.) We also presume that $g(Q(t))$ is continuous for t in $[c, d]$. The function g can be thought of as a **linear mass density** function if you wish. If it is nonnegative, you can think of its value at a point on the curve as representing the mass per unit length along the curve at the point. The mass of a little segment of the curve from $Q(t)$ to $Q(t + \Delta t)$ will be approximately the linear density $g(Q(t))$ at the spot $Q(t)$ times the length along the curve. In our case that is nearly $g(Q(t))\ |\Delta Q|$ which, itself, is not too far from $g(Q(t))\ |Q'(t)\Delta t|$. Adding up all these contributions and passing to the integral

yields the **total mass of a wire** laid along the curve with this density function.

$$\textbf{Mass of the Wire} \quad = \quad \int_c^d g(Q(t))\ |Q'(t)|\ dt.$$

If g has both positive and negative values you could think of it as **linear charge density**, with some positively and some negatively charged places on the wire. The integral would represent the **total charge on the wire.**

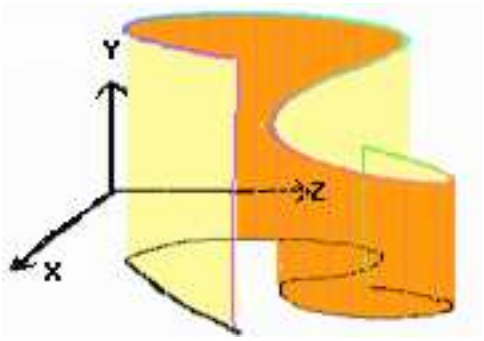
22.3. ***Exercise***. *Show that the integral for the mass or charge on a wire does not depend on the parameterization. Specifically, consider what happens to the integral if a new parameterization traces over the same piece of the curve but in the opposite direction.*

Whatever the interpretation as mass or charge or something else from a different application, an integral such as $\int_c^d g(Q(t))\ |Q'(t)|\ dt$ is called the **integral of g weighted by arclength** or simply the **line integral of g on the curve** and in the last exercise you showed that line integrals depend on the piece of curve traversed during the integration and the values of g on that piece but **not the specific parameterization** of the curve.

The integral $\int_c^d g(Q(t))\ |Q'(t)|\ dt$ may be written $\int_C^D g(Q)\ |dQ|$ or even $\int_{\mathfrak{C}} g(s)\ ds$ to de-emphasize the apparent dependence on the parameterization. In a practical sense, however, the parameterization is not irrelevant: in most cases *some* parameterization is needed to calculate a line integral.

22.4. ***Exercise***. *Consider the helix $Q(t) = \langle cos(t), sin(t), t\rangle$ for t between 0 and 2π, where all distances are in meters. Suppose that the linear density of a wire laid along this curve increases with height above the XY plane according to the function $g(X, Y, Z) = 3Z$ kilograms per meter. What is the arclength of the curve between $\langle 1, 0, 0\rangle$ and $\langle 1, 0, 2\pi\rangle$? What is the mass of this piece of wire?*

Here is another example. We consider a curve parameterized by $Q(t) = \langle X(t), Z(t)\rangle$ for t in the interval $[c, d]$ wandering around in the XZ plane. We will presume that $Y(Q(t))$ is nonnegative and continuous for t in $[c, d]$. The integral $\int_c^d Y(Q(t))\ |Q'(t)|\ dt$ can be interpreted as the **area of a curtain** hanging from a bent "curtain rod" along the curve $\widetilde{Q}(t) = \langle X(t),\ Y(Q(t)),\ Z(t)\rangle$. The value $Y(Q(t))$ is the length of the curtain hanging from each spot on the curve to the "floor" on the XZ plane. Although this scenario is nothing more than a reinterpretation of Y as "length of curtain" rather than "mass" it does bring a different idea into the mix.



23. Orientation of Curves and Line Integrals

In many applications it is important to distinguish a preferred direction along a good curve $\mathcal{C}$.

Suppose $\mathcal{C}$ is not a loop. In the last section you showed that if W and Q are two good parameterizations of $\mathcal{C}$ and if u is defined for each t by $W(u) = Q(t)$ then u is a differentiable function of t with continuous nonzero derivative. So this derivative must have constant sign in (c, d).

This means that good parameterizations of $\mathcal{C}$ fall into two groups: those that parameterize $\mathcal{C}$ in "one direction" and those that traverse $\mathcal{C}$ in "the other direction."

23.1. ***Exercise***. ** If $\mathcal{C}$ is a good loop, show that the function u from above is defined except for at most three values of t and that u' has constant sign elsewhere. So parameterizations fall into two groups, just as above.*

A selection of a preferred direction for a good curve corresponds to a choice of one of these two groupings of good parameterizations, and is called an **orientation for the curve**. The curve together with this orientation is called an **oriented curve**.

An orientation is often specified, to downplay the significance of any particular parameterization, by giving a unit tangent vector pointing in the preferred direction for all, or all but at most two, points on $\mathcal{C}$. If Q is any particular good parameterization that belongs to the specified orientation we can define the **unit tangent vector along the curve for the orientation** as

$$\mathcal{T}(Q(t)) = \frac{Q'(t)}{|Q'(t)|}$$

for each t in (c, d). Had Q belonged to the other orientation, $\mathcal{T}$ would be defined as the negative of the ratio shown above. We emphasize that, though the unit vector $\mathcal{T}$ is calculated for the points along $\mathcal{C}$ by a formula involving a parameterization, it does not depend on this parameterization but only on the geometry of $\mathcal{C}$ and a choice of orientation.

23.2. ***Exercise***. * *$\mathfrak{T}(Q(t))$ is a continuous vector valued function of t on (c,d). It will be possible to define it to be a continuous function of t on all of $[c,d]$ unless $\mathfrak{C}$ is a loop and $Q(c)$ is at a cusp of $\mathfrak{C}$.*

*Recall that a point K of $\mathfrak{C}$ is called a **cusp** if, whenever P is any differentiable parameterization defined on an interval (q,r) whose values are contained in $\mathfrak{C}$, and if $P(t) = K$ for some t, then $P'(t) = 0$.*

We illustrate some of the ideas of this section by recalling from Section 9 that mysterious "something" called work (trust me, you will see that it is important later) for a linear displacement and a constant force vector. It is a number defined to be the scalar projection of the force in the direction of displacement times the magnitude of the displacement and calculated as the dot product of force by displacement.

We can use the ideas here to define work for a curvy movement and where the force might vary from place to place or from time to time.

Suppose $Q(t)$ is any continuous parameterization of a good curve $\mathfrak{C}$ on the interval $[c,d]$. Suppose that for each time t in $[c,d]$ we are given a force vector $F(t)$ and suppose the vector valued function F is continuous. Let $c = t_0 < t_1 < \cdots < t_N = d$ be a partition of the interval and define $\Delta Q_i = Q(t_i) - Q(t_{i-1})$ for i between 1 and N as before. ΔQ_i is close to the movement along the curve if $\Delta t_i = t_i - t_{i-1}$ is small enough. F cannot vary much on a small interval $[t_{i-1}, t_i]$. So it makes sense to say that the work done moving along the curve in the time interval $[t_{i-1}, t_i]$ is close to $F(t_i) \cdot \Delta Q_i$ and the work done during the entire motion is nearly $\sum_{i=1}^{N} F(t_i) \cdot \Delta Q_i$. Under our conditions this is nearly $\int_c^d F(t) \cdot Q'(t)\ dt$ provided that the mesh of the partition is small. This integral is the **work** done in traversing the curve with this parameterization subject to this force function.

Sometimes F is related to the velocity at various places of a fluid within which the curve is immersed rather than a force. In this case the same integral is called the **flow** or, if the curve is a loop, the **circulation** along the oriented curve.

In either case, this number is not a line integral on the curve: it definitely could depend on the parameterization.

However, it often happens that the change in F as you move along with a parameterization of the curve is caused by the change in position, not really the change in time: that is, if $P(t)$ is any other parameterization of the same curve on the interval $[a,b]$ then the force felt—or the velocity of the fluid, in the second interpretation—at $P(t)$ is the same as that at $Q(s)$ whenever $P(t) = Q(s)$.

In that event, we can define work done by the force as a line integral along the curve, provided that we first specify an orientation $\mathfrak{T}$ for the curve. We calculate the work done under the influence of F when you move along the curve $\mathfrak{C}$ with orientation $\mathfrak{T}$ as

$$\textbf{Work or Flow due to F along Oriented } \mathfrak{C}: \int_c^d F(Q(t)) \cdot \mathfrak{T}(Q(t))\ |Q'(t)|\ dt$$

where Q is any good parameterization of the curve.

This integral is frequently written in shorthand as $\int_{\mathfrak{C}} F(s) \cdot \mathfrak{T}(s)\ ds$ to de-emphasize the parameterization.

However, except in unusual circumstances, a parameterization is required to carry out the calculation, which becomes

$$\begin{aligned}\int_c^d F(Q(t))\cdot\mathfrak{T}(Q(t))\ |Q'(t)|\ dt& \\ = \pm\int_c^d F(Q(t))\cdot\frac{Q'(t)}{|Q'(t)|}\ |Q'(t)|\ dt &= \pm\int_c^d F(Q(t))\cdot Q(t)\ dt \\ = \pm\int_c^d F(Q)\cdot dQ&\end{aligned}$$

with "plus" chosen if the orientation $\mathfrak{T}$ agrees with the direction of Q and "minus" otherwise.

23.3. ***Exercise***. ** In this exercise we think about how to define the integrals above when the parameterizations are not quite so nice.*

Suppose that Q is a continuous parameterization of a curve $\mathfrak{C}$ with domain $[c,d]$ and that Q is one-to-one on $[c,d]$ or Q is one-to-one on $(c,d]$ and $Q(c)=Q(d)$. We suppose that the domain $[c,d]$ can be broken into a finite number of pieces $[t_{i-1},t_i]$ for $i=1,\dots,n$ so that Q is continuously differentiable with nonzero derivative on each interval (t_{i-1},t_i). A parameterization of this kind is called a ***piecewise good parameterization*** *of $\mathfrak{C}$. The curve $\mathfrak{C}$ is called a* ***piecewise good curve*** *by virtue of the existence of any piecewise good parameterization of the curve. If there is any piecewise good parameterization of $\mathfrak{C}$ with $Q(c)=Q(d)$ we call $\mathfrak{C}$ a* ***piecewise good loop****.*

Now suppose that P and Q are any two piecewise good parameterizations of $\mathfrak{C}$ with domains $[a,b]$ and $[c,d]$.

We suppose that g is a real valued function defined along the curve and $g\circ P$ and $g\circ Q$ are continuous. We suppose that F is a vector valued function defined along the curve and $F\circ P$ and $F\circ Q$ are continuous.

(i) Show that $\int_c^d g(Q(t))|Q'(t)|\ dt=\int_a^b g(P(t))|P'(t)|\ dt$.

(ii) Both $\mathfrak{T}_Q=\dfrac{Q'}{|Q'|}$ and $\mathfrak{T}_P=\dfrac{P'}{|P'|}$ are defined except for a finite number of points in their respective domains.

Show that either

$\mathfrak{T}_P(u)=\mathfrak{T}_Q(t)$ whenever both vectors exist and $Q(t)=P(u)$

or

$\mathfrak{T}_P(u)=-\mathfrak{T}_Q(t)$ whenever both vectors exist and $Q(t)=P(u)$.

With this fact in hand, we can break the piecewise good parameterizations into two groups, just as before and define an ***oriented piecewise good curve*** *to be a piecewise good curve together with a selection of one of these two groups to provide a preferred direction for traversing the curve. This selection is called an* ***orientation for the curve****, and is often specified by giving a vector valued function $\mathfrak{T}$ along*

the curve $\mathfrak{C}$ *defined by* $\mathfrak{T}(P(t)) = \mathfrak{T}_P(t)$ *for a piecewise good parameterization* P *belonging to the orientation.*

The ***work or flow or circulation*** *of* F *along the oriented piecewise good curve is then defined, just as before, to be*

$$\int_a^b F(P(t)) \cdot \mathfrak{T}(P(t)) \ |P'(t)| \ dt = \int_a^b F(P) \cdot dP = \int_{\mathfrak{C}} F(s) \cdot \mathfrak{T}(s) \ ds.$$

Note that if you use a parameterization from the other orientation a minus sign is introduced into the second integral in this formula.

23.4. ***Exercise***. * ***Mr. Bacon's Train***[28] *We have a train track set up on the* XY *plane on the circle* $(X-1)^2 + Y^2 = 4$*, where distances are measured in kilometers. The entire track is covered by a solid loop of rail cars. Each car is identical and each has a flat thin board sticking up to act as a sail, with normal parallel to the length of the car. We presume the car body itself offers no resistance to the wind, and there is no friction between wheels and track. Suppose that a constant (over time) wind is blowing across the* XY *plane with velocity vector* $W(X,Y) = \langle Y, -X \rangle$ *kilometers per hour at each point* (X,Y) *on the track. This is an "inside out" tornado, where the wind far from the center is faster than that near the center. When the brakes are released, the wind pushes on the sails, trying to make some cars move clockwise and others counterclockwise, but since they are all attached they must move together.*

(i) When the train starts to move, will it turn clockwise or counterclockwise? (Hint: Let Q *be a differentiable parameterization of the track and let* ΔQ *be a vector representing a short piece of the track, pointing in the direction of the parameterization. The force caused by the wind on the cars on this piece is* $kW \cdot \Delta Q$*, where* k *is a constant of proportionality depending on the units we are using but not on the position of* ΔQ *along the track. If this number is positive, the force on this piece will try to move the train in the same direction as the parameterization.)*

(ii) The train will continue to speed up as time passes. Is there an "equilibrium" speed, at which it would have no tendency to move faster. What is this speed? What do the integrals in this problem have to do with work? (Hint 1: The apparent wind velocity to a passenger on a car moving with velocity V *is* $W - V$*. Hint 2: When the train is moving at constant speed, the velocity of a car at each point is a fixed multiple of the unit tangent vector at that point.)*

(iii) Suppose we want to increase the equilibrium speed of the train, so we put sensor devices and computers on each car that allow each sail to be rotated independently any way we want. How fast could the train's equilibrium speed get, and how would you program the onboard computers to approach that speed most efficiently? Could you make the train move in either direction?

---

24. Flux Past a Curve in Two Dimensions

We now consider an application in two dimensions. Suppose you are given an oriented good curve $\mathcal{C}$ in the plane with unit tangent selection $\mathcal{T}$ along $\mathcal{C}$. Suppose also that we are given a continuous vector field $F = \langle M, N\rangle$ defined on an open set containing the curve and a good parameterization $Q = \langle X, Y\rangle$ which agrees with the orientation.

The vector $\langle Y'(t), -X'(t)\rangle$ is perpendicular to $\mathcal{T}(Q(t))$ everywhere along the curve. Let $\mathcal{N}(Q(t))$ denote the

$$\textbf{Unit Normal for this Orientation:} \quad \mathcal{N}(Q(t)) = \frac{\langle Y'(t), -X'(t)\rangle}{|Q'(t)|}.$$

Unit normals corresponding to the opposite orientation point in the opposite direction.

The unit normal for this orientation is a vector valued function defined on the curve $\mathcal{C}$. Though $\mathcal{N}$ is initially defined using a parameterization, it depends only on $\mathcal{C}$ itself and the chosen orientation.

We define:

$$\begin{aligned}\textbf{Flux of F through } \mathcal{C}\text{:} \int_{t=a}^{t=b} & F(Q(t)) \cdot \mathcal{N}(Q(t))\ |Q'(t)|\ dt \\ &= \int_{t=a}^{t=b} F(Q(t)) \cdot \frac{\langle Y'(t), -X'(t)\rangle}{|Q'(t)|}\ |Q'(t)|\ dt \\ &= \int_{t=a}^{t=b} MY' - NX'\ dt\end{aligned}$$

It is the last integral one uses to calculate the flux.

One frequently sees the notation

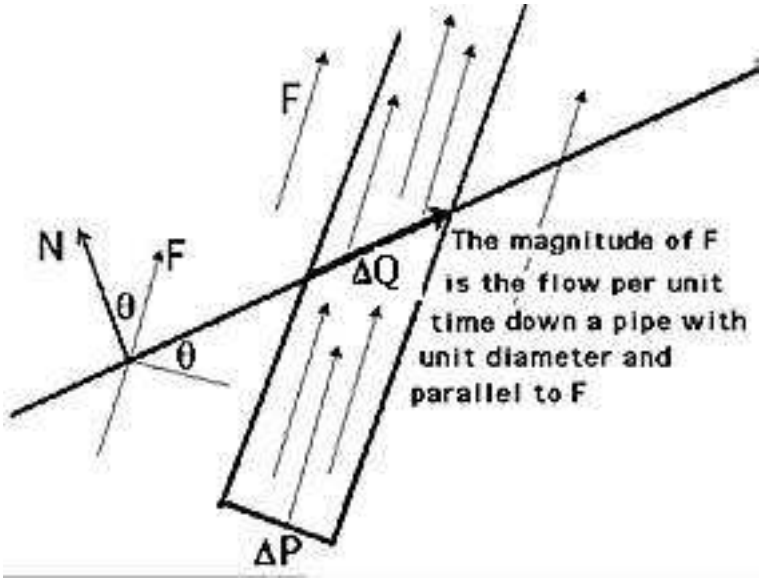
$$\int_{\mathcal{C}} F(s) \cdot \mathcal{N}(s)\ ds$$

to denote the flux of F past oriented $\mathcal{C}$. The orientation is built into $\mathcal{N}$.

Recall for comparison the flow integral along $\mathcal{C}$ with this orientation:

$$\int_{\mathcal{C}} F(s) \cdot \mathcal{T}(s)\ ds = \int_{t=a}^{t=b} F(Q(t)) \cdot \mathcal{T}(Q(t))\ |Q'(t)|\ dt = \int_{t=a}^{t=b} MX' + NY'\ dt.$$

To understand the meaning of flux consider the following diagram.



Here we imagine that F represents the direction of a flow of a layer of fluid on the surface of a plane. The magnitude of F represents the flow rate per unit time down a channel of unit width parallel to the field. We choose a piece of the curve so tiny that it appears straight and the field appears constant in the vicinity. Let Q be a parameterization consistent with the orientation.

In unit time, an amount $|F|\ \Delta P$ will flow past ΔQ, where ΔP is the width of a pipe parallel to F which crosses the curve at ΔQ.

But $|\Delta Q|\ cos(\theta) = \Delta P$ where $cos(\theta) = \dfrac{\mathcal{N} \cdot F}{|F|}$.

So the amount fluid moving past ΔQ per unit time is

$$|F|\ \Delta P = |F|\ cos(\theta)\ |\Delta Q| = |F|\ \frac{\mathcal{N} \cdot F}{|F|}\ |\Delta Q| = \mathcal{N} \cdot F\ |\Delta Q|.$$

Let Δt denote the change in parameter that causes ΔQ. Note: we often think of t as representing time, but it is not the same as the "flow per unit time" we talk about in the context of F. The parameter t is simply a set of labels that serves to identify the points on the curve. Q is differentiable with respect to t. With that in mind

Fluid Past ΔQ in unit time is $\mathcal{N} \cdot F\ |\Delta Q| = \ \mathcal{N} \cdot F\ \dfrac{|\Delta Q|}{\Delta t}\ \Delta t \approx \mathcal{N} \cdot F\ \left|\dfrac{dQ}{dt}\right|\ \Delta t.$

Adding these miniscule contributions over all the little pieces of the curve (i.e. forming a Riemann sum and converting it to an integral) gives the flux integral $\int_{t=a}^{t=b} \mathcal{N}(r(t)) \cdot F(Q(t))\ |Q'(t)|\ dt$, which we *interpret* to be the net amount of fluid crossing the whole curve in the indicated direction per unit time.

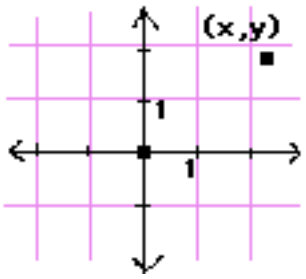
The flux integral is a number created from a vector field, a curve and an orientation. It is not the flow of any real fluid, just a calculation, a number. The argument from above gives one plausible *interpretation* of this number which can

be an aid to intuition and guide us to create one integral or another in an application. The practitioner must take care not to confound intuition with calculation - sometimes intuition will tell us that the calculation must be all wrong. Other times a surprising calculation makes us suspect that our intuition is leading us astray. Each illuminates the other. In this business you need both tools and you need to be aware that they *are different.*

24.1. ***Exercise***. *Convince yourself that the arguments given above for flux across an oriented good curve make sense for an oriented piecewise good curve.*

24.2. ***Exercise***. *Calculate the outward flux across the unit circle for the field given by* $F(X,Y) = \langle -Y, X \rangle$.

25. Calculus in Polar Coordinates



The usual XY coordinates for a point in the plane are a route description of how to get to a point from a designated center using a path that is parallel to the chosen coordinate axes. The coordinate pair (X, Y) is a list of directions: move from the origin along the X axis and then vertically to the point. The rectangular coordinate grid is an aid to finding your way to "the spot." Once you have decided on an origin and axes, each point has only one pair of XY coordinates.

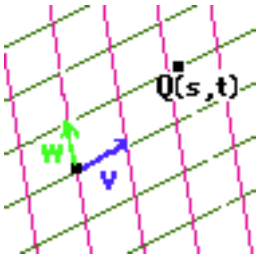
You might recall from Section 14 that we discussed an alternative naming scheme for points in the plane. If V and W are nonzero $2D$ vectors and V is not a multiple of W the function

$$Q(s,t) = sV + tW$$

takes ordered pairs of parameters s and t to distinct points in the plane using the vectors V and W.

For a pair (s,t) you get to $Q(s,t)$ by going from the origin to sV and then along a line containing W to $sV + tW$.

You might call this example a "parallelogram grid" coordinate system rather than a "rectangular grid" coordinate system. There is exactly one ordered pair (s,t) for each point in the plane.



In this section we consider a different kind of parameterization of the plane called **polar coordinates.** This parameterization also requires you to pick an origin and axes first and also uses ordered pairs of real numbers to describe a route to the points on the plane. The ordered pair is called polar coordinates for the point.

If (r, θ) is interpreted as the polar coordinates of a point, the point we mean is found as follows.

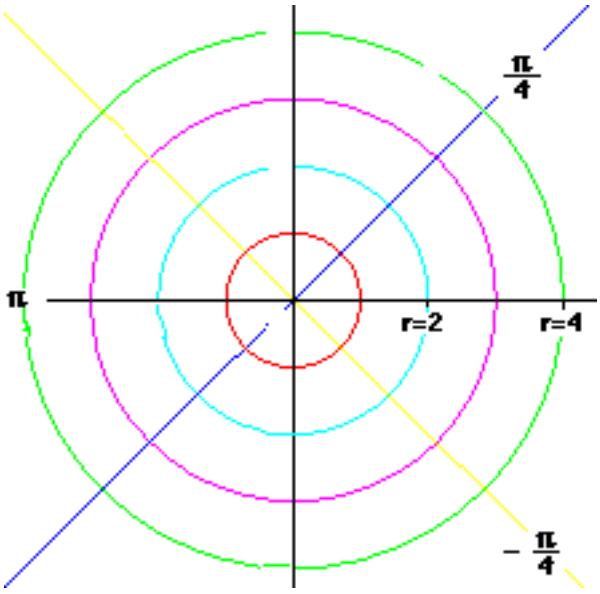
Stand at the origin with your nose pointing in the direction of the positive X axis. If θ is positive, rotate toward the positive Y axis by angle θ. If θ is negative rotate the other way by an angle $|\theta|$. Your nose is now pointing in the required direction. If r is positive walk forward a distance r and stop. You are there. If r is negative walk backwards a distance $|r|$ and stop. You are there.

In the coordinates from above each point in the plane corresponded to exactly one ordered pair—not so in polar coordinates! In polar coordinates you can add any integer multiple of 2π to the angle and you will wind up at the same place! Also if you add an odd multiple of π to the angle and replace r by $-r$ you will end up at the same place. For some purposes this is useful, for others it is annoying. In any case, it is a **feature** of polar coordinates.

The description of how to get to a spot using polar coordinates also is aided by a coordinate grid, but this grid is different. The "constant r" grid consists of circles centered at the origin, while the "constant θ" grid members are lines through the origin at various angles.

In our description of how to get to a point we first located the correct angle gridline and moved along it till we crossed the grid circle with the correct radius.

Polar coordinates are most useful when there is circular symmetry in the situation you are describing, or when considering paths that retrace themselves repeatedly. Orbiting satellites pop to mind as examples.



Using vectors, we can encapsulate most of the above discussion in:

$$\begin{array}{rcl} (r,\theta) & \longleftrightarrow & r\,\langle cos(\theta), sin(\theta)\rangle = \langle r\ cos(\theta), r\ sin(\theta)\rangle \\ \text{polar coordinates} & \longleftrightarrow & \text{position vector} \end{array}$$

This tells us how to convert from polar to rectangular coordinates. Converting from rectangular to polar coordinates is pretty easy too. If (X, Y) represents the rectangular coordinates of a point and $X > 0$ then $r = \sqrt{X^2 + Y^2}$ and $\theta = arctan\left(\frac{Y}{X}\right)$ will do the trick.

25.1. ***Exercise***. *How should you choose polar coordinates r and θ for the point with rectangular coordinates $(0, Y)$? How should you choose polar coordinates r and θ for the point with rectangular coordinates (X, Y) when X is negative?*

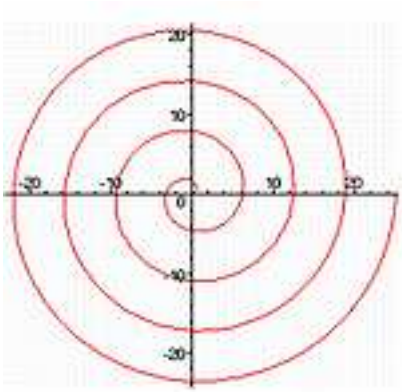
25.2. ***Exercise***. *Give three different polar coordinate pairs for the point with rectangular coordinates $(2\sqrt{3}\ ,\ 2)$. At least one should have positive r and one negative r. At least one should have positive θ and one negative θ.*

Convert polar coordinates $(-7, -135^o)$ to (exact) rectangular coordinates.

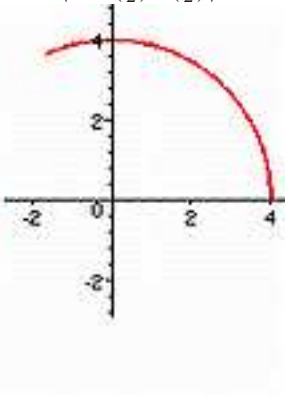
Curves are often described in polar coordinates by giving r and θ as functions of another parameter t. Other times θ itself is the parameter.

Find below sketches of several curves whose description involves polar coordinates in various ways. You should plot as many r and θ pairs as necessary to convince yourself that the stated equation yields the graph beneath.

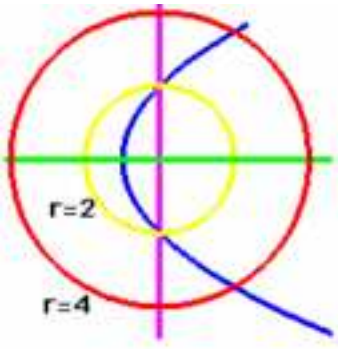
$r = \theta$ for θ in $[0, 8\pi]$
This can also be written as
$Q(\theta) = \theta \langle cos(\theta), sin(\theta) \rangle$.



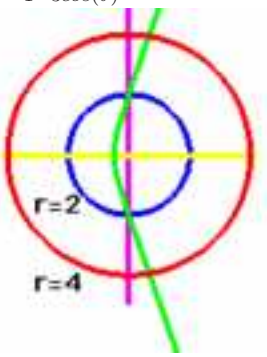
$r = 4$ and $\theta = t/2$ for t in $[0, 4]$
This can also be written as
$Q(t) = 4 \left\langle cos\left(\frac{t}{2}\right), \left(\frac{t}{2}\right) \right\rangle$.



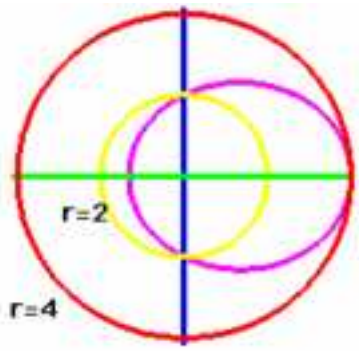
$r = \frac{1}{1-cos(\theta)}$ for θ in $[1, 5.5]$.



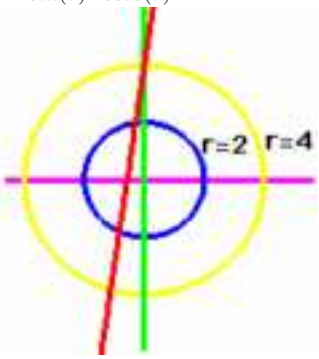
$r = \frac{1}{1-3cos(\theta)}$ for θ in $[1.3,\ 5]$.



$r = \frac{2}{2-cos(\theta)}$ for θ in $[0, 7]$.



$r = \frac{2}{sin(\theta)-7cos(\theta)}$ for θ in $[-1.65, 1.35]$.



The curve you see above corresponding to $r = \dfrac{1}{1 - cos(\theta)}$ looks a lot like a parabola. If you follow the calculation below you might recognize the last line as the equation for a parabola with vertex at the point with rectangular coordinates $(\frac{-1}{2}, 0)$.

$$
\begin{aligned}
& r = \frac{1}{1 - cos(\theta)} \\
& r(1 - cos(\theta)) = 1 \\
& r - X = 1 && \text{because } X = r\ cos(\theta) \\
& r^2 = (X + 1)^2 \\
& X^2 + Y^2 = X^2 + 2X + 1 && \text{because } X^2 + Y^2 = r^2 \\
& Y^2 = 2\left(X + \frac{1}{2}\right)
\end{aligned}
$$

25.3. ***Exercise***. ** Satisfy yourself that the equations for the last three graphics above yield equations in rectangular coordinates which are recognizable as what they appear to be: equations for a hyperbola, an ellipse and a line.*

There are a couple of vectors which will pop up repeatedly when working with parametric equations of the form $Q = r\,\langle cos(\theta),\ sin(\theta)\rangle$. One is the direction

vector at angle θ, $U_\theta = \langle cos(\theta),\ sin(\theta)\rangle$. It points in the direction of increasing r for constant θ. The other is the vector $\langle -sin(\theta),\ cos(\theta)\rangle$ which we denote V_θ. It points in the direction of increasing θ for constant r, and is a unit tangent vector to the circle at that radius centered at the origin. These vectors are not defined at the origin. Note that:

$$\frac{d}{d\theta}U_\theta = V_\theta \quad \text{and} \quad \frac{d}{d\theta}V_\theta = -U_\theta \quad \text{and} \quad U_\theta \cdot V_\theta = 0.$$

If θ is a function of t, a rather common situation, we have:

$$\frac{d}{dt}U_\theta = \theta' V_\theta \quad \text{and} \quad \frac{d}{dt}V_\theta = -\theta' U_\theta.$$

Suppose that a curve $Q(t)$ is given as

$$Q(t) = r(t)\,\langle cos(\theta(t)),\ sin(\theta(t))\rangle = r(t)U_{\theta(t)}.$$

$U_{\theta(t)}$ is called the **radial direction vector** at the point $Q(t)$ on the curve and $V_{\theta(t)}$ is called the **tangential direction vector** at that point.

Taking derivatives with respect to t we have:

$$\begin{aligned} Q' &= r'U_\theta + r\theta' V_\theta \\ \text{and } Q'' &= r''U_\theta + r'\theta' V_\theta + r'\theta' V_\theta + r\theta'' V_\theta - r(\theta')^2 U_\theta \\ &= (r'' - r(\theta')^2)U_\theta + (r\theta'' + 2r'\theta')V_\theta. \end{aligned}$$

The two terms in each derivative are called the **radial and tangential components** of the velocity and acceleration vectors.

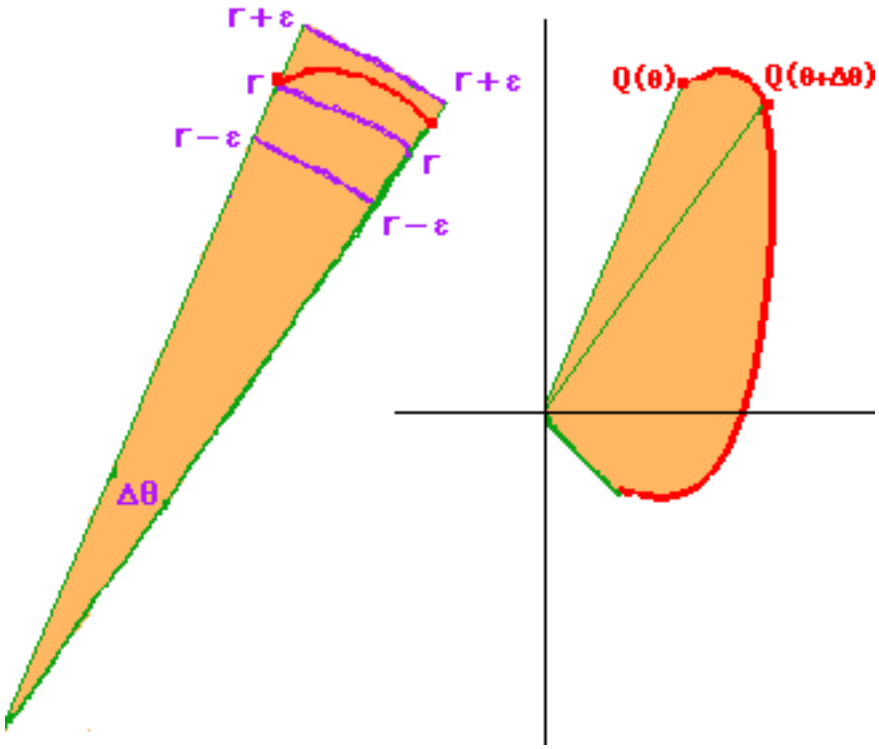
In case $\theta = t$, so the curve is parameterized by angle as frequently happens, we have the simpler looking equations:

$$\begin{aligned} Q' &= r'U_\theta + rV_\theta \\ \text{and } Q'' &= (r'' - r)U_\theta + 2r'V_\theta. \end{aligned}$$

25.4. ***Exercise***. *Show that the speed of a parametric motion $Q(t)$ as above is given by $\sqrt{(r')^2 + (r\theta')^2}$. In case $\theta = t$ the speed is simply $\sqrt{\left(\frac{dr}{d\theta}\right)^2 + r^2}$.*

In addition to the usual things you can do with derivatives and integrals involving parameterizations, some calculations take advantage of the special nature of polar coordinates.

For example in the picture below on the right we want to find a number which we can interpret as the area of the pie shaped region inside the pie with "crust" at the curve with continuous parameterization $Q(\theta) = r(\theta)U_\theta$ where $\alpha \le \theta \le \beta$. We will add up many "pie slices" corresponding to small increments $\Delta\theta$ of the parameter θ. For convenience we will consider positive r and insist that the minimum value of r on $[\alpha, \beta]$ is greater than 0, as in the picture. We also presume that the interval is not longer than 2π so the curve does not wind around the origin more than once.



Let M denote the maximum value of r on the interval and suppose ε is a positive number. Since Q is continuous on $[\alpha, \beta]$ we can find δ so small that whenever $\Delta\theta < \delta$ and both θ and $\theta + \Delta\theta$ are in $[\alpha, \beta]$ then $|r(\theta + \Delta\theta) - r(\theta)| < \varepsilon$.

For any such $\Delta\theta$ the pie slice (on the left above) from Q down to the origin is entirely inside the circle sector of radius $r+\varepsilon$ and, when $r-\varepsilon > 0$, entirely contains the circle sector of radius $r - \varepsilon$.

So if A is the area inside the curve on the left,

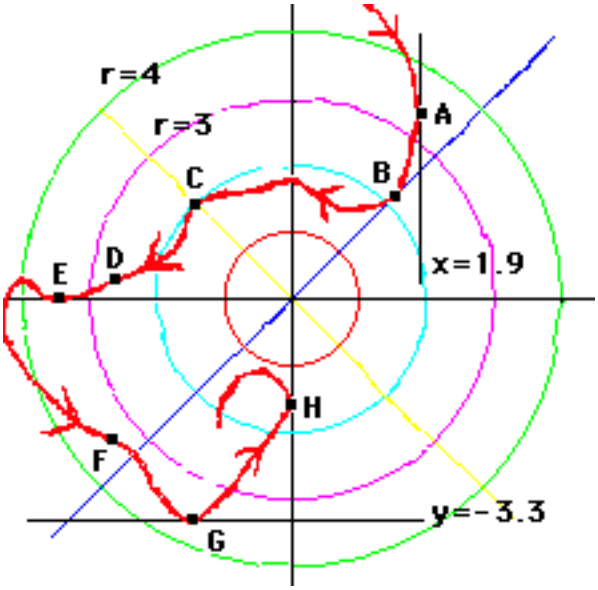
$$\left| A - \pi r^2 \frac{\Delta\theta}{2\pi} \right| < \pi(r+\varepsilon)^2 \frac{\Delta\theta}{2\pi} - \pi(r-\varepsilon)^2 \frac{\Delta\theta}{2\pi} = 4\pi r\varepsilon \frac{\Delta\theta}{2\pi} \leq 2M\varepsilon\Delta\theta.$$

Terms like $\pi r^2 \frac{\Delta\theta}{2\pi} = \frac{r^2\Delta\theta}{2}$ can be used to form a Riemann sum $\sum_{i=1}^{N} \frac{r_i^2 \Delta\theta_i}{2}$ for the whole area on the right. For partitions with mesh less than δ this sum cannot be more than $\sum_{i=1}^{N} 2M\varepsilon\Delta\theta_i = 2M\varepsilon(\beta - \alpha)$ away from any sensible definition of area for the whole pie shaped region, and since ε can be chosen to be as small as we like we have a representation for the area as:

$$\textbf{Area in the Pie Shape} = \int_{\alpha}^{\beta} \frac{1}{2} r^2 \, d\theta.$$

25.5. ***Exercise***. ** How do you modify the discussion of area if r can be negative? If r can be 0? How does the integral look if both r and θ are functions of a parameter t?*

Finally we get to an exercise that will help us think about the signs of various derivatives and how they are related to each other on a graphic found below. We will presume that the parameterization of the curve moves in the direction of the arrow differentiably and never slows to a stop.



At point A the curve seems to be going straight down. At this spot this is in the direction of decreasing angle and passing through circles of diminishing radius. So $\frac{dX}{dt} = 0$, $\frac{dY}{dt} < 0$, $\frac{dr}{dt} < 0$ and $\frac{d\theta}{dt} < 0$.

As you move on the curve through the point A in the direction of increasing θ, r is increasing, Y is decreasing and X is not changing. So $\frac{dr}{d\theta} > 0$, $\frac{dY}{d\theta} < 0$ and $\frac{dX}{d\theta} = 0$.

As you pass through A on the curve in the direction of increasing r, which is opposite to the direction of increasing t, θ is increasing, Y is increasing and X is not changing. So $\frac{d\theta}{dr} > 0$, $\frac{dY}{dr} > 0$ and $\frac{dX}{dr} = 0$.

Derivatives with respect to X will not exist as you pass through the point A on the curve.

Obviously there is considerable redundancy in the sign pattern and an opportunity for consistency checks through the chain rule. For example $\frac{d\theta}{dr} = \frac{d\theta}{dY}\,\frac{dY}{dr}$.

25.6. ***Exercise***. *Making reasonable assumptions about the shape of the curve, fill in the sign chart for the various derivatives at the other points on the graph. Check for consistency via the chain rule. Why are there question marks in two of the boxes?*

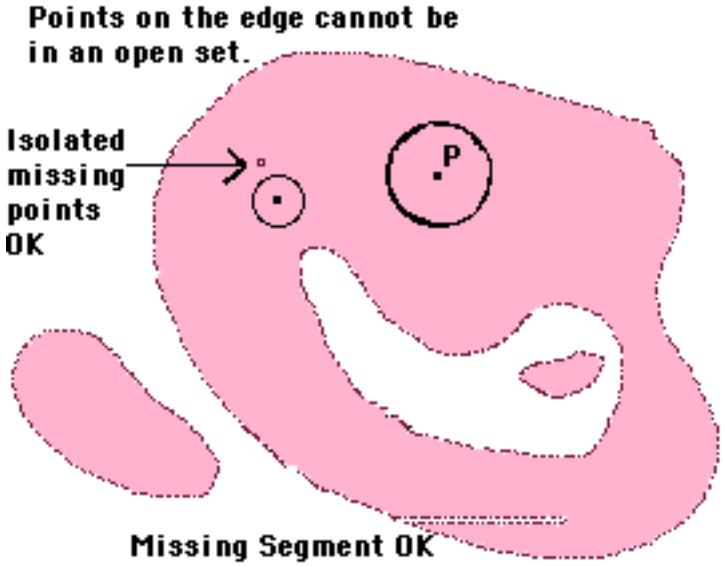
Point	$\frac{dX}{dt}$	$\frac{dY}{dt}$	$\frac{dr}{dt}$	$\frac{d\theta}{dt}$	$\frac{dY}{dX}$	$\frac{dr}{d\theta}$	$\frac{dX}{d\theta}$	$\frac{dY}{d\theta}$	$\frac{dX}{dr}$	$\frac{dY}{dr}$
A	*0*	$-$	$-$	$-$	*DNE*	$+$	*0*	$+$	*0*	$+$
B				*0*						
C			*0*							
D										
E		*0*		*0*				*?*		
F										
G		*0*								
H	*0*			*0*			*?*			

CHAPTER IV

The Gradient May 27, 2005

26. Functions of Two Variables: Continuity

Suppose $\mathfrak{O}$ is a set in $\mathbb{R}^2$. A set such as $\mathfrak{O}$ is called **open** if for each P in $\mathfrak{O}$ there is some disk (possibly tiny) centered at P entirely inside $\mathfrak{O}$. In other words, $\mathfrak{O}$ does not contain any points on an "edge." This concept generalizes the idea of an open set of real numbers.[29]



The **complement of a set $\mathfrak{O}$ in** $\mathbb{R}^2$ is the collection of all points in $\mathbb{R}^2$ which are **not** in $\mathfrak{O}$.

A set $\mathfrak{O}$ is called **closed** if its complement is open: that is, the set of points in the plane **not** in $\mathfrak{O}$ constitute an open set.

In this section we will be interested in real valued functions defined for points in the plane. These are called **real functions of two variables.** Many of the surfaces we examined in Section 13 were the graphs of functions of this type. Usually the domain of our functions will be an open set.

Suppose g is such a function and $A = (P_1, P_2)$ is a point in its domain, located at the tip of the vector $P = \langle P_1, P_2 \rangle$. We will be moving around in the domain of g using vector operations, naming the X and Y coordinates of domain members and so forth. We want a notational convention that will make working with g less cumbersome as we do this so we refer to the value of g at the point A by any of the following: $g(A)$ or $g(P)$ or $g(P_1, P_2)$. We will prefer these to the more correct but ugly $g((P_1, P_2))$ or to $g(\langle P_1, P_2 \rangle)$. This usage will be extended in later sections to functions of three or more variables.

If g is a function defined for points in an open set $\mathfrak{O}$ and P is in $\mathfrak{O}$ we write $\lim_{Q \to P} g(Q) = L$ if and only if for each $\varepsilon > 0$ there is some $\delta > 0$ such that if $0 < |Q - P| < \delta$ then $|g(Q) - L| < \varepsilon$.

In words: You can force $g(Q)$ to be as close to L as you wish by requiring Q to be close enough to P (P itself excluded.) L is called the **limit of** g **at** P.

Letting $\Delta P = Q - P$ we can see that $\lim_{\Delta P \to 0} g(P + \Delta P) = L$ is identical in meaning to $\lim_{Q \to P} g(Q) = L$, and sometimes it is more useful to think of limits using the second notation.

26.1. ***Exercise***. *Let $\Delta P = \langle \Delta X, \Delta Y \rangle$ and define $S_{\Delta P}$ to be $|\Delta X| + |\Delta Y|$ and let $M_{\Delta P}$ be the larger of $|\Delta X|$ or $|\Delta Y|$.*

Show that:

$$M_{\Delta P} \leq |\Delta P| \leq S_{\Delta P} \leq 2M_{\Delta P}.$$

26.2. ***Exercise***. *We could rephrase the definition of limit found above to any of the following equivalent conditions:*

For any $\varepsilon > 0$ there is a $\delta > 0$ so that

(i) if $0 < |\Delta P| < \delta$ then $|g(P + \Delta P) - L| < \varepsilon$.

(ii) if $0 < M_{\Delta P} < \delta$ then $|g(P + \Delta P) - L| < \varepsilon$.

(iii) if $0 < S_{\Delta P} < \delta$ then $|g(P + \Delta P) - L| < \varepsilon$.

(iv) if $0 < r < \delta$ then for any angle θ, $|g(P + rU_\theta) - L| < \varepsilon$.

A function such as g is called **continuous at** P if $\lim_{Q \to P} g(Q)$ exists and is $g(P)$. g is called **continuous on** $\mathfrak{O}$ if it is continuous at every point in $\mathfrak{O}$.

From this point, when we use the word **surface** we will mean the graph of a continuous function g defined on an open set in the plane. Most often we will represent this surface as a collection of points (X, Y, Z) in space with $Z = g(X, Y)$, though from time to time surfaces formed as $Y = g(X, Z)$ or $X = g(Y, Z)$ will be encountered.

Continuity is harder to understand in higher dimensions, and one reason to introduce $\varepsilon - \delta$ proofs early is to ease the transition when you get to this setting, where they (or something equivalent) are required.

Here is a function that provides an interesting example:

$$g(X, Y) = \begin{cases} \frac{XY}{X-Y}, & \text{if } X \neq Y; \\ 0, & \text{if } X = Y. \end{cases}$$

The vertical slice through the graph of this function by the plane $Y = mX$ (with $m \neq 1$) is the line $Q(X) = \left\langle X, mX, \frac{m}{1-m} X \right\rangle$. This line, parameterized by X, is continuous and passes through $(0, 0, 0)$. Also g is constantly zero on the X and Y axes and on the line $Y = X$ in the XY plane: that is g is continuous when you look only at its values on **any particular line** through the origin in its domain.

From this one might conclude that g is continuous at the origin, but it is not. That is because no matter how tiny δ might be, you can find a slope m near 1 for

which $\frac{m}{1-m}\delta > 1$. In other words, if ε is a positive number less than 1 we can find a point (X, mX) arbitrarily close to $(0,0)$ for which $g(X, mX) > \varepsilon$. So g is not continuous at $(0,0)$.

---

27. Functions of Two Variables: Differentiability

We suppose g is a real valued function of two variables as in the last section and defined on an open set $\mathfrak{O}$. We will use the notation of the last section with $\Delta g = g(P + \Delta P) - g(P)$.

g is called **differentiable at** P in $\mathfrak{O}$ if there is a vector A so that

$$\lim_{\Delta P \to 0} \frac{\Delta g - A \cdot \Delta P}{|\Delta P|} = 0.$$

When this limit exists and is 0 the vector A is unique: that is, there can be no other such vector, for if $\lim_{\Delta P \to 0} \frac{\Delta g - B \cdot \Delta P}{|\Delta P|}$ also exists and equals 0 for $A \neq B$ then we could let ΔP be a tiny positive multiple of $A - B$. So

$$\begin{aligned} &\frac{(A-B)\cdot \Delta P}{|\Delta P|} \\ &\qquad = \frac{\Delta g - B\cdot\Delta P - \Delta g + A\cdot\Delta P}{|\Delta P|} \\ &\qquad = \frac{\Delta g - B\cdot\Delta P}{|\Delta P|} - \frac{\Delta g - A\cdot\Delta P}{|\Delta P|}. \end{aligned}$$

The first line would be a positive constant while the last line must converge to 0 as ΔP becomes smaller. This contradictory calculation shows that $A = B$.

When the vector A as above exists we call it the **gradient of** g **at** P. This vector is denoted (usually) in these notes $\nabla \mathbf{g}(\mathbf{P})$. It is also common to see **grad g(P)** used to denote the gradient.

The function g is continuous wherever the gradient exists: the existence of the gradient implies that the numerator in the limit from its definition must converge to 0. Since $\lim_{\Delta P \to 0} A \cdot \Delta P = 0$ we must have $\lim_{\Delta P \to 0} \Delta g = 0$ also, which is the requirement for continuity.

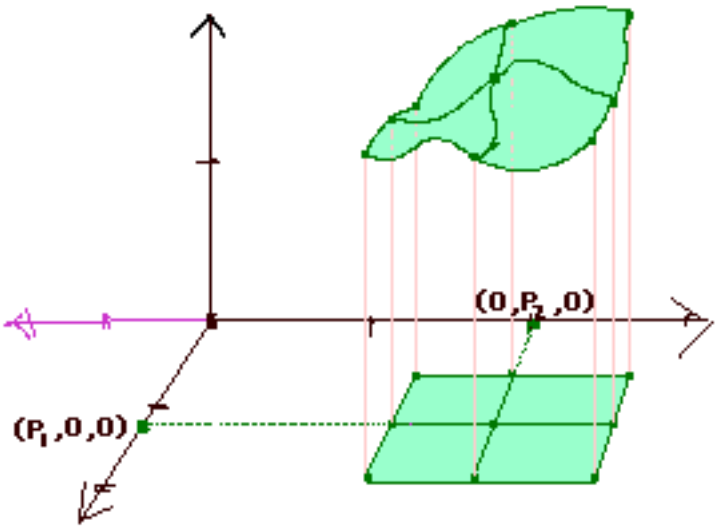
The entries of the gradient vector A, when it exists, have their own notation and meaning. For $A = \langle A_1, A_2 \rangle$ let's calculate the limit using ΔP of the form $\langle \Delta X, 0 \rangle$. So

$$\begin{aligned} 0 &= \lim_{\Delta X \to 0} \frac{g(P + \Delta X\ \vec{i}\,) - g(P) - A_1 \Delta X}{|\Delta X|} \\ &= \lim_{\Delta X \to 0} \frac{\Delta X}{|\Delta X|} \left(\frac{g(P + \Delta X\ \vec{i}\,) - g(P)}{\Delta X} - A_1 \right). \end{aligned}$$

This means that

$$A_1 = \lim_{\Delta X \to 0} \frac{g(P + \Delta X\ \vec{i}\,) - g(P)}{\Delta X}.$$

This number is therefore an ordinary derivative of the real valued function $f(X) = g(X, P_2)$ at $X = P_1$. If you prefer, you can form the vector function $h(X) = \langle X, P_2, g(X, P_2)\rangle$ parameterized by X in an interval around P_1. $h'(P_1) = \langle 1, 0, A_1\rangle$. The picture you should retain from this is the following: The graph of g is a surface above or below the XY plane. If you slice through this surface with a plane $Y = P_2$ the curve formed by the surface at the cut looks like an ordinary function from first year calculus. A_1 is the derivative of this curve at $X = P_1$.



You can cut through the graph of g with the plane $X = P_1$ to obtain a similar interpretation of A_2 as the "X constant, allow Y to vary" derivative.

These numbers are called the **partial derivatives of g at P**. There are several notations in common use for A_1 and A_2. Among them are:

$$A_1 = D_1 g(P) = D_X g(P) = \frac{\partial g}{\partial X}(P) = \left.\frac{\partial g}{\partial X}\right|_{\langle X,Y\rangle=\langle P_1,P_2\rangle}$$

and

$$A_2 = D_2 g(P) = D_Y g(P) = \frac{\partial g}{\partial Y}(P) = \left.\frac{\partial g}{\partial Y}\right|_{\langle X,Y\rangle=\langle P_1,P_2\rangle}.$$

We will use $D_1 g(P)$ and $D_2 g(P)$ to denote the partial derivatives, which avoids explicit mention of the idiosyncratic choice of axis names.

We have shown that, when it exists the gradient is

$$\nabla g(P) = \langle D_1 g(P),\ D_2 g(P)\rangle\,.$$

Now is the time for us to define **vector valued functions in the plane.** Sometimes they are also called **vector fields.** They assign a vector to each point in their domain, which will now be a subset of the plane. A vector valued function defined on an open set such as $\mathfrak{O}$ is called **continuous** at P if its coordinate functions are continuous at P. It is called continuous on $\mathfrak{O}$ if it is continuous at each point of $\mathfrak{O}$. The reason to put this definition here is because we have just created an example of a vector valued function.

With g as above we define ∇g to be the function whose value is the gradient of g at P for every P in $\mathfrak{O}$ for which the gradient exists. ∇g is an example of a vector valued function defined for points in the plane, and is defined here for some (or perhaps all) of the points in $\mathfrak{O}$. $\nabla g = \langle D_1 g,\ D_2 g \rangle$.

It is not true that the existence of the partial derivatives implies that the gradient exists. The example from Section 26 provides a counterexample. We saw that the function g defined there was not continuous at the origin, a requirement for differentiability. But $D_1 g(0) = D_2 g(0) = 0$.

However, if g is a function defined around a point P and if both $D_1 g$ and $D_2 g$ are defined and continuous on any open set containing P then g **is** differentiable at P. This follows from the Mean Value Theorem, and the proof is found below.

Suppose that both $D_1 g$ and $D_2 g$ are defined and continuous on a disk of radius δ centered at P. Suppose $\Delta P = \langle \Delta X, \Delta Y \rangle$ and $|\Delta P| < \delta$ so $P + \Delta P$ is in this disk.

Note that

$$\begin{aligned} g(P + \Delta P) - g(P) & \\ &= g(P + \Delta X\ \vec{i}\,) - g(P) \\ &\quad + g(P + \Delta X\ \vec{i} + \Delta Y\ \vec{j}\,) - g(P + \Delta X\ \vec{i}\,). \end{aligned}$$

Since the partial derivatives exist and are continuous in this disk the Mean Value Theorem says that there are numbers α between 0 and ΔX and β between 0 and ΔY so that

$$\frac{g(P + \Delta X\ \vec{i} + \Delta Y\ \vec{j}\,) - g(P + \Delta X\ \vec{i}\,)}{\Delta Y} = D_2 g(P + \Delta X \vec{i}\ + \beta \vec{j}\,)$$

$$\text{and} \qquad \frac{g(P + \Delta X\ \vec{i}\,) - g(P)}{\Delta X} = D_1 g(P + \alpha \vec{i}\,).$$

This implies that

$$\begin{aligned} &\frac{|g(P + \Delta P) - g(P) - D_1 g(P)\Delta X - D_2 g(P)\Delta Y|}{|\Delta P|} \\ &= \frac{\left| \left(D_1 g(P + \alpha \vec{i}\,) - D_1 g(P) \right) \Delta X + \left(D_2 g(P + \Delta X \vec{i}\ + \beta \vec{j}\,) - D_2 g(P) \right) \Delta Y \right|}{|\Delta P|} \\ &\leq |D_1 g(P + \alpha \vec{i}\,) - D_1 g(P)| + |D_2 g(P + \Delta X \vec{i}\ + \beta \vec{j}\,) - D_2 g(P)|. \end{aligned}$$

Both of the last terms converge to 0 along with $|\Delta P|$ so g is differentiable at P.

We finish this section with a final definition.

When $D_1 g$ or $D_2 g$ exist it is possible that these functions might, themselves, be differentiable. The notation $D_{i,j} g = D_j(D_i g)$ is used for these **second partial derivatives**. When $i \neq j$ they are called **mixed partial derivatives.** It often happens that the mixed partials $D_{i,j} g$ and $D_{j,i} g$ are equal, and we note a condition under which this happy state pertains in Exercise 33.4.

One can keep going with this and, when the functions involved are nice enough, calculate **third** and **higher order** partial derivatives such as

$$D_{1,1,2,1}g = D_1(D_2(D_1(D_1g))).$$

So, for example, if $g(X,Y) = X^3 sin(Y)$ then $D_{1,1,2,1}g(X,Y) = 6cos(Y)$. The **order** of a partial derivative is the number of "differentiations" required to calculate it. So this partial derivative is mixed and of order 4.

Partial derivatives come up in many of the same situations that ordinary derivatives do from beginning Calculus.

---

28. The Chain Rule and The Tangent Plane

When $\nabla g(P)$ exists we define

$$L_{g,P}(Q) = g(P) + \nabla g(P) \cdot (Q - P).$$

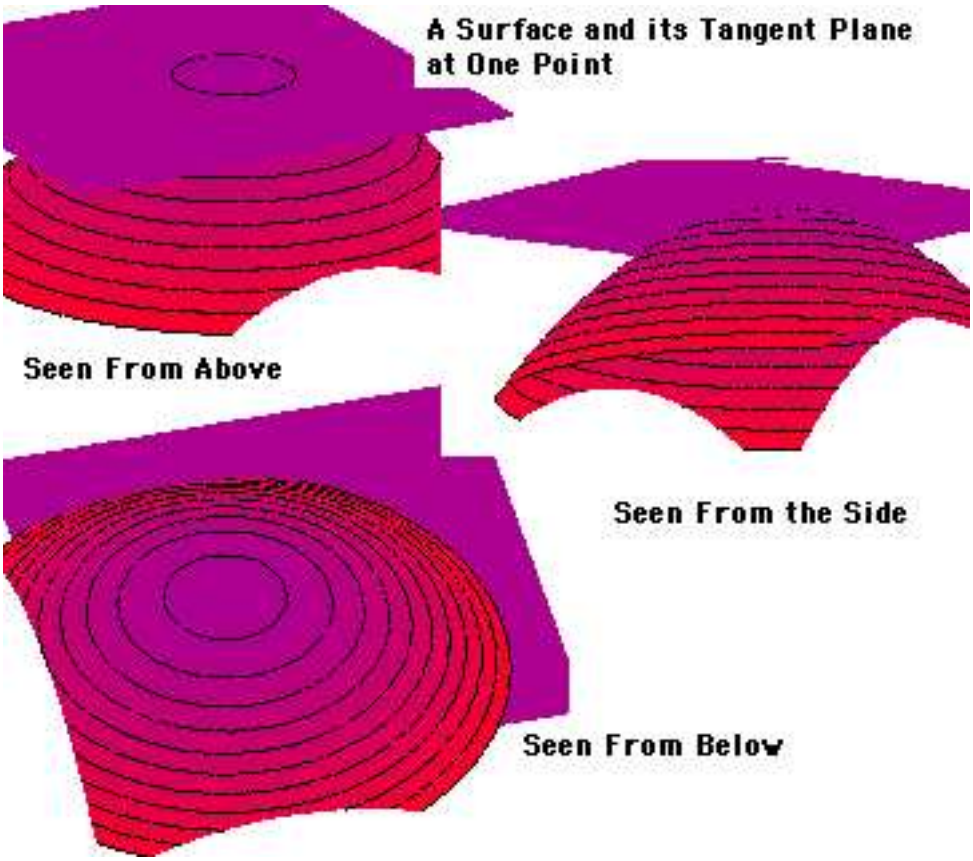
$L_{g,P}$ is called the **linearization of** g **at** P. For such P we have

$$\lim_{Q \to P} \frac{g(Q) - L_{g,P}(Q)}{|Q - P|} = 0.$$

So when Q is close to P, not only is $g(Q)$ near to $L_{g,P}(Q)$ but the difference between them is small **even in comparison to** $|Q - P| = |\Delta P|$. So the graph of $Z = L_{g,P}(Q)$ is glued to the graph of $Z = g(Q)$ near P and provides a good approximation to g in that vicinity. The graph of $L_{g,P}$ is a plane with normal form $(\langle X, Y, Z\rangle - \langle P_1, P_2, g(P)\rangle) \cdot \langle -D_1g(P), -D_2g(P), 1\rangle = 0$.

This plane is called the **tangent plane to this surface at** $(P_1, P_2, g(P))$. The vector $\langle -D_1g(P), -D_2g(P), 1\rangle$ is perpendicular to this plane, as is any nonzero multiple of this vector. Because the plane and the surface are so closely associated we say that a nonzero multiple of this vector is normal to the surface at $(P_1, P_2, g(P))$ as well.

Let's suppose that the gradient exists in the vicinity of P and that $Q(t) = \langle X(t), Y(t)\rangle$ is a differentiable parameterization of a curve in the plane with nonzero derivative and $Q(c) = P$. Then $H(t) = \langle X(t), Y(t), g \circ Q(t)\rangle$ is a parameterization of a curve above or below the track of $\langle X(t), Y(t), 0\rangle$ in the XY plane and wanders around on the surface formed as the graph of g. We will say in a situation like this that the **curve is in the surface**. Our curve shares the point $(P_1, P_2, g(P))$ with both the tangent plane and the graph of g.



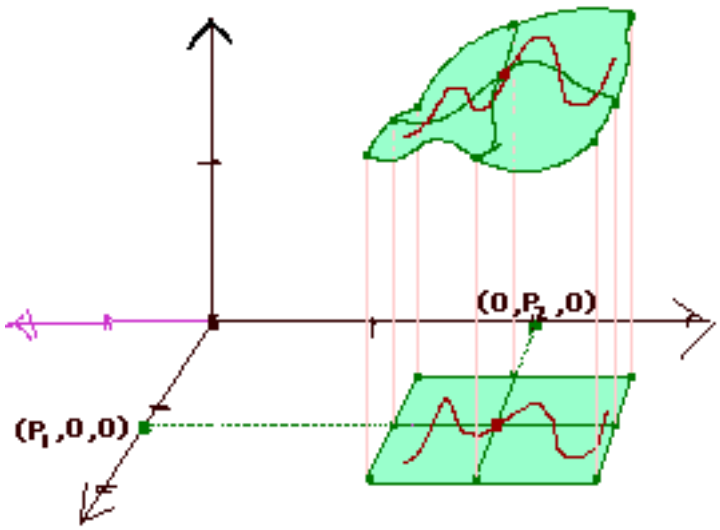
What is the derivative of H at c? It is obvious that the first and second components of the derivative are the same as for Q. As for the third, we define $\Delta P = Q(c+h) - Q(c) = \langle \Delta X, \Delta Y \rangle$ and note that because Q is continuous at c then $\lim_{h\to 0} \Delta P = 0$. Since g is differentiable, this implies that

$$\lim_{h\to 0} \frac{g(P+\Delta P) - g(P) - \nabla g(P)\cdot \Delta P}{|\Delta P|} = 0.$$

Rewriting this we have

$$\lim_{h\to 0} \left|\frac{\Delta P}{h}\right|^{-1} \left[\frac{g(P+\Delta P)-g(P)}{h} - \left(D_1 g(P)\frac{\Delta X}{h} + D_2 g(P)\frac{\Delta Y}{h}\right)\right] = 0.$$

By assumption, the factor on the left converges to a nonzero number and the terms in the inner parentheses on the right converge to $D_1 g(P)X'(c) + D_2 g(P)Y'(c)$ so we have discovered that $\lim_{h\to 0} \frac{g(P+\Delta P)-g(P)}{h} = \frac{d}{dt}(g\circ Q)(c)$ exists and equals $\nabla g(Q(c))\cdot Q'(c)$. We have also found that $H'(c) = \langle X'(c), Y'(c), \nabla g(Q(c))\cdot Q'(c)\rangle$.



28.1. ***Exercise***. *With conditions as above except that* $Q'(c) = 0$ *show that* $\frac{d}{dt}(g \circ Q)(c) = \nabla g(Q(c)) \cdot Q'(c) = 0$ *and also* H *is differentiable at* c *and* $H'(c) = 0 = \langle X'(c), Y'(c), \nabla g(Q(c)) \cdot Q'(c) \rangle$ *in this case too.*

We have just proved a version of **The Chain Rule:**

$$\frac{d}{dt}(g \circ Q)(c) = \nabla g(Q(c)) \cdot Q'(c).$$

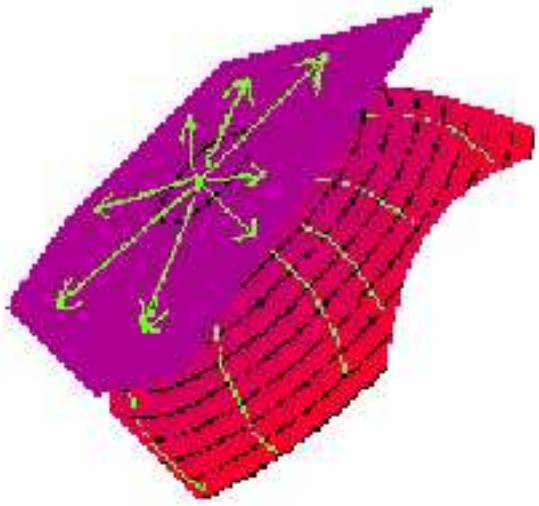
28.2. ***Exercise***. *We can pull many interesting facts out of the calculations found above. Suppose* g *is differentiable at* $P = \langle P_1, P_2 \rangle$

(i) If $H(t) = \langle X(t), Y(t), Z(t) \rangle$ *is* ***any*** *differentiable parameterization of a curve in the surface which passes through* $(P_1, P_2, g(P))$ *at time* c *then* $Q(t) = \langle X(t), Y(t) \rangle$ *is a differentiable parameterization of a curve in the plane and* $Q(c) = P$. *Also,* $Z(t) = g(X(t), Y(t))$. *So* ***any*** *differentiable curve in the surface is of exactly the type we have just considered.*

(ii) If H *is any differentiable parameterization of a curve in the surface and passes through* $(P_1, P_2, g(P))$ *at time* c *then* $H'(c)$ *is a vector that lies in the tangent plane to the surface at* $(P_1, P_2, g(P))$. *(hint: Dot* $H'(c)$ *against the normal* $\langle D_1 g(P), D_2 g(P), -1 \rangle$ *to the tangent plane.)*

(iii) If $V = \langle v_1, v_2, v_3 \rangle$ *is any vector that that lies in the tangent plane to the surface at* $(P_1, P_2, g(P))$ *then there is a differentiable parameterization* H *of a curve in the surface that passes through* $(P_1, P_2, g(P))$ *at time* 0 *and for which* $H'(0) = V$. *(hint: Let* $Q(t) = \langle P_1, P_2 \rangle + t \langle v_1, v_2 \rangle$.*)*

In the exercise above you proved that **the vectors in the tangent plane at a point consist of all tangent vectors to differentiable parameterizations of curves in the surface through that point, and to no other vectors.**



Here is another application of the chain rule. Recall that when the two vectors $\nabla g(Q(c))$ and $Q'(c)$ are nonzero

$$\nabla g(Q(c)) \cdot Q'(c) = |\nabla g(Q(c))||Q'(c)|cos(\theta)$$

where θ is the angle between $\nabla g(Q(c))$ and $Q'(c)$.

Examining this formula we see that among all differentiable curves Q in the plane which pass through $Q(c)$ **with a given speed**, the rate at which the function $g \circ Q$ is increasing or decreasing depends only on the angle θ. If $Q'(c)$ points in the same direction as $\nabla g(Q(c))$ then $(g \circ Q)(t)$ is increasing at maximum possible rate. Up on the graph of g, if you head in this direction you are going straight up hill. The opposite direction is straight down hill on the surface. If $cos(\theta) = 0$, so θ is $\frac{\pi}{2}$, you are heading straight across the side of the hill, neither rising nor falling.

We define the **directional derivative for the vector** V at a point P in the domain of g to be $\nabla \mathbf{g}(\mathbf{P}) \cdot \mathbf{V}$. The notation $\nabla_{\mathbf{V}} \mathbf{g}(\mathbf{P})$ is used for this number. It represents the rate at which you are rising or falling on the graph of g if you are moving along any parameterized curve passing through P with velocity V. In particular, if V is a unit vector this is how fast you are rising or falling up on the surface if your shadow in the XY plane is moving at unit speed in direction corresponding to V as it passes through P in the XY plane.

28.3. ***Exercise***. *You are a* ***Weird Alien*** *swimming in a layer of scum on a volcano and neutron star ravaged cess-pool, hereafter known as* $\mathbb{R}^2$. *The temperature*

at each point is given by $T(X,Y) = \frac{\sqrt{2}}{4}X^2 - \frac{\sqrt{2}}{2}Y$*, where distances are measured in meters and temperatures in degrees Celsius.*

Due to your need to pump nutrients past your dorsal cilia, you are doomed to swim at exactly 20 *meters per second so long as you live.*

Also, you cannot stand temperature changes greater than 10 *degrees per second, but since you mindlessly crave heat you will try to go toward hotter scum up to that limiting rate.*

Find a velocity vector your primitive instincts prompt you to use as you pass through the point $(1,2)$*.*

28.4. ***Exercise***. ***Susan's Hill:*** *Susan is hiking around on a smooth spherical hill above the* XY *plane which is the graph of* $g(X,Y) = -75 + \sqrt{100^2 - X^2 - Y^2}$ *where distances are in meters. She is trying to get to the top as fast as possible, but doesn't want to climb at an angle more than* 30^o *above horizontal. When she is at* $(60,0)$ *she finds that going straight toward the top is too steep. Give a direction vector in the* XY *plane to describe her best heading. (There are two possible answers.)*

28.5. ***Exercise***. *Suppose you have a block of rubber bounded above by the graph of a positive differentiable function* g *and below by* XY *plane. You have a "cookie cutter" knife in the shape of a curve parameterized by one-to-one and differentiable* $Q(t) = \langle X(t), Y(t)\rangle$ *in the domain of* g *for* $-1 \leq t \leq 1$*.*

Slicing vertically, you punch through your block of rubber and, moving the knife a tiny distance, punch through the block again to yield a very thin, curvy sheet of rubber.

You gently unroll this sheet, flatten it out, and lay it on the XY *plane with bottom edge on the* X *axis, the edge corresponding to* $Q(-1)$ *on the left and* $Q(1)$ *on the right.*

What is the slope of the upper edge at the point cut by $Q(0)$*?*

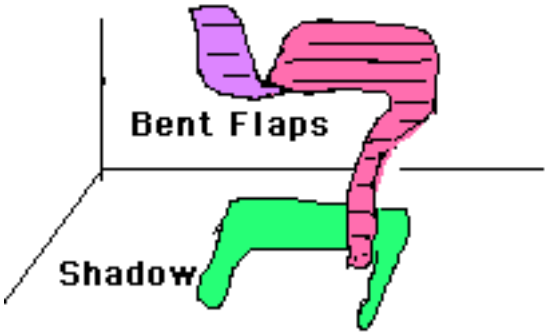
28.6. ***Exercise***. *Suppose* Q *is a differentiable parameterization of a curve in the domain of* g*. Then* $g \circ Q$ *is an ordinary real valued function. Show that if* $g \circ Q$ *has a local extreme value at* t *then the velocity of* Q *at* t *is perpendicular to the gradient of* g *at* $Q(t)$ *if that gradient is nonzero: the motion is at right angles to the direction of maximum increase of* g*, if there is one.*

Though this latter condition is necessary for a local extreme value of $g \circ Q$ *at* t *it is not sufficient to require a local extreme value at* t*.*

28.7. ***Exercise***. ** g is said to have a **local maximum value at** P if there is some circle centered at P for which $g(P) \geq g(A)$ for all A inside this circle. A similar definition is used to define a **local minimum value at** P. When one or the other holds, g is said to have a **local extreme value at** P.*

Prove that if ∇g exists everywhere on the open set $\mathfrak{O}$ then the only places where g could attain a local extreme value in $\mathfrak{O}$ are at points P for which $\nabla g(P) = 0$. This condition, though necessary, is not sufficient for the existence of a local extreme value.

28.8. ***Exercise***. *Suppose g is a differentiable real valued function on an open set in $\mathbb{R}^2$ and $D_2 g$ is always 0. One might suspect that g does not depend on the second coordinate: that $g(x, y_1) = g(x, y_2)$ whenever (x, y_1) and (x, y_2) are in the domain of g. This is false (produce a counterexample.)*



*However it **is** true that if the entire line segment connecting (x, y_1) and (x, y_2) is in the domain of g and $D_2 g(P) = 0$ all along the line segment then $g(x, y_1) = g(x, y_2)$.*

*This could be rephrased in special cases as follows: If $\mathfrak{H}$ is an open disk or an open rectangle inside the domain of g and if $D_2 g$ is always 0 in $\mathfrak{H}$ then g does not depend on the second coordinate **inside** $\mathfrak{H}$.*

In elementary Calculus the second derivative test can be used to decide if a function has a local maximum or minimum at a critical point. A similar **second derivative test** holds in higher dimensions.

If g is a twice continuously differentiable real valued function on an open set and $\nabla g(P) = 0$ we let $A = D_{1,1} g(P)$, $B = D_{1,2} g(P)$ and $C = D_{2,2} g(P)$.

The second derivative test depends on the sign of $B^2 - AC$. Note that if $B^2 - AC < 0$ it must be that A and C are both nonzero and have the same sign.

If $B^2 - AC < 0$ and $A < 0$ then g has a local maximum at P.
If $B^2 - AC < 0$ and $A > 0$ then g has a local minimum at P.
If $B^2 - AC > 0$ there is definitively not a local extreme of g at P.
If $B^2 - AC = 0$ the test is inconclusive.

This matter is discussed in considerable detail in the endnotes.[30]

28.9. ***Exercise***. ***Pete's World***: *Pete lives on a surface in a strange universe with one lake trapped in a low spot and one hilltop. The surface is the graph of* $g(X,Y) = \frac{X^3}{3} - \frac{X^2}{2} - 2X + \frac{Y^3}{3} - 9Y$ *where gravity pulls in the negative* Z *direction and distances are in meters. What is the greatest depth of water (from lake bottom to lake surface) that Pete's lake could hold?*

28.10. ***Exercise***. *Suppose* $P(t) = A + tB$ *and* $Q(t) = C + tD$ *are parametric vector equations for the position of two particles moving with constant velocity as time* t *passes.*

(i) How close will the particles get to each other?

(ii) Consider $H(s,t) = (Q(s) - P(t)) \cdot (Q(s) - P(t))$ *and determine the minimum distance between the geometrical tracks of the two particles.*

We note, for reference, the result from Exercise 10.2 which did not use methods of calculus.

28.11. ***Exercise***. ** Suppose you have a surface given as the graph of a twice differentiable function* g *on the plane. Suppose that for every vector* V *in the plane the function* $g(tV)$ *has a local minimum at* $t = 0$*: that is,* g *has a local minimum at the origin whenever we restrict attention to straight lines through the origin.*

It is not true that g *must have a local minimum at the origin, though it is rather tricky to provide an example of a formula for* g *which demonstrates this counterintuitive behaviour.*

Can you sketch a picture or describe in some other way a graph that behaves like this?

You may recall from Calculus that sufficiently differentiable functions can be used to form polynomials, called the **Taylor Polynomials**[31] which can be used to approximate the function. Specifically, if f can be differentiated n times at a then:

$$f(t) = \sum_{m=0}^{n} \frac{f^{(m)}(a)}{m!}(t-a)^m + R_n(t).$$

The polynomial $\sum_{m=0}^{n} \frac{f^{(m)}(a)}{m!}(t-a)^m$ is called the nth Taylor Polynomial at a and frequently denoted $P_n(t)$. The term $R_n(t) = f(t) - P_n(t)$ is called the **remainder** and if this is small $P_n(t)$ is a good approximation to $f(t)$. In case $f^{(n)}$ exists and is continuous on an interval containing a and t and $f^{(n+1)}$ exists on (a,t) (or on (t,a) should t be less than a) there is a c between a and t with $R_n(t) = \frac{f^{(n+1)}(c)}{(n+1)!}(t-a)^{n+1}$. If $|f^{(n+1)}|$ is bounded above by a number M on an interval containing a and t this tells us that $|R_n(t)| \leq \frac{M}{(n+1)!}|t-a|^{n+1}$, an inequality used most commonly to get an upper estimate for the mistake we make by using $P_n(t)$ instead of $f(t)$ itself.

We will use this fact from Calculus to create a similar formula for functions defined on the plane.

Suppose g is differentiable and all partial derivatives of g up to order $n+1$ exist and are continuous on a rectangle containing points with position vectors A and $A+V$ inside (not on the edge) of the rectangle.

Define $f(t) = g(A+tV)$. So $f'(t) = (\sum_{i=1}^{2} v_i D_i\ g)(A+tV)$. Differentiating once again gives $f''(t) = \left(\sum_{i,j=1}^{2} v_i v_j D_{i,j}\ g\right)(A+tV)$, where in a sum with multiple indices (in this case i and j) it is understood that each counter traverses its range independently.

In general we have

$$f^{(n)}(0) = \left(\sum_{i_1,i_2,\ldots,i_n=1}^{2} v_{i_1}\cdots v_{i_n} D_{i_1,\cdots,i_n}\ g\right)(A).$$

Observe that $f(1) = g(A+V)$ and denote $R_n(1) = R_n^A(V)$, so that

$$\begin{aligned} g(A+V) &= \sum_{m=0}^{n} \frac{f^{(m)}(0)}{m!} + R_n^A(V) \\ &= \sum_{m=0}^{n} \frac{\left(\sum_{i_1,\ldots,i_m=1}^{2} v_{i_1}\cdots v_{i_m} D_{i_1,\cdots,i_m}\ g\right)(A)}{m!} + R_n^A(V). \end{aligned}$$

The first part of the last line is an *nth* degree polynomial in the coordinates of V, the difference vector between the place where the derivatives are calculated and the place where we are approximating g. We will call it $P_n^A(V)$. The error in the approximation, $R_n^A(V) = g(A+V) - P_n^A(V)$, cannot exceed in magnitude the number $\frac{M}{(n+1)!}$ where M is the maximum magnitude, over the interval $0 \le t \le 1$, of

$$\left(\sum_{i_1,i_2,\ldots,i_{n+1}=1}^{2} v_{i_1}\cdots v_{i_{n+1}} D_{i_1,\cdots,i_{n+1}}\ g\right)(A+tV).$$

Obviously if n is bigger than 3 (or possibly 4) the calculations become intractable and a person would have to be **highly motivated** to deal with all the terms. But for small n it is not so bad. Let's look at an example with $n=2$ to see how this all works.

Let $g(X,Y) = \frac{cos(Y)}{X^3}$. Let $A = \langle 3, 0\rangle$. We will find $P_2^A(V)$ and an estimate for $R_2^A(1)$ when $-.1 < v_1 < .1$ and $-.1 < v_2 < .1$. First we calculate the partial

derivatives on the rectangle:

$$D_1\ g(X,Y) = \frac{-3cos(Y)}{X^4} \quad D_2\ g(X,Y) = \frac{-sin(Y)}{X^3} \quad D_{1,1}\ g(X,Y) = \frac{12cos(Y)}{X^5}$$
$$D_{1,2}\ g(X,Y) = D_{2,1}\ g(X,Y) = \frac{3sin(Y)}{X^4} \quad D_{2,2}\ g(X,Y) = \frac{-cos(Y)}{X^3}$$
$$D_{1,1,1}\ g(X,Y) = \frac{-60cos(Y)}{X^6} \quad D_{2,2,2}\ g(X,Y) = \frac{sin(Y)}{X^3}$$
$$D_{1,1,2}\ g(X,Y) = D_{1,2,1}\ g(X,Y) = D_{2,1,1}\ g(X,Y) = \frac{-12sin(Y)}{X^5}$$
$$D_{1,2,2}\ g(X,Y) = D_{2,1,2}\ g(X,Y) = D_{2,2,1}\ g(X,Y) = \frac{3cos(Y)}{X^4}.$$

So for the possible vectors V the formula for the remainder, in this case

$$\frac{1}{3!}\left(v_1^3 D_{1,1,1}\ g + 3v_1^2 v_2 D_{1,1,2}\ g + 3v_1 v_2^2 D_{1,2,2}\ g + v_2^3 D_{2,2,2}\ g\right)$$

evaluated somewhere between A and $A+V$ cannot exceed in magnitude

$$\frac{.1^3}{6}\left(\frac{60}{2.9^6} + 3\frac{1.2}{2.9^5} + 3\frac{3}{2.9^4} + \frac{.1}{2.9^3}\right) < 4.2 \times 10^{-5}.$$

If this is good enough for your purpose you can use the polynomial

$$P_2^A(V)$$
$$= g(A) + v_1 D_1\ g(A) + v_2 D_2\ g(A) + v_1^2 D_{1,1}\ g(A) + 2v_1 v_2 D_{1,2}\ g(A) + v_2^2 D_{2,2}\ g(A)$$
$$= \frac{1}{27} - \frac{v_1}{27} + \frac{4v_1^2}{81} - \frac{v_2^2}{27}$$

in place of $g(A+V)$.

28.12. ***Exercise***. *(i) Show that the graph of $Z = P_1^A(\langle X,Y\rangle - A)$ is the tangent plane to the graph of g at A.*

*(ii) * Show also that if g has bounded second partials in the vicinity of A that*

$$\lim_{V\to 0} \frac{|R_1^A(V)|}{|V|^2} \text{ exists.}$$

*Rephrasing, this means that when g has bounded second derivatives in the vicinity of A (this would happen if g had continuous second partials, for example) then near A the tangent plane at A is extremely close to the graph of g. The distance between them is proportional to the **square** of the distance to A.*

*(iii) * Generalize (ii) to higher order remainders.*

---

29. Functions of Three or More Variables

In the discussion above we thought about real and vector functions of one or two variables. In this section we will extend the discussion, usually by merely making an observation, to functions of three variables. The reader is invited at each step to refer back to the $2D$ versions of these statements, and consider any further adaptation which would be needed to treat functions of four or more variables. Though some important ideas are singled out as exercises, the section really should be thought of as a long exercise, placed here to make sure that the earlier ideas are solidified.

Suppose $\mathfrak{O}$ is a set in the space. A set such as $\mathfrak{O}$ is called **open** if for each P in $\mathfrak{O}$ there is a sphere centered at P entirely inside $\mathfrak{O}$. A set in space is called **closed** if its complement in $\mathbb{R}^3$ is open: that is, the set of points in space **not** in the set constitute an open set.[32]

We will consider **real functions of three variables.** Usually the domain of our functions will be an open set.

If g is a function defined for points in $\mathfrak{O}$ and P is in $\mathfrak{O}$ we write $\lim_{Q\to P} g(Q) = L$ if and only if for each $\varepsilon > 0$ there is some $\delta > 0$ such that if $0 < |Q - P| < \delta$ then $|g(Q) - L| < \varepsilon$.

When the limit exists, you can force $g(Q)$ to be as close to L as you wish by requiring Q to be close enough to P (P itself excluded.) L is called the **limit of** g **at** P.

Letting $\Delta P = Q - P$ we can see once again that $\lim_{\Delta P\to 0} g(P + \Delta P) = L$ is identical in meaning to $\lim_{Q\to P} g(Q) = L$.

29.1. ***Exercise***. *Let $\Delta P = \langle \Delta X, \Delta Y, \Delta Z\rangle$ and define $S_{\Delta P}$ to be $|\Delta X|+|\Delta Y|+|\Delta Z|$ and let $M_{\Delta P}$ be the largest of $|\Delta X|$ or $|\Delta Y|$ or $|\Delta Z|$.*

Show that:

$$M_{\Delta P} \le |\Delta P| \le S_{\Delta P} \le 3M_{\Delta P}.$$

29.2. ***Exercise***. *We could rephrase the definition of limit found above to any of the following equivalent conditions:*

For any $\varepsilon > 0$ there is a $\delta > 0$ so that

(i) if $0 < |\Delta P| < \delta$ then $|g(P + \Delta P) - L| < \varepsilon$.

(ii) if $0 < M_{\Delta P} < \delta$ then $|g(P + \Delta P) - L| < \varepsilon$.

(iii) if $0 < S_{\Delta P} < \delta$ then $|g(P + \Delta P) - L| < \varepsilon$.

(iv) if $0 < r < \delta$ and if U is any $3D$ unit vector then $|g(P + rU) - L| < \varepsilon$.

A function such as g is called **continuous at** P if $\lim_{Q\to P} g(Q)$ exists and is $g(P)$. g is called **continuous on** $\mathfrak{O}$ if it is continuous at every point in $\mathfrak{O}$.

If P and ΔP are $3D$ vectors and both P and $Q = P + \Delta P$ are in the domain of g we sometimes use the shorthand Δg for $g(P + \Delta P) - g(P)$ or, equivalently, $g(Q) - g(P)$.

g is called **differentiable at** P in $\mathfrak{O}$ if there is a $3D$ vector A so that

$$\lim_{\Delta P \to 0} \frac{\Delta g - A \cdot \Delta P}{|\Delta P|} = 0.$$

When this limit exists and is 0 the vector A is unique: that is, there can be no other such vector.

When the vector A as above exists we call it the **gradient of** g **at** P. This vector is denoted $\nabla g(P)$. In some books you will see $grad\ g(P)$ used to denote the gradient.

The function g is continuous wherever the gradient exists.

You can, as before, form the function $h(X) = g(X, P_2, P_3)$ parameterized by X in an interval around P_1. This function might be differentiable at P_1. If it is, the derivative is denoted $D_1 g(P)$ and similar definitions allowing the other variables to vary one at a time yields derivatives $D_2 g(P)$ and $D_3 g(P)$. These are called **partial derivatives of** g **at** P. **Higher, mixed** and **second partial derivatives** are defined just as in the $2D$ case.

The entries of $\nabla g(P)$, when it exists, are $D_1 g(P),\ D_2 g(P)$ and $D_3 g(P)$.

The gradient ∇g is an example of a **vector valued function in space**. Generally, such functions are also called **vector fields**.

Any vector valued function defined on an open set such as $\mathfrak{O}$ is called **continuous** at P if its coordinate functions are continuous at P, just as before. It is called continuous on $\mathfrak{O}$ if it is continuous at each point of $\mathfrak{O}$.

It is not true that the existence of $D_1 g,\ D_2 g$ and $D_3 g$ imply that $\nabla g(P)$ exists. However, if g is a function defined around a point P and if $D_1 g,\ D_2 g$ and $D_3 g$ are defined and continuous on some open set containing P then g is differentiable at P.

When $\nabla g(P)$ exists we define

$$L_{g,P}(Q) = g(P) + \nabla g(P) \cdot (Q - P).$$

$L_{g,P}$ is called the **linearization of** g **at** P. For such P we have

$$\lim_{Q \to P} \frac{g(Q) - L_{g,P}(Q)}{|Q - P|} = 0.$$

So when Q is close to P, not only is $g(Q)$ near to $L_{g,P}(Q)$ but the difference between them is small **even in comparison to** $|Q - P| = |\Delta P|$.

Let's suppose that the gradient exists in the vicinity of P and that $Q(t) = \langle X(t), Y(t), Z(t) \rangle$ is a differentiable parameterization of a curve in space with nonzero derivative and $Q(c) = P$. Then $H(t) = g \circ Q(t)$ is an ordinary real valued function.

What is the derivative of H at c? We define $\Delta P = Q(c + h) - Q(c) = \langle \Delta X, \Delta Y, \Delta Z \rangle$ and note that because Q is continuous at c then $\lim_{h \to 0} \Delta P = 0$. Since g is differentiable, this implies that

$$\lim_{h \to 0} \frac{g(P + \Delta P) - g(P) - \nabla g(P) \cdot \Delta P}{|\Delta P|} = 0.$$

Rewriting this we have

$$\lim_{h\to 0}\left|\frac{\Delta P}{h}\right|^{-1}\left[\frac{g(P+\Delta P)-g(P)}{h}-\left(D_1g(P)\frac{\Delta X}{h}+D_2g(P)\frac{\Delta Y}{h}+D_3g(P)\frac{\Delta Z}{h}\right)\right]=0.$$

By assumption, the factor on the left converges to a nonzero number and the terms in the inner parentheses on the right converge to $D_1g(P)X'(c)+D_2g(P)Y'(c)+D_2g(P)Z'(c)$ so we have discovered that $H'(c)=\lim_{h\to 0}\frac{g(P+\Delta P)-g(P)}{h}=\frac{d}{dt}(g\circ Q)(c)$ exists and equals $\nabla g(Q(c))\cdot Q'(c)$.

29.3. ***Exercise***. *With conditions as above except that $Q'(c)=0$ show that*

$$\frac{d}{dt}(g\circ Q)(c)=\nabla g(Q(c))\cdot Q'(c)=0.$$

So H is differentiable at c and $H'(c)=\nabla g(Q(c))\cdot Q'(c)$ in this case too.

We have just proved a $3D$ version of **The Chain Rule:**

$$\frac{d}{dt}(g\circ Q)(c)=\nabla g(Q(c))\cdot Q'(c).$$

Recall that in space just as in the plane when the two vectors $\nabla g(Q(c))$ and $Q'(c)$ are nonzero

$$\nabla g(Q(c))\cdot Q'(c)=|\nabla g(Q(c))||Q'(c)|cos(\theta)$$

where θ is the angle between $\nabla g(Q(c))$ and $Q'(c)$.

So among all differentiable curves Q in space which pass through $Q(c)$ **with a given speed**, the rate at which the function $g\circ Q$ is increasing or decreasing depends only on the angle θ. If $Q'(c)$ points in the same direction as $\nabla g(Q(c))$ then $(g\circ Q)(t))$ is increasing at maximum rate. If you move in the opposite direction $(g\circ Q)(t))$ is decreasing at maximum rate. If $cos(\theta)=0$, so θ is $\frac{\pi}{2}$, we have $\frac{d}{dt}(g\circ Q)(c)=0$. The collection of all vectors for which $cos(\theta)=0$ constitute a plane with normal $\nabla g(Q(c))$.

We define the **directional derivative for the vector** V at a point P in the domain of g to be $\nabla g(P)\cdot V$. The notation $\nabla_V g(P)$ is used for this number. It represents the rate at which $g\circ Q$ is increasing at c if you are moving along **any** parameterized curve with $Q(c)=P$ and $Q'(c)=V$.

29.4. ***Exercise***. *Suppose Q is a differentiable parameterization of a curve in the domain of differentiable g. Then $g\circ Q$ is an ordinary real valued function. Show that if $g\circ Q$ has a local extreme value at t then the velocity of Q at t is perpendicular to the gradient of g at $Q(t)$ if that gradient is nonzero: the motion is at right angles to the direction of maximum increase of g, if there is one.*

Though this latter condition is necessary for a local extreme value of $g\circ Q$ at t it is not sufficient to require a local extreme value at t.

*g is said to have a **local maximum value at** P if there is some sphere centered at P for which $g(P)\geq g(A)$ for all A inside this sphere. A similar definition is*

used to define a ***local minimum value at*** P. *When one or the other holds, g is said to have a* ***local extreme value at*** P.

Prove that if ∇g exists everywhere on the open set $\mathfrak{O}$ then the only places where g could attain a local extreme value in $\mathfrak{O}$ are at points P for which $\nabla g(P) = 0$. This condition, though necessary, is not sufficient for the existence of a local extreme value.

29.5. ***Exercise***. *Find any local extreme values of*

$$g(X, Y, Z) = X^2 - 2XY + 2Y^2 + Z^2 + 2Z.$$

29.6. ***Exercise***. ** Suppose g is a differentiable real valued function on an open set $\mathfrak{O}$ in $\mathbb{R}^3$ and ∇g is the zero vector on all of $\mathfrak{O}$. One might suspect that g is constant. This is false (produce a counterexample.)*

However it ***is*** *true that if the entire line segment connecting P and Q is in the domain of g and $\nabla g = 0$ all along the line segment then $g(P) = g(Q)$.*

This could be rephrased in special cases as follows: If $\mathfrak{H}$ is an open ball or an open rectangular solid inside the domain of g and if ∇g is always 0 in $\mathfrak{H}$ then g is constant ***inside*** *$\mathfrak{H}$.*

As a further extension, g will be constant when restricted to any set that can be built as the union of sets like $\mathfrak{H}$ where new sets are added one at a time and each new set ***overlaps*** *at least one set which was added previously.*

29.7. ***Exercise***. ** Create* ***Taylor Polynomials*** *with error estimate for functions defined in $3D$. Then apply your ideas to form an approximating polynomial P_2^A for the function $g(X, Y, Z) = \frac{X^2 Y}{cos(Z^2)}$ on the cube with edge length .2 centered at $A = \langle 3, 0, 0 \rangle$.*

30. Implicit Functions

In this section we will discuss curves and surfaces defined "implicitly" rather than "explicitly" as the graph of a function. Examples of such surfaces abound. The collection of points defined by $X^2 + Y^2 + Z^2 = 1$, the sphere of unit radius, is an example. Z cannot be solved for explicitly as a function of X and Y because there are multiple Z values for each X and Y pair inside the unit circle. The sphere is, however, the graph of two functions: the upper and lower hemisphere. They are patched together at the equator. In many cases the formula relating the variables has them inextricably entangled and not even this explicit "patching" is possible. Even so, it is often possible to do quite a bit with them.

We will first work through the (rather extensive) setup in a special case.

Suppose g is a continuously differentiable real function of three variables defined on an open set $\mathfrak{O}$ containing the origin and $g(0,0,0) = 0$.

Suppose that $D_3 g(0,0,0) = D > 0$. We will first confine attention to the interior of a sphere centered at the origin and so small that $D_3 g(Q) > \frac{D}{2}$ for all Q in the sphere.

We will further require the ball to be small enough so that

$$\frac{|g(Q) - \nabla g(0) \cdot Q|}{|Q|} \leq \frac{D}{2} \quad \text{for all } Q \text{ in the ball.}$$

We now confine our attention to a cylinder with axis parallel with $\vec{k}$ and centered at the origin with height above and below the XY plane equal to its radius, r, and contained inside this ball.

For every point Q in this cylinder we have

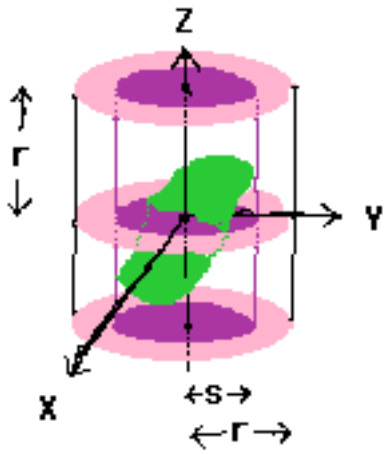
$$-\frac{D}{2}|Q| \leq g(Q) - \nabla g(0) \cdot Q \leq \frac{D}{2}|Q|.$$

This means that	$-\frac{D}{2}r \leq g(0,0,r) - Dr \leq \frac{D}{2}r$
and	$-\frac{D}{2}r \leq g(0,0,-r) + Dr \leq \frac{D}{2}r$
so	$\frac{D}{2}r \leq g(0,0,r) \leq \frac{3D}{2}r$
and	$-\frac{3D}{2}r \leq g(0,0,-r) \leq -\frac{D}{2}r.$

We have shown that $g(0,0,r)$ is positive and $g(0,0,-r)$ is negative. Since g is continuous, there is a smaller concentric cylinder $\mathfrak{H}$ of the same height and radius s so that g is positive on every point on the top of the cylinder and negative on every point on the bottom.

If (X,Y,r) is any point on the top of $\mathfrak{H}$ and $(X,Y,-r)$ the corresponding point on the bottom then $Q(t) = g(X,Y,t)$ is continuous with $Q(-r) < 0 < Q(r)$ so there exists at least one t with $-r < t < r$ and $Q(t) = g(X,Y,t) = 0$.

We wish to rule out the possibility that there is more than one such t for each (X,Y) pair.



Suppose $g(X,Y,t_1) = g(X,Y,t_2) = 0$ for (X,Y,t_1) and (X,Y,t_1) in $\mathfrak{H}$. So

$$|g(X,Y,t_2) - g(X,Y,t_1) - \nabla g(X,Y,t_1) \cdot \langle 0,0,t_2 - t_1\rangle\,| \leq \frac{D}{2}|\,\langle 0,0,t_2 - t_1\rangle\,|$$

But then

$$|D_3 g(X,Y,t_1)(t_2 - t_1)| \leq \frac{D}{2}|t_2 - t_1|.$$

This is impossible unless $t_2 = t_1$ in light of the fact that the original sphere was chosen so that $D_3 g(Q) > \frac{D}{2}$ for all Q in the sphere.

So for each (X,Y) inside a disk $\mathfrak{D}$ of radius s centered at $(0,0)$ there is a unique $Z = Z(X,Y)$ for which (X,Y,Z) is in $\mathfrak{H}$ and $g(X,Y,Z(X,Y)) = 0$. We will let $\mathfrak{S}$ be the collection of all these $(X,Y,Z(X,Y))$ in $\mathfrak{H}$. We will show that Z is differentiable.

To that end we first look at the following calculation. Suppose $(X,Y,Z) \neq (0,0,0)$ is in $\mathfrak{S}$. Then

$$|\nabla g(0,0,0) \cdot \langle X,Y,Z\rangle\,| \leq \frac{D}{2}|\,\langle X,Y,Z\rangle\,|.$$

Expanding this gives

$$|D_1 g(0,0,0)X + D_1 g(0,0,0)Y + DZ| \leq \frac{D}{2}|\,\langle X,Y,0\rangle + \langle 0,0,Z\rangle\,|$$

and so

$$D|Z| - |D_1 g(0,0,0)X| - |D_1 g(0,0,0)Y| \leq \frac{D}{2}|\,\langle X,Y\rangle\,| + \frac{D}{2}|Z|.$$

This gives

$$\frac{D}{2}|Z| \leq \frac{D}{2}|\,\langle X,Y\rangle\,| + |D_1 g(0,0,0)X| + |D_1 g(0,0,0)Y|.$$

Dividing everywhere by nonzero $|\,\langle X,Y\rangle\,|$ we have

$$\frac{D}{2}\frac{|Z|}{|\,\langle X,Y\rangle\,|} \leq \frac{D}{2} + |D_1 g(0,0,0)|\frac{|X|}{|\,\langle X,Y\rangle\,|} + |D_1 g(0,0,0)|\frac{|Y|}{|\,\langle X,Y\rangle\,|}$$

which yields

$$\frac{|Z|}{|\langle X,Y\rangle|} \le 1 + \frac{2}{D}\left(|D_1g(0,0,0)| + |D_1g(0,0,0)|\right).$$

This little calculation implies that whenever $(X,Y,Z) \neq (0,0,0)$ is in $\mathfrak{S}$ then the fraction $\frac{|\langle X,Y,Z\rangle|}{|\langle X,Y\rangle|}$ is bounded by a fixed number which we will call K.

We are now ready to show that Z is differentiable at $(0,0)$ and

$$\nabla Z(0,0) = \left\langle \frac{-D_1g(0,0,0)}{D_3g(0,0,0)}, \frac{-D_2g(0,0,0)}{D_3g(0,0,0)} \right\rangle.$$

This is because

$$\begin{aligned}
&\frac{K}{D}\frac{|\nabla g(0,0,0)\cdot\langle X,Y,Z\rangle|}{|\langle X,Y,Z\rangle|}\\
=&\frac{K}{D}\frac{|D_1g(0,0,0)X + D_2g(0,0,0) + DZ|}{|\langle X,Y,Z\rangle|}\\
\ge&\frac{1}{D}\frac{|\langle X,Y,Z\rangle|}{|\langle X,Y\rangle|}\frac{|D_1g(0,0,0)X + D_2g(0,0,0)Y + DZ|}{|\langle X,Y,Z\rangle|}\\
=&\frac{|\frac{D_1g(0,0,0)}{D}X + \frac{D_2g(0,0,0)}{D}Y + Z|}{|\langle X,Y\rangle|}\\
=&\frac{|Z - \left\langle -\frac{D_1g(0,0,0)}{D}, -\frac{D_2g(0,0,0)}{D}\right\rangle\cdot\langle X,Y\rangle|}{|\langle X,Y\rangle|}.
\end{aligned}$$

Since the first line has limit 0 as $\langle X,Y\rangle$ approaches 0 the last must converge to 0 too and the result is proved.

30.1. ***Exercise***. ** *Suppose g is as above and $C = \langle c_1,c_2,c_3\rangle$ is a vector and ω is a constant. Let $\widetilde{\mathfrak{H}} = \{P + C \mid P \in \mathfrak{H}\}$ and $\widetilde{\mathfrak{S}} = \{P + C \mid P \in \mathfrak{S}\}$ and $\widetilde{\mathfrak{D}} = \{P + C \mid P \in \mathfrak{D}\}$. Define $h(Q) = g(Q - C) + \omega$ whenever $Q \in \mathfrak{O}$.*

(i) h is continuously differentiable and defined on an open set.

(ii) $D_3h(C) = D > 0$.

(iii) $\widetilde{\mathfrak{S}} = \{Q \in \widetilde{\mathfrak{H}} \mid h(Q) = \omega\}$ and $C \in \widetilde{\mathfrak{S}}$.

(iv) For each $(X,Y) \in \widetilde{\mathfrak{D}}$ there is a unique Z so that $(X,Y,Z) \in \widetilde{\mathfrak{S}}$. So Z is a function defined for all points in the disk $\widetilde{\mathfrak{D}}$.

(v) Z is continuously differentiable at (c_1,c_2) and

$$\nabla Z(c_1,c_2) = \left\langle \frac{-D_1h(C)}{D_3h(C)}, \frac{-D_2h(C)}{D_3h(C)} \right\rangle.$$

(vi) Z is continuously differentiable at every point

$$P = (p_1,p_2,p_3) = \Big(p_1,p_2,Z(p_1,p_2)\Big) \in \widetilde{\mathfrak{D}}$$

$$\text{and} \qquad \nabla Z(p_1,p_2) = \left\langle \frac{-D_1h(P)}{D_3h(P)}, \frac{-D_2h(P)}{D_3h(P)} \right\rangle.$$

(vii) The same final result (vi) is true if all conditions are the same except that $D_3h(C) < 0$.

If h is any function defined on a set $\mathfrak{O}$ in the plane or in space and $h(C) = \omega$ the set $\{ P \in \mathfrak{O} \mid h(P) = \omega \}$ is called a **level set of** h.

We have just discovered that if h is continuously differentiable on an open set $\mathfrak{O}$ in space and $D_3h(C) \neq 0$ then locally (that is, on some small disk centered at (c_1, c_2)) the level set is the graph of a differentiable function Z and we have an explicit representation of the gradient of Z in terms of the entries in the gradient of h.

The curves $T_1(X) = \langle X, C_2, Z(X, C_2)\rangle$ and $T_2(Y) = \langle C_1, Y, Z(C_1, Y)\rangle$ are curves in this surface passing through C. These curves have tangent vectors

$$T_1'(C_1) = \langle 1, 0, D_1Z(C_1, C_2)\rangle \text{ and } T_2'(C_2) = \langle 0, 1, D_2Z(C_1, C_2)\rangle$$

at C, which lie in the tangent plane to the surface at C.

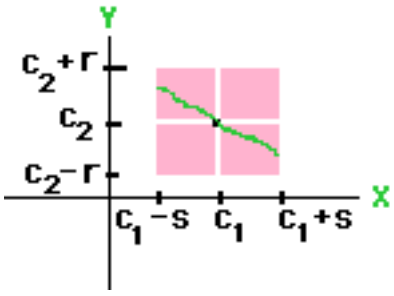
The vector

$$\begin{aligned} T_1(C_1) \times T_2(C_2) & \\ &= \langle -D_1Z(C_1, C_2), -D_2Z(C_1, C_2), 1\rangle \\ &= \left\langle \frac{D_1h(C)}{D_3h(C)}, \frac{D_2h(C)}{D_3h(C)}, 1 \right\rangle \\ &= \frac{1}{D_3h(C)} \langle D_1h(C), D_2h(C), D_3h(C)\rangle \end{aligned}$$

is normal to the surface and is an explicit nonzero multiple of $\nabla h(C)$ which is, therefore, itself normal to the surface (that is, to the tangent plane) at C.

30.2. ***Exercise***. *Prove that the analogous results for functions h defined on an open set $\mathfrak{O}$ in the plane is also true:*

Suppose $C = (c_1, c_2)$ is a point in $\mathfrak{O}$ and h is continuously differentiable on $\mathfrak{O}$ and $h(C) = \omega$ and $D_2h \neq 0$.



Then there is a rectangle $[c_1 - s, c_1 + s] \times [c_2 - r, c_2 + r]$ so that for every x in $[c_1 - s, c_1 + s]$ there is a unique y in $[c_2 - r, c_2 + r]$ with $h(x, y) = \omega$. So the part of the level set inside the rectangle is the graph of a function.

The function y is differentiable and $\frac{dy}{dx}(x, y) = \frac{-D_1h(x,y)}{D_2h(x,y)}$ for each x in $(c_1 - s, c_1 + s)$.

Moreover, $\nabla h(x, y)$ is normal to the tangent line at each (x, y) in that part of the level set inside the rectangle.

Some level sets do not exactly fall into the pattern discussed above, but still form what we would want to call surfaces. Consider, for example, the function $F(P) = P \cdot P$ where P is a point in $\mathbb{R}^3$ and the related function $G(Q) = Q \cdot Q$ where Q is a point in $\mathbb{R}^2$. The level set $F(P) = 1$ is the unit sphere, and we certainly think of this as a surface. However $D_3F = 0$ on the equator (that is, on the XY plane) so those points do not conform to the pattern we looked at above. Similarly, the level set $G(Q) = 1$ is the unit circle, and any piece of this curve around $(\pm 1, 0)$ cannot be the graph of a function of the first coordinate.

We solve this problem by simply changing point of view. We note that ∇F is never the zero vector, so for every point in the sphere at least one of the partial derivatives is nonzero. So the variable corresponding to this derivative can be solved for in terms of the others as above. Locally, the sphere can be formed by glueing together overlapping "patches" each of which is the graph of a function of two of the three variables.

Similarly, the unit circle can be formed by piecing together overlapping graphs of functions of either Y or X.

30.3. ***Exercise***. *How many patches, at least, would you need to cover the unit circle? How many would you need to cover the unit sphere? How many would you need to cover the torus from Section 13?*

Using the ideas from above we come to the following conclusion: If F is a continuously differentiable function defined on an open set in $\mathbb{R}^2$ or $\mathbb{R}^3$ let $\mathfrak{M}$ be a level set of F. If $\nabla F(P)$ is never the zero vector for any point in $\mathfrak{M}$ then $\mathfrak{M}$ can be formed as overlapping patches, each of which is the graph of one variable as a continuously differentiable function of the other(s).

Sets with this last property are called (differentiable) one or two dimensional **manifolds.** These manifolds, and their higher dimensional brethren, are very important in applications. After all, we seem to live in one and on one.

As a final remark, we note that every graph can be thought of as a level set. Suppose $\mathfrak{O}$ is an open set in the plain and F is a differentiable function defined on $\mathfrak{O}$. Let $\mathfrak{U}$ be the set of all ordered triples (X, Y, Z) for which (X, Y) is in $\mathfrak{O}$. The set $\mathfrak{U}$ is open.

Define the function H on $\mathfrak{U}$ by $H(X, Y, Z) = Z - F(X, Y)$.

So $\nabla H = \langle -D_1F, -D_2F, 1\rangle$ and the level set where $H = 0$ is the graph of F.

30.4. ***Exercise***. *(i) Prove that the set $\mathfrak{U}$ defined above is open.*

(ii) Go through an analogous discussion when F is a function of one variable to prove that the graph of differentiable function of one variable is the level set of a function H of two variables. What is the gradient of H in this case?

31. Derivatives as Matrices

This section is devoted to a useful consolidation of notation and an extension of the chain rule. It requires that you know how to multiply matrices of various sizes and we remind the reader of the particulars here.

We will suppose that L, M and N are taken from the integers $1, 2$ or 3. An $M \times N$ **matrix** (read as "M by N") is a rectangular array of numbers with M rows and N columns and matrices are usually surrounded by curved parentheses. For example the six matrices

$$\begin{pmatrix} 1 & 2 \end{pmatrix}\ ,\ \begin{pmatrix} 1 & 2 & -3 \end{pmatrix}\ ,\ \begin{pmatrix} 1 \\ 2 \end{pmatrix}\ ,\ \begin{pmatrix} 1 \\ 2 \\ -3 \end{pmatrix}\ ,\ \begin{pmatrix} 1 & 2 & -3 \\ 6 & 1 & 5 \\ -6 & 9 & -3 \end{pmatrix}\ ,\ \begin{pmatrix} 1 & 2 \\ 6 & 5 \end{pmatrix}$$

are of sizes 1×2, 1×3, 2×1, 3×1, 3×3, and 2×2 respectively.

If A is an $M \times N$ matrix the entries of A will be denoted $a_{i,j}$ where i denotes the row number counting from the top and j the column number counting from the left. So if A is the last matrix above we have $a_{1,1} = 1,\ a_{2,1} = 6,\ a_{1,2} = 2$ and $a_{2,2} = 5$.

Not every pair of matrices can be multiplied. They must be of the correct shape. In particular, the number of columns of the left matrix in a product must equal the number of rows of the right one.

For example the products

$$\begin{pmatrix} 1 & 2 & -3 \\ 6 & 1 & 5 \\ -6 & 9 & -3 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ -3 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

are formed from matrices of the right shape (in **that** order) but

$$\begin{pmatrix} 1 \\ 2 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 6 & 5 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & 2 & -3 \end{pmatrix} \begin{pmatrix} 1 & 2 & -3 \end{pmatrix}$$

are not compatible for purposes of multiplication.

If A is an $M \times L$ matrix and B is an $L \times N$ matrix the **product matrix** $C = AB$ is the $M \times N$ matrix with entries $c_{i,j} = \sum_{k=1}^{L} a_{i,k} b_{k,j}$.

So

$$\begin{pmatrix} 1 & 2 & -3 \\ 6 & 1 & 5 \\ -6 & 9 & -3 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ -3 \end{pmatrix} = \begin{pmatrix} 14 \\ -7 \\ 21 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 5 \end{pmatrix} = 5.$$

Vectors look a lot like matrices. For reasons that will become more clear after you take a Linear Algebra class, when vectors are thought of in this way they are represented as single column vectors. For example

$$\langle 3, 8 \rangle = \begin{pmatrix} 3 \\ 8 \end{pmatrix} \quad \text{and} \quad \langle -6,\ 9,\ -3 \rangle = \begin{pmatrix} -6 \\ 9 \\ -3 \end{pmatrix}.$$

A 1×1 matrix is usually not distinguished from a real number.

Matrices of the same shape can be **added** by adding corresponding entries. There is an operation called **scalar multiplication** in which matrices are multiplied by real constants by multiplying every entry in the matrix by the constant. This agrees with the operation of multiplication by a 1×1 matrix when that operation is defined.

There is an operation called **transposition** defined for matrices which acts to tip a matrix on its side by switching row and column numbers of the entries. It is denoted by a T exponent on a matrix. So, for example

$$\begin{pmatrix} 1 & 2 & 3 \\ 6 & 1 & 5 \\ -6 & 9 & -3 \end{pmatrix}^T = \begin{pmatrix} 1 & 6 & -6 \\ 2 & 1 & 9 \\ 3 & 5 & -3 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} -6 \\ 9 \\ -3 \end{pmatrix}^T = \begin{pmatrix} -6 & 9 & -3 \end{pmatrix}.$$

The second matrix is called the **transpose** of the first.

As an example of the notation in action, we can define the dot product of two vectors using the associated column matrices by

$$\langle 3, -7 \rangle \cdot \langle 4, 2 \rangle = 12 - 14 = -2 \begin{pmatrix} 5 \end{pmatrix} = \begin{pmatrix} 3 \\ -7 \end{pmatrix}^T \begin{pmatrix} 4 \\ 2 \end{pmatrix}.$$

Now we arrive at the reason for this notational reminder.

If f is a real valued differentiable function in the plane we define

$$f'(X, Y) = \begin{pmatrix} D_1 f(X, Y) & D_2 f(X, Y) \end{pmatrix}.$$

If f is a real valued differentiable function in space we define

$$f'(X, Y, Z) = \begin{pmatrix} D_1 f(X, Y, Z) & D_2 f(X, Y, Z) & D_3 f(X, Y, Z) \end{pmatrix}.$$

If $Q(t) = \langle X(t), Y(t) \rangle$ is a curve in the plane we define

$$Q'(t) = \begin{pmatrix} X'(t) \\ Y'(t) \end{pmatrix}$$

and If $Q(t) = \langle X(t), Y(t), Z(t) \rangle$ is a curve in space we define

$$Q'(t) = \begin{pmatrix} X'(t) \\ Y'(t) \\ Z'(t) \end{pmatrix}.$$

With the convention of identifying vectors with column matrices, this definition of $Q'(t)$ is the same as before and $\nabla f = (f')^T$.

These definitions are special cases of a more general formula for derivatives as matrices. Suppose M and N are $1, 2$ or 3. Suppose for $1 \leq i \leq M$ the function g_i is differentiable and defined for points in $\mathbb{R}^N$.

In the case of Q above, M is 2 or 3 and N is 1. In the case of f above we have $M = 1$ and $N = 2$ or 3. When both N and M are 1 we have a single ordinary real valued function.

Let us suppose $g = \langle g_1, \dots \rangle$ where $\dots$ means that you keep going until the subscript on the entries is M.

g is a function whose domain is an open set in $\mathbb{R}^N$ and whose range is in $\mathbb{R}^M$.

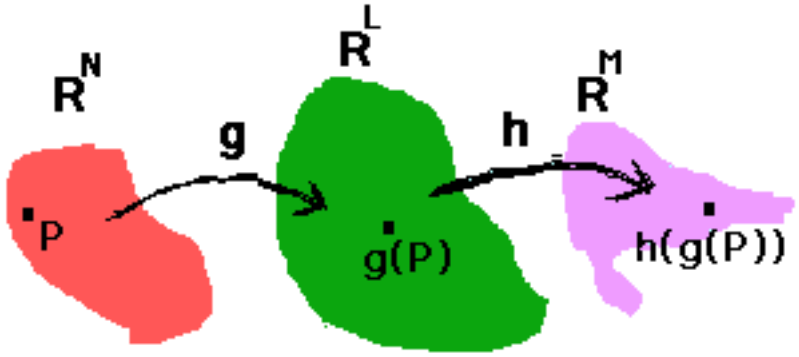
We define g' to be the matrix with entries $g'_{i,j} = D_j g_i$ and where, if $N = 1$ we interpret D_1 to be the ordinary derivative. In some texts this matrix is called the **Jacobian matrix** of g.

As an illustration, if g operates on points in $\mathbb{R}^3$ and has values in $\mathbb{R}^2$ we have:

$$g' = \begin{pmatrix} D_1 g_1 & D_2 g_1 & D_3 g_1 \\ D_1 g_2 & D_2 g_2 & D_3 g_2 \end{pmatrix}.$$

31.1. ***Exercise***. *Show that this definition agrees with the previous one involving Q and f as above.*

31.2. ***Exercise***. ** Suppose h and g are differentiable functions as defined above and the composite function $(h \circ g)(P)$ exists for each P in the domain of g. There are 27 different combinations of range and domain dimensions.*



Prove ***The Chain Rule****:* $(h \circ g)'(P) = h'(g(P))g'(P)$ *for each P in the domain of g.*

Observe that this formulation subsumes all earlier versions of the chain rule which you might have learned.

31.3. ***Exercise***. ** Show that if $N = L = M$ and h and g are differentiable and inverse functions (to each other) then $(h \circ g)'(P) = h'(g(P))g'(P)$ is the* ***identity matrix****: that is, one of the following three, depending on dimension:*

$$1 \qquad \textit{or} \qquad \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \textit{or} \qquad \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

(These are called identity matrices because whenever a matrix is of the right shape to multiply by one of them, the act of multiplication does not change the matrix.)

Observe that when $N = L = M = 1$ this yields the usual formula relating the derivatives of a function and its inverse:

$$g'(P) = \frac{1}{h'(g(P))}$$

If you know what an inverse matrix is, it yields the corresponding formula for the derivative of an ***inverse function****:*

$$\text{If } Q = g(P) \quad \text{then} \quad (g'(P))^{-1} = (g^{-1})'(Q).$$

31.4. ***Exercise****. Suppose $\mathfrak{O}$ is the set $(0,\infty)\times(0,2\pi)$. Define F for these points by $F(r,\theta) = (r\ cos(\theta), r\ sin(\theta))$.*

r and θ are polar coordinates of the point $F(r,\theta)$ and the domain of F has been chosen so that F is one-to-one and the range of F is the open set $\mathfrak{U}$ consisting of all points in the plane except those on the positive X axis and the origin.

We define $G(X,Y)$ to be:

$$\begin{cases} \left(\sqrt{X^2+Y^2}\ ,\ arccos\left(\frac{X}{\sqrt{X^2+Y^2}}\right)\right), & \text{If } (X,Y) \text{ is above the } X \text{ axis;} \\ \left(\sqrt{X^2+Y^2}\ ,\ 2\pi - arccos\left(\frac{X}{\sqrt{X^2+Y^2}}\right)\right), & \text{If } (X,Y) \text{ is below the } X \text{ axis;} \\ (\ |X|\ ,\ \pi\ ), & \text{If } X<0 \text{ and } Y=0. \end{cases}$$

G is defined on $\mathfrak{U}$ and is the inverse function to F.

Evaluate $G'(X,Y)$ and $F'(r,\theta)$ and show that both $(G\circ F)'$ and $(F\circ G)'$ are the identity matrices.

31.5. ***Exercise****. * Suppose that h and f are continuously differentiable functions on an open set $\mathfrak{O}$ and $Z = Z(X,Y)$ is a function defined on open $\mathfrak{D}$ in the plane and arises as a piece of a level set of h, where $(D_3h)(X,Y,Z(X,Y))$ is never 0 on $\mathfrak{D}$.*

Show that if the composite function $f(X,Y,Z(X,Y))$ has a local extreme value at $(X,Y)\in\mathfrak{D}$ then there is a real number λ so that $(\nabla f)(X,Y,Z(X,Y)) = \lambda(\nabla h)(X,Y,Z(X,Y))$. In other words, ∇f must be the zero vector or ∇f is perpendicular to the surface. The number λ is called a ***Lagrange Multiplier.***

hint: Let $W(X,Y) = (X,Y,Z(X,Y))$. So $f\circ W$ is a differentiable real valued function on an open set in the plane. It can only have extreme values at places where $\nabla(f\circ W)$, and hence $(f\circ W)'$, is zero. Use the chain rule to conclude that if P is a point on this level surface and is a local extrema of f among points on this level surface then $\nabla f(P) = \frac{D_3f(P)}{D_3h(P)}\nabla h(P)$.

31.6. ***Exercise***. ** Suppose that h and f are continuously differentiable functions on an open set $\mathfrak{O}$ where $\mathfrak{O}$ is in the plane or in space. Let $\mathfrak{M}$ be a level set of h and suppose that ∇h is never the zero vector on $\mathfrak{M}$.*

If f has a local extreme value at P among points in $\mathfrak{M}$ then there must be a number λ for which $\nabla f(P) = \lambda \nabla h(P)$.

The idea of Lagrange multipliers is as follows. Suppose we have a point P on a level surface $\mathfrak{M}$ of a function h defined for points in an open set in space. The condition that P come from $\mathfrak{M}$ is called a **constraint** for the problem. Suppose we have another function f, sometimes called the **objective function**, defined on this open set and we wish to consider the possibility that f has a local extreme value at P when we confine attention to points in $\mathfrak{M}$.

If f has a local maximum value at P when confined to $\mathfrak{M}$ then any direction you can go when confined to $\mathfrak{M}$ as you pass through P is a direction in which f is not increasing. The whole tangent plane at P must be perpendicular to $\nabla f(P)$. The same is true if f has a local minimum at P when confined to $\mathfrak{M}$.

Taking this a bit farther, suppose that we have a second constraint generated as a level set of differentiable g, and P is also in this second level set. Suppose we want a condition satisfied by points at which f has a local extreme value among points subject to **both** constraints—that is, among points in the intersection of the two level sets. One would expect a crossing of two level surfaces to be a curve, though it might be hard or impossible to write a formula for this curve.

When it is a curve, the tangent line for the curve at each point will be the line of crossing of the tangent planes of the two surfaces. So $\nabla g(P) \times \nabla h(P)$ will lie in the tangent line. If f has a local extreme value here, the tangent line must be perpendicular to the direction of increase of f.

So we come to the conclusion: If f has a local extreme value at P among points in the intersection of a level surface for g and a level surface for h then

$$\nabla f(P) \cdot \nabla g(P) \times \nabla h(P) = 0.$$

But when will the intersection be a curve? Intuition says it **should** be, but intuition can sometimes lead one astray. We discuss the matter in an endnote.[33]

Note that neither this condition for a point P nor the Lagrange multiplier condition of the problem above suffice, by themselves, for us to conclude that P is the location of a local extreme value. They are only necessary conditions.

However, we know that a continuous function on a closed and bounded domain actually attains both a maximum and minimum value on that domain, and these conditions can be used to reduce the search for extrema to fewer locations.

31.7. ***Exercise***. *(i) Use Lagrange Multipliers to help you determine how big $f(X, Y) = X - Y$ could get on the curve $X^2 + Y^2 = 4$.*

(ii) Use Lagrange Multipliers to help you determine how big $f(X, Y, Z) = XY + YZ + XZ$ could get in the all-positive octant when $X + 2Y + 3Z = 10$.

31.8. ***Exercise***. ***Pete's World Revisited:*** *Pete lives on a surface in a strange universe. The surface is the graph of* $g(X,Y) = \frac{X^3}{3} - \frac{X^2}{2} - 2X + \frac{Y^3}{3} - 9Y$. *Pete moves around on or inside the cylinder* $X^2 + Y^2 = 100$ *on Pete's world. This is Pete's property. What is the high point on Pete's property? The low spot?*

32. Potentials

We will consider here two related issues having to do with either differential equations or line integrals, depending on point of view.

We suppose that $F = F_1\vec{i} + F_2\vec{j} + \dots$ is a continuous vector field defined on an open set $\mathfrak{O}$.

We will be dealing with **piecewise good parameterizations** of **piecewise good curves** inside $\mathfrak{O}$.

Recall from Exercise 23.3 that this means the curve $\mathfrak{C}$ has a continuous parameterization Q with domain $[c,d]$ for which

- Q is one-to-one on $[c,d]$ or is one-to-one on $(c,d]$ and $Q(c) = Q(d)$
- $[c,d]$ can be broken into a finite number of pieces $[t_{i-1}, t_i]$ for $i = 1, \dots, n$ so that Q is continuously differentiable with nonzero derivative on each interval (t_{i-1}, t_i).
- In case $Q(c) = Q(d)$ the curve is called a piecewise good loop.

All of the curves and parameterizations under consideration in this section will have these properties.

If P is another parameterization of $\mathfrak{C}$ belonging to the same orientation as Q we have:

$$\int_c^d F(Q(t)) \cdot Q'(t)\ dt = \int_a^b F(P(t)) \cdot P'(t)\ dt$$

If $\int_c^d F(Q(t)) \cdot Q'(t)\ dt = \int_a^b F(P(t)) \cdot P'(t)\ dt$ **whenever** $P(a) = Q(c)$ **and** $P(b) = Q(d)$ **even if they parameterize different curves** so long as both curves remain in $\mathfrak{O}$, the field F is called a **conservative vector field.**

This implies immediately that the circulation of F around any loop in $\mathfrak{O}$ is zero. (Show this.)

If for any pair of points C and D in $\mathfrak{O}$ there is at least one curve entirely in $\mathfrak{O}$, parameterizable by one of our piecewise good parameterizations Q, for which $Q(c) = C$ and $Q(d) = D$ the open set $\mathfrak{O}$ is called **path connected,** and a parameterization of a curve in $\mathfrak{O}$ that starts at one of the points and ends at the other is called a **path connecting the two points in $\mathfrak{O}$**. We won't prove it, but it is a fact that if open $\mathfrak{O}$ is path connected then you can actually connect any two points with a path consisting of a finite number of straight line segments, or by a path with a

good parameterization. We could construct such a curve, in the end, by patching together the Bezier curves we thought about in Section 21.

Suppose S is a particular point in path connected $\mathfrak{O}$ and the field F is conservative. We will define for generic T in $\mathfrak{O}$ the number $g(T) = \int_c^d F(Q(t)) \cdot Q'(t)\ dt$ when Q is any one of our piecewise good parameterizations chosen to start at S and end at T. Since the parameterization Q used is irrelevant for each T, and any T in $\mathfrak{O}$ can be connected to S by one of our parameterizations, we have created a function g defined on all of $\mathfrak{O}$.

If h is a function defined just as g was but by using a different starting point $\widetilde{S}$ then $h(T) = g(T) - g(\widetilde{S})$ for all T in $\mathfrak{O}$: that is, g and h differ by a constant. Also, if $\widetilde{T}$ is another point in $\mathfrak{O}$ then

$$g(T) - g(\widetilde{T}) = h(T) - h(\widetilde{T}).$$

We are going to show that g is continuously differentiable and $\nabla g = F$.

Since $\mathfrak{O}$ is open for each T in $\mathfrak{O}$ there is a (possibly tiny) disk or ball centered at T and entirely inside $\mathfrak{O}$. Choose ε small enough so that $T + \varepsilon\vec{i}$ is inside this disk or ball. So the parameterization $Q(t) = T + t\vec{i}$ is entirely inside $\mathfrak{O}$ for $-\varepsilon \le t \le \varepsilon$ and for each t, $Q'(t) = \vec{i}$.

Recall that $D_1 g(T) = \lim_{h\to 0} \frac{g(T+h\vec{i}\,) - g(T)}{h}$ when this limit exists. If we let $\widetilde{g}$ be the function defined just as g was but starting at $T - \varepsilon\vec{i}$ rather than S we have

$$\begin{aligned} g(T + h\vec{i}\,) - g(T) &= \widetilde{g}(T + h\vec{i}\,) - \widetilde{g}(T) \\ &= \int_{-\varepsilon}^{h} F(Q(t)) \cdot Q'(t)\ dt - \int_{-\varepsilon}^{0} F(Q(t)) \cdot Q'(t)\ dt \\ &= \int_0^h F(Q(t)) \cdot Q'(t)\ dt = \int_0^h F_1(Q(t))\ dt. \end{aligned}$$

The continuity of F implies that $D_1 g(T) = \lim_{h\to 0} \frac{g(T+h\vec{i}\,) - g(T)}{h} = F_1(T)$. Other partial derivatives are handled in the same way yielding $\nabla g = F$.

Any function g for which $\nabla g = F$ for any vector field F defined on any open set $\mathfrak{O}$ is called a **potential for** F. We have just shown how to construct many potentials when F is continuous and conservative and $\mathfrak{O}$ is path connected.

32.1. ***Exercise***. ** Show that if the domain of F is path connected, any two potentials for continuous conservative F differ by a constant.*

Looked at another way, let us suppose that g is any continuously differentiable real valued function defined on $\mathfrak{O}$. We will define the vector field F to be ∇g. Then F is conservative. This is, essentially, the chain rule. If S and T are in $\mathfrak{O}$ and Q is a

differentiable path in $\mathfrak{O}$ with $Q(c) = S$ and $Q(d) = T$ then the real valued function $g \circ Q$ is differentiable and $(g \circ Q)'(t) = \nabla g(Q(t)) \cdot Q'(t)$ and so

$$\begin{aligned} g(T) - g(S) &= g(Q(d)) - g(Q(c)) \\ &= \int_c^d (g \circ Q)'(t)\ dt = \int_c^d (\nabla g)(Q(t)) \cdot Q'(t)\ dt \\ &= \int_c^d F(Q(t)) \cdot Q'(t)\ dt. \end{aligned}$$

Since the left side depends only on the endpoints of the parameterization we can conclude (after some thought and the next exercise) that F is conservative.

32.2. ***Exercise***. *In the last paragraph we showed that F is independent of the path when the path is differentiable. What happens if the path is only "semi-nice" as in the rest of this section?*

So we have shown the following for open sets $\mathfrak{O}$:

- If g is a continuously differentiable real valued function on any open set $\mathfrak{O}$ then the field $F = \nabla g$ is a continuous conservative field.
- If F is a continuous conservative field on path connected $\mathfrak{O}$ then there is a family of potentials for F, any two of which differ by a constant.

It remains to be seen how one might compute a potential if there is one, or infer that a given field is not conservative. Here are a couple of examples.

First consider the field

$$F(X, Y, Z) = \langle F_1(X, Y, Z), F_2(X, Y, Z), F_3(X, Y, Z) \rangle = \left\langle 2XY, Z^2, 0 \right\rangle.$$

If $F = \nabla g$ then we must have $D_1 g(X, Y, Z) = 2XY$ and $D_2 g(X, Y, Z) = Z^2$. But these derivatives are themselves continuously differentiable so the mixed partials created from them should be equal, and they are not in this case. This field has no potential.

So when the components of F are continuously differentiable we have a necessary condition for the existence of a potential. We must have

$$D_2 F_1 = D_1 F_2 \text{ and } D_2 F_3 = D_3 F_2 \text{ and } D_3 F_1 = D_1 F_3.$$

Another approach with this same example would be to exhibit a path dependency of $\int_c^d F(Q(t)) \cdot Q'(t)\ dt$ for Q connecting two specific points. Almost any two paths will do. The path that moves on the coordinate axes from $\langle 1, 0, 0 \rangle$ to $\langle 0, 1, 0 \rangle$ yields 0, while the path that goes from $\langle 1, 0, 0 \rangle$ to $\langle 0, 1, 0 \rangle$ along the curve $\langle cos(t), sin(t), 0 \rangle$ yields the integral

$$\int_0^{\frac{\pi}{2}} \langle cos(t)sin(t), 0, 0 \rangle \cdot \langle -sin(t), cos(t), 0 \rangle\ dt = \int_0^{\frac{\pi}{2}} -cos(t)sin^2(t)\ dt = \frac{-1}{2}.$$

As a second example we look at the field

$$F(X, Y, Z) = \langle F_1(X, Y, Z), F_2(X, Y, Z), F_3(X, Y, Z) \rangle = \left\langle 2XY, X^2, 2Z \right\rangle.$$

$D_2F_1(X,Y,Z) = 2X = D_1F_2(X,Y,Z)$ and $D_3F_1(X,Y,Z) = D_1F_3(X,Y,Z) = 0 = D_2F_3(X,Y,Z) = D_3F_2(X,Y,Z)$. So at least there is the possibility that a potential exists.

If $\langle X,Y,Z\rangle$ is any point let $Q(t) = t\langle X,Y,Z\rangle$. So $Q'(t) = \langle X,Y,Z\rangle$. Also, $F(Q(t)) = t\left\langle 2t^2XY, t^2X^2, 2tZ\right\rangle$. Therefore

$$g(X,Y,Z) \;=\; \int_0^1 F(Q(t))\cdot Q'(t)\;dt \;=\; \int_0^1 3t^2X^2Y + 2tZ^2\;dt = X^2Y + Z^2.$$

We can check by differentiating that $\nabla g = F$.

To form the potential you create paths connecting a "center" point to a generic point of $\mathfrak{O}$ and integrate. Check that it actually is a potential for f by differentiating.

Another approach to this example is to integrate as follows: If a potential g exists then $D_1g(X,Y,Z) = 2XY$ so it must be that $g(X,Y,Z) = X^2Y + H$ where H is an unknown function that depends on Y and Z but not X. Similarly $D_2g(X,Y,Z) = X^2$ so $g(X,Y,Z) = X^2Y + K$ where K is a function of X and Z only. Comparing these implies that K must be a function of Z alone. Finally, $D_3g(X,Y,Z) = 2Z$ so $g(X,Y,Z) = Z^2 + W$ where W is a function of X or Y but not Z. Comparing these yields $g(X,Y,Z) = X^2Y + Z^2 + C$ for any constant C, and the fact that $\nabla g = F$ can be verified. This argument depends explicitly on the fact that the domain of F contains all line segments parallel to the coordinate axes connecting all points in the domain. (See Exercise 28.8.)

32.3. ***Exercise***. ** Suppose that $\mathfrak{O}$, the domain of F, contains all line segments parallel to the coordinate axes whenever the endpoints of these segments are in $\mathfrak{O}$. Suppose also that the components of F are themselves continuously differentiable and $D_2F_1 = D_1F_2$ and $D_2F_3 = D_3F_2$ and $D_3F_1 = D_1F_3$. Show that F has a potential.*

32.4. ***Exercise***. *Go over the details of the material of this section once more in the case where F is a vector field on an open set $\mathfrak{O}$ in $\mathbb{R}^2$ rather than $\mathbb{R}^3$.*

32.5. ***Exercise***. *An engineer is attempting to move a machine from one configuration to another. The desired change is equivalent to moving from (0,0,0) to (2,2,2) within the first octant of $\mathbb{R}^3$, where distances are in feet. The motion must take place along straight line segments, and at most one change in direction is allowed. During the motion various forces act in aid or opposition corresponding to a vector force field $F = -X\vec{i} - 2\vec{j} - 2\vec{k}$ pounds. In addition to work done against this force there is a work penalty of 1 foot pound for each foot traversed, making longer routes less attractive by that amount. How should the engineer proceed to minimize the work cost? How would the answer change if movement along any coordinate axis was "free?"*

CHAPTER V

Integration Involving Surfaces and Volumes May 27, 2005

33. Area and Integrals in the Plane

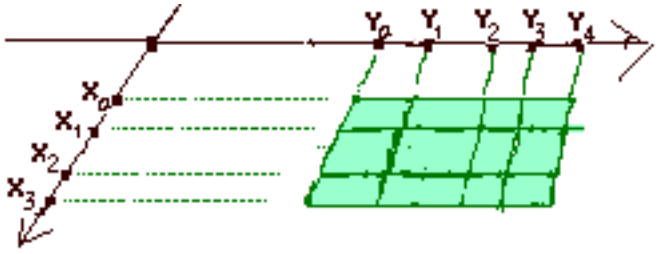
Suppose $[a, b]$ and $[c, d]$ are intervals. The **closed rectangle** formed from these intervals in the plane will be denoted $[a, b] \times [c, d]$, and consist of those ordered pairs (r, s) with $a \leq r \leq b$ and $c \leq s \leq d$.

A set $\mathfrak{O}$ is called **bounded** when there is a rectangle which contains $\mathfrak{O}$.

We are going to describe how to define and calculate a number that is reasonable to think of as the area of a bounded open set $\mathfrak{O}$ in the plane. We will also define integrals of bounded continuous functions defined on $\mathfrak{O}$ and indicate how they might be calculated. We will then consider integrals over certain closed sets.

Throughout these notes we have tried to outline justification for our results carefully, leaving only issues dependent on a careful construction of the Real Numbers for later classes. In this chapter we will try to be clear and precise but the detailed justification of a couple of our results is better left for later classes.[34] These include, in particular, common generalizations of the results we do prove. The student will need to revisit these issues several times. As mathematical maturity increases, he or she will become confused about increasingly deeper points. It is, however, quite possible to understand the main results of the section and apply them to practical problems without "dotting each i." We will provide numerous practical tools and survey the terrain in preparation for the more thorough second pass some readers will require.

Suppose $a = X_0 < \cdots < X_n = b$ is a partition of the interval $[a, b]$ and $c = Y_0 < \cdots < Y_m = d$ is a partition of the interval $[c, d]$. The collection P of closed rectangles formed from all the subintervals from consecutive partition members of $[a, b]$ and $[c, d]$ form what is called a **partition of the rectangle** $[a, b] \times [c, d]$. There are mn of these smaller rectangles. The **mesh** of this partition is the length of the longest edge of any rectangle in the partition.



A set of points C with members $C_{i,j}$ for $i = 1 \ldots n$ and $j = 1 \ldots m$ in the plane is called **subordinate to the partition** P if $C_{i,j}$ is in the subrectangle $[X_{i-1}, X_i] \times [Y_{j-1}, Y_j]$ for each i and j.

We suppose h is a **bounded continuous real valued function** defined on $\mathfrak{O}$.

Consider the sum $\sum_{\mathfrak{O}} h(C_{i,j}) \Delta X_i \Delta Y_j$ where this notation indicates that the sum is over those subscripts corresponding to rectangles in P which are entirely inside $\mathfrak{O}$. A sum of this kind, which depends on h, C and P, is called a **Riemann sum**.

It is a fact that under these conditions there is a number denoted

$$\int_{\mathfrak{O}} h(X,Y)\ dX\ dY$$

to which every Riemann sum is arbitrarily close provided only that the mesh of P is small enough. This number is called the **the double integral of h over $\mathfrak{O}$**.

If you are willing to accept this, or check the endnotes[35] for more, you are free to use a regular grid to form a partition and a uniform choice of points subordinate to the partition to calculate integrals.

For a positive integer n let $\mathbb{B}_n$ be the collection of all rectangles of the form $\left[\frac{i}{2^n}, \frac{i+1}{2^n}\right] \times \left[\frac{j}{2^n}, \frac{j+1}{2^n}\right]$ where i and j are any integers. Let $C^n_{i,j}$ denote the point $\left(\frac{i}{2^n}, \frac{j}{2^n}\right)$. So

$$\lim_{n\to\infty} \sum_{\mathbb{B}_n} \frac{f(C^n_{i,j})}{4^n} = \int_{\mathfrak{O}} h(X,Y)\ dX\ dY$$

where the sum is over all i and j for which the rectangle $\left[\frac{i}{2^n}, \frac{i+1}{2^n}\right] \times \left[\frac{j}{2^n}, \frac{j+1}{2^n}\right]$ is entirely in $\mathfrak{O}$.

The following are now fairly easy to show:

33.1. ***Exercise***. *If f and g are continuous and bounded on bounded open $\mathfrak{O}$ and c is a real number then*

$$\int_{\mathfrak{O}} f(X,Y)\ dX\ dY\ +\ c\ \int_{\mathfrak{O}} g(X,Y)\ dX\ dY\ =\ \int_{\mathfrak{O}} f(X,Y) + c\ g(X,Y)\ dX\ dY.$$

If $m \le f(P) \le M$ for all P in $\mathfrak{O}$ for constants m and M then

$$\int_{\mathfrak{O}} m\ dX\ dY \le \int_{\mathfrak{O}} f(X,Y)\ dX\ dY \le \int_{\mathfrak{O}} M\ dX\ dY.$$

Also, if $f \ge g$ and $f(P) > g(P)$ for even one point P in $\mathfrak{O}$ then

$$\int_{\mathfrak{O}} f(X,Y)\ dX\ dY > \int_{\mathfrak{O}} g(X,Y)\ dX\ dY.$$

*** If f is nonnegative and the sets $\mathfrak{O}$ and $\mathfrak{U}$ are both open and contained in the domain of f then the union of these two sets, $\mathfrak{O} \cup \mathfrak{U}$, is an open set and*

$$\int_{\mathfrak{O}\cup\mathfrak{U}} f(X,Y)\ dX\ dY \le \int_{\mathfrak{O}} f(X,Y)\ dX\ dY + \int_{\mathfrak{U}} f(X,Y)\ dX\ dY.$$

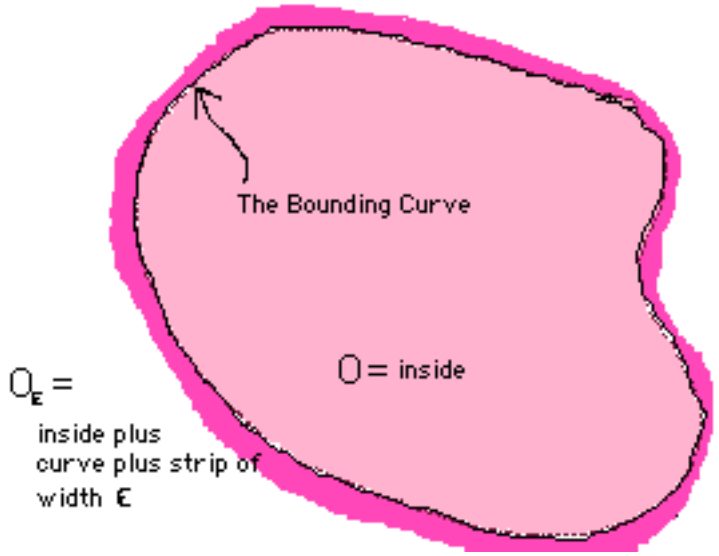
If, further, $\mathfrak{O} \cap \mathfrak{U} = \varnothing$, equality holds in the last line.

Quite often the region $\mathfrak{O}$ from above arises as the bounded part of the plane surrounded by a piecewise good loop $\mathfrak{C}$. To make this precise we need the concept of "boundary."

A point P is said to be on the **boundary** of a set $\mathcal{A}$ of points in the plane provided that every disk centered at P contains at least one point in $\mathcal{A}$ and a point not in $\mathcal{A}$.

In our case we are saying that $\mathcal{O}$ is a bounded open set and $\mathcal{C}$ consists precisely of the points on the boundary of $\mathcal{O}$.

When $\mathcal{O}$ arises in this way we let $\overline{\mathcal{O}}$ be the union of $\mathcal{O}$ and $\mathcal{C}$. $\overline{\mathcal{O}}$ is a closed set. (Can you show this?)



We define the set $\mathcal{C}_\varepsilon$ to be those points in the plane less than a distance ε away from **some** point on the bounding curve $\mathcal{C}$ and let $\mathcal{O}_\varepsilon$ be those points in the union of $\mathcal{O}$ and $\mathcal{C}_\varepsilon$.

$\mathcal{C}$ has a finite length L and an easy estimate shows that the open set $\mathcal{C}_\varepsilon$ can be covered by (roughly) $\frac{4L}{\varepsilon}$ squares each of area ε^2. So any reasonable measure of the area of $\mathcal{C}_\varepsilon$ cannot exceed $4L\varepsilon$.

So if h is continuous on some $\mathcal{O}_\varepsilon$ and $|h|$ is bounded by M then no matter how tiny ε might be,

$$\left| \int_{\mathcal{O}_\varepsilon} h(X,Y)\ dX\ dY - \int_{\mathcal{O}} h(X,Y)\ dX\ dY \right| \ \leq\ 4ML\varepsilon.$$

Since $\overline{\mathcal{O}}$ consists of exactly those points in the plane in **every** $\mathcal{O}_\varepsilon$ it makes sense to define:

$$\int_{\overline{\mathcal{O}}} h(X,Y)\ dX\ dY = \lim_{\varepsilon\to 0} \int_{\mathcal{O}_\varepsilon} h(X,Y)\ dX\ dY = \int_{\mathcal{O}} h(X,Y)\ dX\ dY.$$

So we have defined integrals on certain types of closed sets too: namely, closed sets which are the bounded region inside a piecewise good loop and for bounded functions continuous on some open set containing this closed set.

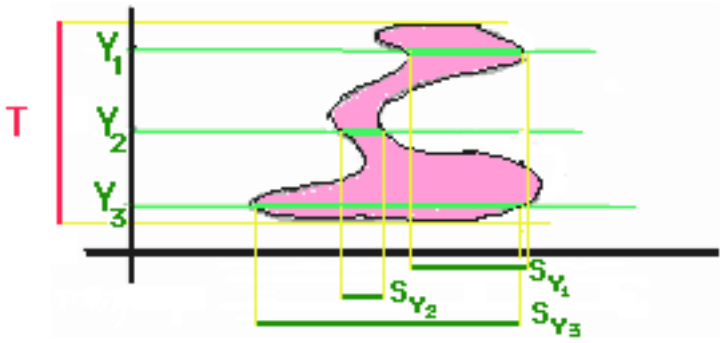
We now come to the issue of how one might actually calculate an integral.

For each Y in $[c,d]$ define S_Y to be the set of those X in $[a,b]$ for which (X,Y) is in $\mathcal{O}$. Each S_Y is an open set and the function $\mathcal{A}_Y$ defined on S_Y by $\mathcal{A}_Y(X) = h(X,Y)$ is bounded and continuous.

In most of the integrals you will run across, each set S_Y is empty (nothing in it) or a finite union of open intervals, and if you wish you can throw in the endpoints of these intervals without changing any of the integrals $\int_{S_Y} \mathcal{A}_Y(X)\, dX$. If $\mathcal{A}_Y(X)$ is not actually defined at an endpoint, you could still define the integral on these closed intervals as an **improper integral**, using a limiting procedure. This is not usually necessary.

Define $\mathcal{B}(Y) = \int_{S_Y} \mathcal{A}_Y(X)\, dX$ for each Y in $[c, d]$, where if S_Y is empty this number is 0.

Let T be the set of those Y in $[c, d]$ for which (X, Y) is in $\mathcal{O}$ for any X. T is open.



The function $\mathcal{B}$ is continuous and bounded on T. It is a fact that under these conditions

$$\int_{\mathcal{O}} h(X, Y)\, dX\, dY = \int_T \mathcal{B}(Y)\, dY = \int_T \left(\int_{S_Y} \mathcal{A}_Y(X)\, dX \right) dY.$$

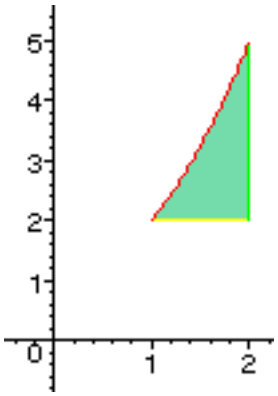
The last integral is called the **iterated integral of** h on $\mathcal{O}$, and is the main tool used to actually calculate a double integral on an open set in the plane, or a closed set surrounded by a piecewise good loop. Under the conditions of this section, iterated integrals can be calculated in either order, by modifying the definition slightly to integrate first with respect to one or the other variable. The theorem which identifies the double integral with the iterated integral in either order is called **Fubini's Theorem,** and is very important. You can find a discussion of the proof of this theorem in an endnote.[36]

When $h = 1$, the constant function, this integral is to be interpreted as the **area of** $\mathcal{O}$. If h is nonnegative, one could interpret the integral as the **volume** between the surface formed as the graph of h and the set corresponding to $\mathcal{O}$ in the XY plane, or the **mass** of a plate in the shape of $\mathcal{O}$ with density function h. For a general h the integral could represent total **charge** on a plate in the shape of $\mathcal{O}$ with charge density h.

33.2. ***Exercise***. *Suppose we have a plate in the shape of the region* $\mathcal{O}$ *in the first quadrant of the* XY *plane bounded by* $X = 1$, $X = 2$, $Y = 2$ *and* $Y = X^2 + 1$ *where* X *and* Y *are measured in meters. This plate has variable density given by*

$h(X,Y) = Y + X^2$ kilograms per square meter on this region. We calculate the mass of this plate to be:

$$\begin{aligned}
&\int_{\mathfrak{O}} h(X,Y)\ dX\ dY \\
&\quad = \int_{X=1}^{X=2} \left(\int_{Y=2}^{Y=X^2+1} Y + X^2\ dY \right)\ dX \\
&\quad = \int_{X=1}^{X=2} \left(\frac{Y^2}{2} + X^2 Y \Big|_{Y=2}^{Y=X^2+1} \right)\ dX \\
&\quad = \int_{X=1}^{X=2} \frac{3X^4}{2} - \frac{3}{2}\ dX \\
&\quad = \frac{3X^5}{10} - \frac{3X}{2}\Big|_{X=1}^{X=2} = 7.8 \text{ kilograms.}
\end{aligned}$$



Calculate this iterated integral in the "other order," first with respect to X and then with respect to Y.

33.3. ***Exercise***. *Calculate the volume of the part of the unit sphere in the first octant. This surface is the graph of $g(X,Y) = \sqrt{1 - X^2 - Y^2}$ where X and Y are both positive. (hint: Consider $\int_{\mathfrak{O}} g(X,Y)\ dX\ dY$. Calculate the iterated integral first with respect to X integrating from $X = 0$ to $X = \sqrt{1-Y^2}$. Then use the substitution $X = \sqrt{1-Y^2}sin(u)$.)*

33.4. ***Exercise***. ** Suppose f is defined on a bounded open set $\mathfrak{O}$ in the plane and the mixed partial derivatives $D_{1,2}f$ and $D_{2,1}f$ of f exist and are continuous in $\mathfrak{O}$. Use Fubini's Theorem to show that $D_{1,2}f = D_{2,1}f$ in $\mathfrak{O}$. (Hint: If they differ at a point P, say $D_{1,2}f(P) > D_{2,1}f(P)$, then by continuity there is some little rectangle and $\varepsilon > 0$ for which $D_{1,2}f > D_{2,1}f + \varepsilon$ everywhere on the rectangle.)*

---

34. Area of a Curved Surface and Surface Integrals

We will discuss an application of these results to define surface area on the surface formed as the graph of differentiable g with continuous gradient and also integrals of functions defined on such a surface.

We will start out with the part of the graph consisting of points $\langle X, Y, g(X,Y)\rangle$ for $\langle X, Y\rangle$ in the open set $\mathfrak{O}$ contained in the rectangle $[a,b] \times [c,d]$. We will often visualize the set $\mathfrak{O}$ as being in the XY plane in $3D$ beneath the graph of g though, of course, that last set, the "shadow" of the graph, consists of points with three components with 0 third component while the points in $\mathfrak{O}$ have only two.

∇g will appear in many formulas in this chapter, so we will often denote $D_1 g$ by g_1 and $D_2 g$ by g_2. So $\nabla g = \langle g_1,\ g_2 \rangle$.

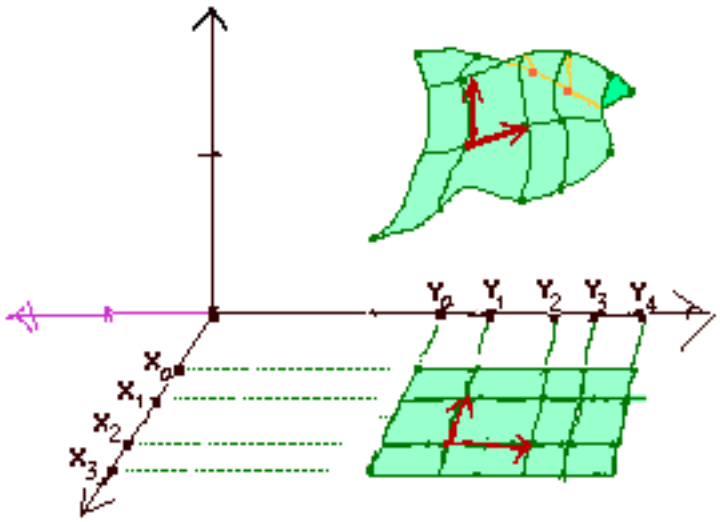
At each (α, β) in $\mathfrak{O}$ the functions B_α and A_β defined by $B_\alpha(Y) = \langle \alpha, Y, g(\alpha, Y) \rangle$ and $A_\beta(X) = \langle X, \beta, g(X, \beta) \rangle$ are differentiable parameterizations (over possibly small intervals) of curves through $\langle \alpha, \beta, g(\alpha, \beta) \rangle$ and these curves are in the surface.

$B'_\alpha(\beta) = \langle 0, 1, g_2(\alpha, \beta) \rangle$ and $A'_\beta(\alpha) = \langle 1, 0, g_1(\alpha, \beta) \rangle$ are in the tangent plane to the surface at $\langle \alpha, \beta, g(\alpha, \beta) \rangle$.

The vector $K(\alpha, \beta) = A'_\beta(\alpha) \times B'_\alpha(\beta) = \langle -g_1(\alpha, \beta), -g_2(\alpha, \beta), 1 \rangle$ is normal to the surface at $\langle \alpha, \beta, g(\alpha, \beta) \rangle$. The vector $\mathfrak{N}(\alpha, \beta) = K(\alpha, \beta)/|K(\alpha, \beta)|$ is the (upward) unit normal to the surface at that spot.

Suppose P is a partition of the rectangle $[a, b] \times [c, d]$ containing $\mathfrak{O}$ and $[X_{i-1}, X_i] \times [Y_{j-1}, Y_j]$ is a rectangle from this partition inside $\mathfrak{O}$. The number $\Delta X_i \Delta Y_j$ is the area of this (presumably tiny) rectangle in the XY plane which can be thought of as beneath or above a roughly polygonal patch on the surface.

The line segments $\langle X_i, s \rangle$ for $i = 0 \ldots n$ and $u \le s \le v$, and $\langle t, Y_j \rangle$ for $j = 0 \ldots m$ and $a \le t \le b$ in the XY plane break up the big rectangle into the little ones. If you imagine a light casting a shadow vertically past these segments thought of as in the XY plane onto the underside of the surface of interest, the shadows will break the surface into patches which look like little parallelograms.



When $[X_{i-1}, X_i] \times [Y_{j-1}, Y_j]$ is contained in $\mathfrak{O}$, parameterizations for the shadows on the surface from the four edges of this rectangle are given by $B_{X_k}(s)$ for $k = i - 1$ and i and $A_{Y_k}(t)$ for $k = j - 1$ and j and where the parameters s and t extend (at least a little) beyond the intervals $[Y_{j-1}, Y_j]$ and $[X_{i-1}, X_i]$ respectively.

The tangent vectors to the two curves which cross at the corner

$$\langle X_{i-1}, Y_{j-1}, g(X_{i-1}, Y_{j-1}) \rangle \text{ are}$$

$$T^X_{i,j} = A'_{Y_{j-1}}(X_{i-1}) \text{ and } T^Y_{i,j} = B'_{X_{i-1}}(Y_{j-1}).$$

The two vectors $\Delta X_i T^X_{i,j}$ and $\Delta Y_j T^Y_{i,j}$ lie on the shadows of the edges of a parallelogram on the tangent plane at the corner point $\langle X_{i-1}, Y_{j-1}, g(X_{i-1}, Y_{j-1}) \rangle$. To the extent that the tangent plane is glued tightly to the surface the area of this parallelogram would be a good candidate for an approximation to the area of the nearby surface patch.

The area of the parallelogram is the magnitude of

$$\Delta X_i T^X_{i,j} \times \Delta Y_j T^Y_{i,j} = \Delta X_i \Delta Y_i \left\langle -g_1(X_{i-1}, Y_{j-1}), -g_2(X_{i-1}, Y_{j-1}), 1 \right\rangle .$$

This magnitude is

$$\Delta X_i \Delta Y_j \sqrt{(g_1(X_{i-1}, Y_{j-1}))^2 + (g_2(X_{i-1}, Y_{j-1}))^2 + 1} \ .$$

Because the gradient of g is continuous, the area of the sum of all these patches is nearly (if the mesh of P is small enough):

$$\int_{\mathfrak{O}} \sqrt{(g_1(X,Y))^2 + (g_2(X,Y))^2 + 1} \ dX \ dY \ .$$

Another way of getting at this is through the normal vector. The upward unit normal vector to the surface at $\langle X_{i-1}, Y_{j-1}, g(X_{i-1}, Y_{j-1}) \rangle$ is

$$\mathbf{N}_{i,j} = \frac{\langle g_1(X_{i-1}, Y_{j-1}), g_2(X_{i-1}, Y_{j-1}), -1 \rangle}{\sqrt{(g_1(X_{i-1}, Y_{j-1}))^2 + (g_2(X_{i-1}, Y_{j-1}))^2 + 1}}$$

and the normal vector to the XY plane is $\vec{k}$. We saw in Section 12 that the areas of the slanted and shadow parallelograms were related by

$$\text{Slanted Area (On the Tangent Plane)} = \frac{1}{cos(\theta)} \text{Shadow Area (On the XY Plane).}$$

where θ is the angle between the surface normal and $\vec{k}$.

We have

$$\mathbf{N}_{i,j} \cdot \vec{k} = |\mathbf{N}_{i,j}||\vec{k}| cos(\theta) = cos(\theta)$$

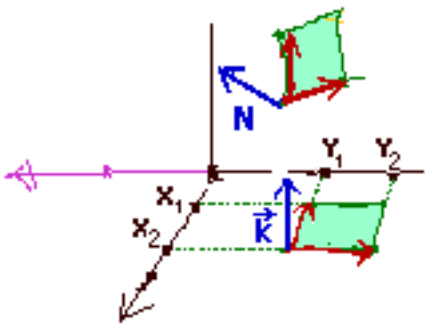
which gives

$$cos(\theta) = \frac{1}{\sqrt{(g_1(X_{i-1}, Y_{j-1}))^2 + (g_2(X_{i-1}, Y_{j-1}))^2 + 1}}.$$

The area of the rectangular bit in the XY plane is $\Delta X_i \Delta Y_j$ so the slant area is, once again,

$$\Delta X_i \Delta Y_j \sqrt{(g_1(X_{i-1}, Y_{j-1}))^2 + (g_2(X_{i-1}, Y_{j-1}))^2 + 1}$$

which generates the same integral as before.



The discussion above constitutes justification for why we think of that integral as the area of the curved surface. In the end, we simply **define area of a piece of surface** to **be** that integral. It is up to the user to decide if it reproduces intuition about what surface area should be. In particular, if you have multiple methods for calculating or estimating numbers each of which **should** be the area, you would want them all, at minimum, to agree.

Finally, suppose h is a continuous function defined on $\mathfrak{O}$. Each point on the surface is associated with a value of the function h. With this interpretation we call the integral

$$\int_{\mathfrak{O}} h(X,Y)\sqrt{(g_1(X,Y))^2+(g_2(X,Y))^2+1}\ dX\ dY$$

the **surface integral of h on this surface.** By analogy with line integrals, we could call this the **integral of h weighted by surface area.**

We could think of h as representing, for example, mass density (mass per unit area) or charge density up on the surface. In those cases the surface integral would correspond to total **mass** or **charge** on the piece of surface above $\mathfrak{O}$.

Frequently a density is initially given as a function k of the three coordinates (X,Y,Z) on the surface. Since $Z = g(X,Y)$ this gives the density $h(X,Y) = k(X,Y,g(X,Y))$ as a function of two variables. Also, given h you can define k by the same formula. The same integral is also called the **surface integral of k on this surface** or the **integral of k weighted by surface area.** There is only a slight philosophical difference between h and k: in the first case you are thinking of density as a function on the domain of g, while in the second k gives the density as an explicit function of the three coordinates of the point on the surface.

One often sees various shorthand forms for integrals of this kind, whose purpose is to suppress the specific variable names and make the formulas quicker to write. Laudable as a goal, this has the effect of separating the user from the calculation, or confusing the user as to which symbols are variables. Be cautious.

For example if you denote $\nabla g(X,Y)\cdot\nabla g(X,Y)$ by $(\nabla g)^2$ the integral formula from above becomes:

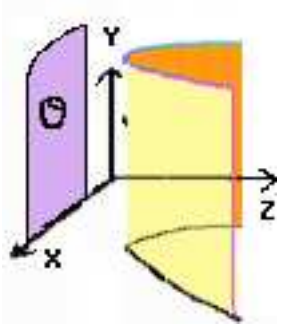
$$\int_{\mathfrak{O}} h\ \sqrt{(\nabla g)^2 + 1}\ dX\ dY.$$

34.1. ***Exercise***. *Use a surface integral to calculate the surface area of the part of the unit sphere in the first octant. This surface is the graph of* $g(X,Y) = \sqrt{1 - X^2 - Y^2}$ *where* X *and* Y *are both positive. (hint: Use the same substitution as in Exercise 33.2.)*

Suppose that this piece of surface has density which varies with height: the density of a part of the surface at height Z *is* $3Z$ *kilograms per unit area. What is the mass of this surface?*
(hint: Consider $\int_{\mathfrak{O}} 3g(X,Y)\sqrt{(g_1(X,Y))^2 + (g_2(X,Y))^2 + 1}\ dX\ dY.$ *)*

34.2. ***Exercise***. *Use a graphing utility such as Maple to examine the surface which is the graph of* $g(X,Y) = cos(X)cos(Y)+3$ *on the square* $[0, 4\pi]\times[0, 4\pi]$. *Set up an integral for the surface area of this surface, think about it until you can get a rough estimate of what the area should be (within* 50 *percent) and then integrate numerically (once again using Maple or Mathematica) to obtain the area integral to greater accuracy.*[37]

34.3. ***Exercise***.



Consider an open set $\mathfrak{O}$ *in the* XY *plane and a positive function* Z *defined on* $\mathfrak{O}$ *depending only on* X*: that is,* $Z(X,Y_1) = Z(X,Y_2)$ *whenever* (X,Y_1) *and* (X,Y_2) *are in* $\mathfrak{O}$.

Prove the following statements:

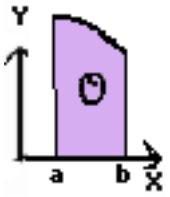
(i) The (upward) unit normal to the surface formed as the graph of Z *is*

$$\frac{\langle 1, 0, D_1 Z\rangle \times \langle 0, 1, 0\rangle}{\sqrt{1 + (D_1 Z)^2}} = \frac{\langle -D_1 Z, 0, 1\rangle}{\sqrt{1 + (D_1 Z)^2}}.$$

(ii) The cosine of the angle this vector makes with the XY *plane is* $\frac{1}{\sqrt{1+(D_1 Z)^2}}$.

(iii) The surface area of this graph is $\int_{\mathfrak{O}} \sqrt{1 + (D_1 Z)^2}\ dY\ dX$.

Suppose that the region $\mathfrak{O}$ *is that trapped inside* $X = a$, $X = b$, $Y \geq 0$ *and* $Y \leq Y(X)$ *where* $Y(X)$ *is a positive continuous function on* $[a, b]$.



So the integral from (iii) becomes:

$$\int_{X=a}^{X=b}\int_{Y=0}^{Y=Y(X)}\sqrt{1+(D_1Z)^2}\,dY\,dX=\int_{X=a}^{X=b}Y(X)\sqrt{1+(D_1Z)^2}\,dX.$$

(v) Note that the upper limit of the region is the graph of $Q(X)=\langle X,Y(X)\rangle$. *Compare this result to the one on page 87.*

---

35. Parametric Descriptions of a Plane Set

We will presume that $\mathcal{W}$ is an open set in the plane and f is a vector valued function with domain $\mathcal{W}$, $f(s,t)=\langle X(s,t),Y(s,t)\rangle$. Wherever it is convenient we will write $X_i(s,t)$ in place of $D_iX(s,t)$ and $Y_i(s,t)$ in place of $D_iY(s,t)$. The only other places in this chapter where subscripts are encountered on X or Y is in the definition of the Riemann sums used to form integrals, and context will make the distinction obvious.

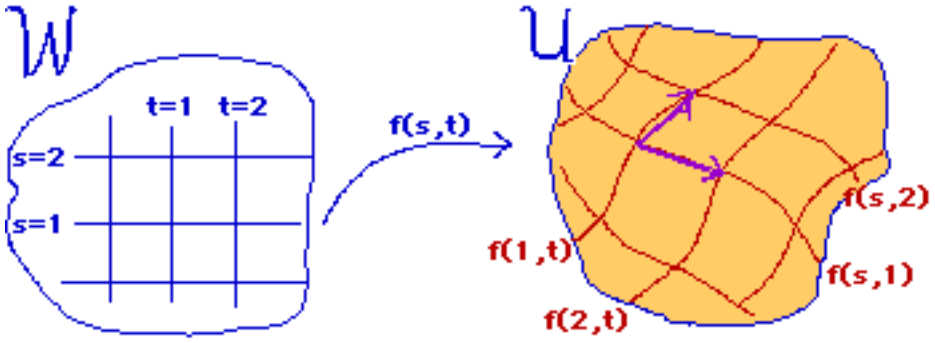
We adopt the notation $A_t(s)=f(s,t)=B_s(t)$.

We will make a number of "niceness" presumptions about f and leave for later classes the extent to which some of these assumptions are redundant or unnecessarily restrictive.[38]

(i) We presume that f is one-to-one and that the collection of values $f(s,t)$ for (s,t) in $\mathcal{W}$ constitute an open subset $\mathcal{U}$ in the plane.

(ii) The functions A_t for each t and B_s for each s are continuously differentiable.

(iii) The vectors $A_t'(s)$ and $B_s'(t)$ are never 0, nor is one a multiple of the other for any fixed (s,t).



We will refer to an f like this as a **good parameterization** of the open plane set $\mathcal{U}$.

The curves parameterized by A_t and B_s constitute a **coordinate grid** on $\mathcal{U}$, and each point in $\mathcal{U}$ can be located uniquely as the crossing place of two of these curves. These curves act as alternative coordinate lines on $\mathcal{U}$, and a mesh of closely spaced grid curves helps one navigate around on $\mathcal{U}$ just as the ordinary rectangular grid would. Small pieces of $\mathcal{U}$ bounded by the grid curves will not be rectangles necessarily but will, when small enough, look like parallelograms.

The condition (iii) ensures that the parallelograms in the grid do not get too "skinny" or, at least, never go entirely "flat."

35.1. ***Exercise***. *Verify:*

Condition (ii) implies that f is continuously differentiable.

In fact

$$A'_t(s) = \big\langle X_1(s,t), Y_1(s,t)\big\rangle$$
$$\text{and } B'_s(t) = \big\langle X_2(s,t), Y_2(s,t)\big\rangle.$$

The area of the parallelogram formed by $A'_t(s)$ and $B'_s(t)$ is

$$|X_1(s,t)Y_2(s,t) - X_2(s,t)Y_1(s,t)|$$

and this magnitude is a continuous function on $\mathcal{W}$.

Condition (iii) and continuity of the derivatives ensures that the sign of $X_1Y_2 - X_2Y_1$ is constant on $\mathcal{W}$, a fact that will be useful in calculations.

For those of you who know about determinants,

$$X_1Y_2 - X_2Y_1 = \det\ f',$$

a fact that shortens notation here and there.

35.2. ***Exercise***. ** Show that if f is a good parameterization of $\mathcal{U}$ then $g = f^{-1}$ is a good parameterization of $\mathcal{W}$.*

---

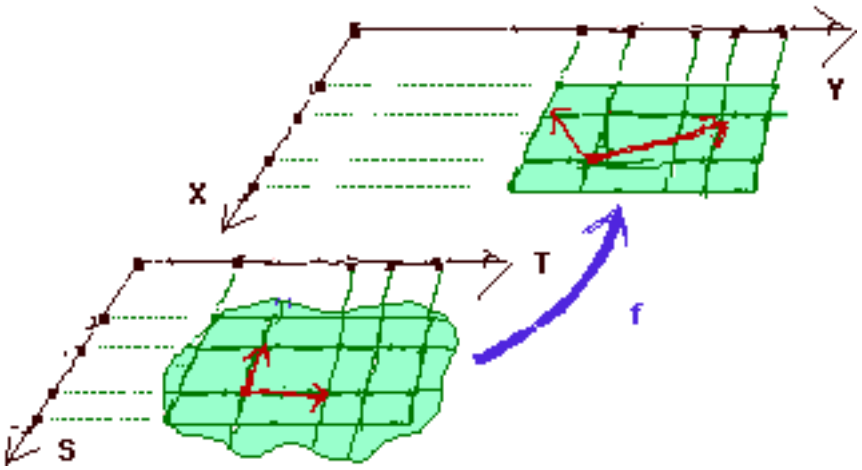
36. Change of Variable in the Plane

We will use the earlier ideas about integrals together with the notation from the last section on parametric descriptions of plane sets to create an integral formula involving the parameterization. This is called a **change of variables** formula.

We presume that f is a good parameterization of $\mathcal{U}$ defined on $\mathcal{W}$. We will presume that the domain $\mathcal{W}$ of f is inside the rectangle $[a,b]\times[c,d]$ in the plane. Suppose P is a partition of this rectangle composed of subrectangles $[s_{i-1}, s_i]\times[t_{j-1}, t_j]$ for$i = 1\ldots n$ and $j = 1\ldots m$.

Suppose $[s_{i-1}, s_i]\times[t_{j-1}, t_j]$ is entirely contained in $\mathcal{W}$.

$\Delta t_j B'_{s_{i-1}}(t_{j-1})$ and $\Delta s_i A'_{t_{j-1}}(s_{i-1})$ are both vectors in the plane.



If the mesh of the partition is small enough, the parallelogram formed by these two vectors in the plane will be very close to the small part of $\mathcal{U}$ covered by f on the rectangle $[s_{i-1}, s_i] \times [t_{j-1}, t_j]$. This parallelogram has area

$$\Delta s_i \Delta t_j |X_1(s_{i-1}, t_{j-1}) Y_2(s_{i-1}, t_{j-1}) - X_2(s_{i-1}, t_{j-1}) Y_1(s_{i-1}, t_{j-1})|.$$

Suppose h is a continuous and bounded function defined on $\mathcal{U}$. We can imagine that h represents **charge** (or **mass**) density, for example, at each point in $\mathcal{U}$. h will be nearly constant on this tiny piece of $\mathcal{U}$, so the total charge on this piece will be nearly the area of the parallelogram multiplied by $h(X(s_{i-1}, t_{j-1}), Y(s_{i-1}, t_{j-1}))$.

Adding together these numbers for all the partition rectangles entirely contained in $\mathcal{W}$ generates a Riemann sum which is an approximation to the total charge on $\mathcal{U}$ and yields the integral

$$\int_{\mathcal{W}} h(X(s,t), Y(s,t)) |X_1(s,t) Y_2(s,t) - X_2(s,t) Y_1(s,t)| \, ds \, dt$$

which **should**, if our thinking is correct, be the same as the double integral

$$\int_{\mathcal{U}} h(X, Y) \, dX \, dY.$$

In this context, this integral on $\mathcal{W}$ is called a **parametric form for the integral of h on $\mathcal{U}$**.

Though the discussion above makes the equality of these two integrals very plausible, we have not proved that the two integrals agree. That must be, alas, reserved for a more advanced treatment elsewhere.

When $h = 1$, the constant function, we have a representation of the **area** of $\mathcal{U}$ as $\int_{\mathcal{W}} |X_1(s,t) Y_2(s,t) - X_2(s,t) Y_1(s,t)| \, ds \, dt$.

It is common to see the collapsed notation

$$\int_{\mathcal{W}} h \circ f \; |det \; f'| \; ds \; dt \; = \; \int_{\mathcal{U}} h \; dX \; dY$$

to express the equality of these two integrals. This notation is fine as long as you don't lose track of which functions involve which variables.

36.1. ***Exercise***. *Consider the function* $f(r,\theta) = \langle rcos(\theta), rsin(\theta)\rangle$ *with domain* $\mathcal{W} = (0,1) \times (0, \frac{\pi}{2})$ *and with range equal to* $\mathcal{U}$*, the "all positive" quadrant of the unit disk.* f *is called* ***polar coordinates*** *on* $\mathcal{U}$*.*

Show that f *is a "good parameterization" of* $\mathcal{U}$*.*

Show that the coordinate grid curves cross at right angles to each other. Coordinate systems like this are called ***orthogonal****.*

Calculate the area of the quarter disk in two ways: directly as an integral over the set $\mathcal{U}$ *and by changing coordinates as an integral over the set* $\mathcal{W}$*.*

Try to calculate

$$\int_{\mathcal{U}} e^{-X^2-Y^2} \; dX \; dY$$

directly and then change variables to polar coordinates and try again.

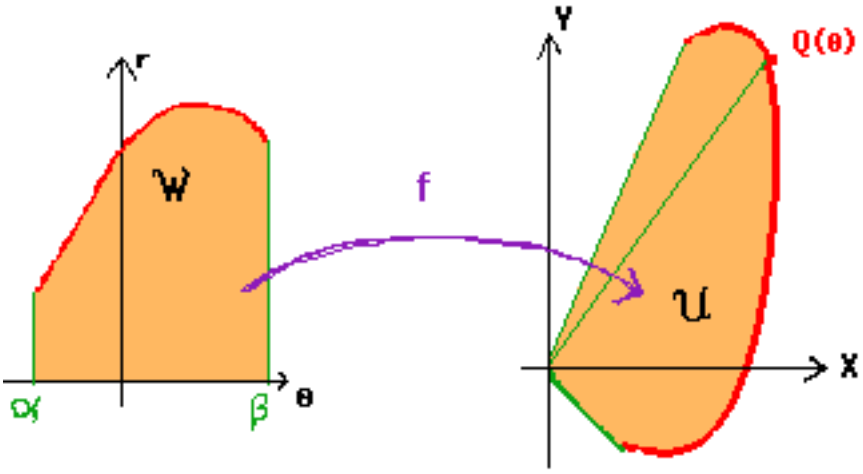
Finally, suppose that your goal is to calculate an ***improper integral*** *where* $\mathcal{U}$ *is not the quarter disk but the entire positive quadrant of the plane. How would you define polar coordinates on this new* $\mathcal{U}$*? How might you define an improper integral for this unbounded domain?*

What is $\int_{\mathcal{U}} e^{-X^2-Y^2} \; dX \; dY$ *over this unbounded region?*

36.2. ***Exercise***. *Recall our discussion of area using polar coordinates from page 99. In that section we had a curve* $Q(\theta) = \langle r(\theta)cos(\theta), r(\theta)sin(\theta)\rangle$ *defined on an interval* $[\alpha, \beta]$*. To simplify things we presume that* $r(\theta) > 0$ *and* $-\pi < \alpha < \beta < \pi$*. We wanted to calculate the area* $\mathcal{U}$ *inside the curve and bounded by lines of constant angle. After breaking the region into many narrow triangles, we concluded that the* ***area*** *should be* $\int_\alpha^\beta \frac{1}{2} r^2 \; d\theta$*.*

Define polar coordinates by $f(r,\theta) = \langle rcos(\theta), rsin(\theta)\rangle$*. If* $\mathcal{W}$ *is the region under the graph of* r *then* f *is a "good parameterization" taking points in* $\mathcal{W}$ *to points in* $\mathcal{U}$*.*

So the pie shaped area should be $\int_{\mathcal{W}} |det \; f'| \; dr \; d\theta$*.*



(i) Show that this gives the same number (the area) as before.

(ii) Can this setup be modified to allow negative values for r?

(iii) Modify the problem (that is, reformulate the question) to deal with curves that wind around the origin more than once.

36.3. ***Exercise***. *Consider the function $f(s,t) = \langle X, Y\rangle = \langle st, t^2 - s^2\rangle$ with domain $\mathcal{W} = (0,\infty)\times(0,\infty)$.*

f is a type of ***parabolic coordinates.***

(i) Show that f is one-to-one with range equal to $\mathcal{U}$, the $X > 0$ half of the plane.

(ii) Draw a sketch containing three constant-t and three constant-s grid curves. Show that these parabolic coordinates form a good coordinate system and an orthogonal coordinate system. Why are these called "parabolic" coordinates?

(iii) Set up a double integral for the area of the region $\mathcal{U}_1$ in the plane beneath $Y = -X^2 + 1$ and above $Y = X^2 - 1$ and with $X > 0$. Calculate the area of this region. Then identify the region $\mathcal{W}_1$ taken by parabolic coordinates to $\mathcal{U}_1$ and calculate the area a second time using parabolic coordinates and the change of variable integration formula.

You will notice in the last exercise that the difficulty in the integrals (minor though they were) changed: in one form the integrand was simple while the limits of integration were more complicated; in the other the limits of integration were trivial at the price of a more complicated integrand. In any case, the calculations were **different.** That is the point of changing variables. In some problems there is a huge advantage to be had by considering symmetries in integrand or region and finding or inventing coordinates to match.

36.4. ***Exercise***. *Consider the function*

$$f(s,t) = \langle X, Y\rangle = \langle cosh(s)cos(t), sinh(s)sin(t)\rangle$$

with domain $\mathcal{W} = (0,\infty)\times(0,\pi)$. *You will recall that* $cosh^2(t) - sinh^2(t) = 1$ *for all* t. *Also note that for any specific* $f(s,t) = \langle X, Y\rangle$ *in the range of* f *with* $X \neq 0$ *that*

$$\frac{X^2}{cos^2(t)} - \frac{Y^2}{sin^2(t)} = 1 \quad \textit{and} \quad \frac{X^2}{cosh^2(s)} + \frac{Y^2}{sinh^2(s)} = 1.$$

We will refer to f *as* ***hyperbolic-elliptic coordinates.***[39] *Why is that a reasonable name for these coordinates?*

(i) Show that f *is one-to-one with range equal to* $\mathcal{U}$, *the half of the plane corresponding to* $Y > 0$.

(hint: Show first that if $\langle X, Y\rangle$ *is in the range of* f *with* $X \neq 0$ *and both*

$$\frac{X^2}{cos^2(t_1)} - \frac{Y^2}{1-cos^2(t_1)} = 1 \quad \textit{and} \quad \frac{X^2}{cos^2(t_2)} - \frac{Y^2}{1-cos^2(t_2)} = 1$$

then $t_1 = t_2$.*)*

(ii) Show that these coordinates form a good coordinate system and an orthogonal coordinate system.

(iii) Note that the upper half-ellipse

$$Y = \sqrt{\left(\frac{15}{8}\right)^2 - \left(\frac{15}{17}\right)^2 x^2}$$

corresponds to the gridline $s = ln(4)$. *Set up a double integral for the area of the region* $\mathcal{U}_1$ *in the upper half-plane inside this ellipse in terms of the* X *and* Y *coordinates. Calculate the area of this region. Then identify the region* $\mathcal{W}_1$ *taken by hyperbolic-elliptic coordinates to* $\mathcal{U}_1$ *and calculate the area a second time using hyperbolic-elliptic coordinates and the change of variable integration formula.*[40]

37. Parametric Descriptions of a Surface

We are going to extend the discussion somewhat to treat surfaces given as the graph of a differentiable function g of two variables as in Section 34. In that section the X and Y coordinates of a point on the graph determine the point uniquely and explicitly - the X and Y coordinates were the parameters. We can (and did) visualize this surface as being above or below its domain, the open set $\mathcal{O}$, thought of as the shadow of the surface on the XY plane in $3D$.

Often surfaces are parameterized by alternative means.

We will presume that $\mathcal{W}$ is an open set in the plane and f is a vector valued function with domain $\mathcal{W}$, $f(s,t) = \langle X(s,t), Y(s,t), Z(s,t)\rangle$. f is a vector valued function on a plane set but whose values are $3D$ vectors.

We adopt the notation $B_s(t) = f(s,t) = A_t(s)$ just as before.

We make, as before, a number of "niceness" presumptions[41] about f.

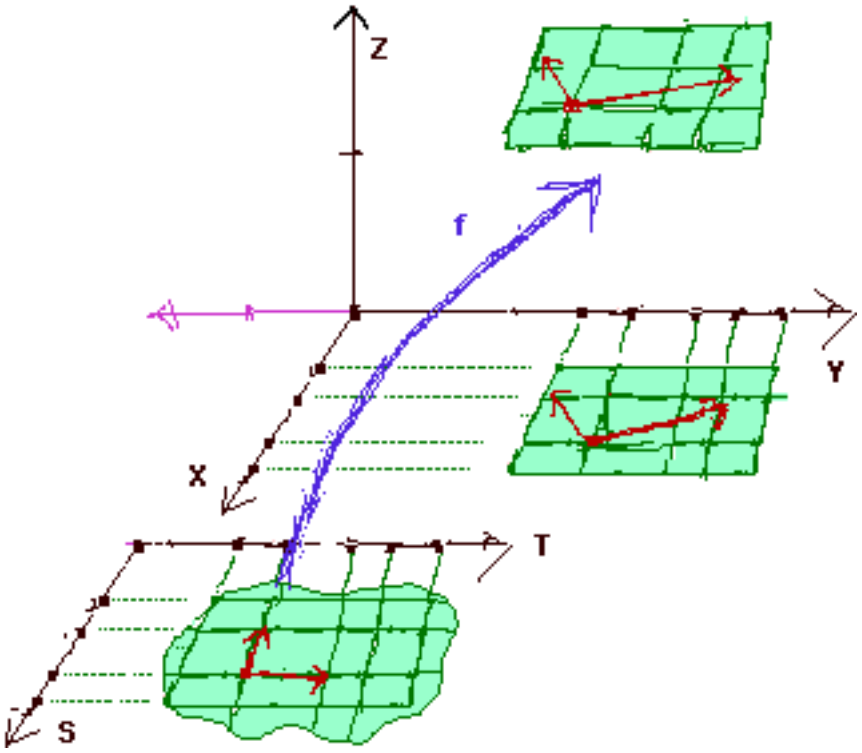
(i) We presume that f is one-to-one and that the values $f(s,t)$ for (s,t) in $\mathcal{W}$ constitute part of the graph of a differentiable function g restricted to an open subset $\mathcal{U}$ of its domain $\mathcal{O}$. This means $Z(s,t) = g(X(s,t), Y(s,t))$.

(ii) The functions A_s for each s and B_t for each t are continuously differentiable.

(iii) The vector $A'_s(t) \times B'_t(s)$ is never 0.

We will refer to an f like this as a **good parameterization of the piece of the surface defined by g above $\mathcal{U}$.**

The curves parameterized by A_s and B_t constitute a **coordinate grid** on the surface above $\mathcal{U}$, and each point on the surface above $\mathcal{U}$ can be located uniquely as the crossing place of two of these. These curves act as coordinate lines on the surface, and a mesh of closely spaced grid curves helps one navigate around up on the surface just as they do on the XY plane with the rectangular or polar grid. Small pieces of surface bounded by the grid curves will not be rectangles necessarily but will, when small enough so that the tangent plane is still very close to the surface, look like tilted parallelograms. The shadow of this piece will also look like a parallelogram down in the XY plane.



The condition that $A'_s(t) \times B'_t(s) \neq 0$ together with the fact that the surface is the graph of g on $\mathcal{U}$ ensures that neither the parallelograms in the grid near the surface nor the shadow parallelograms in the XY plane get too "skinny" or, at least, never go completely "flat."

We extend our "subscript for derivative" notation and write X_i, Y_i or Z_i in place of D_iX, D_iY or D_iZ. Also, since $Z = g(X,Y)$ the chain rule gives

$$Z_i = \nabla g(X,Y) \cdot \langle X_i,\ Y_i\rangle = g_1(X,Y)X_i + g_2(X,Y)Y_i,$$

a fact which can simplify formulas in case you have g as an explicit function of X and Y.

37.1. ***Exercise***. *Condition (ii) implies that the function f is continuously differentiable.*

In fact $A'_s(t) = \big\langle X_2(s,t), Y_2(s,t), Z_2(s,t)\big\rangle$
and $B'_t(s) = \big\langle X_1(s,t), Y_1(s,t), Z_1(s,t)\big\rangle$.

$A'_s(t)$ and $B'_t(s)$ are both vectors in the tangent plane to the surface at $f(s,t)$. The vector $A'_s(t) \times B'_t(s)$ is normal to the tangent plane there.

The magnitude of $A'_s(t) \times B'_t(s)$ is

$$\sqrt{(Y_2Z_1 - Y_1Z_2)^2 + (X_2Z_1 - X_1Z_2)^2 + (X_2Y_1 - X_1Y_2)^2}$$

where all derivatives are evaluated at (s,t), and this magnitude is a continuous function on $\mathcal{W}$ and never 0.

---

38. Change of Variable on a Surface

We presume that $f = \langle X, Y, Z\rangle$ is a good parameterization of the piece of the surface defined by g above $\mathcal{U}$. We will presume that the domain $\mathcal{W}$ of f is inside the rectangle $[a,b]\times[c,d]$ in the plane. Our goal is to create a **change of variables** formula for surface integrals for surfaces given parametrically.

Suppose P is a partition of this rectangle composed of subrectangles $[s_{i-1}, s_i]\times[t_{j-1}, t_j]$ for $i = 1\ldots n$ and $j = 1\ldots m$.

If $[s_{i-1}, s_i]\times[t_{j-1}, t_j]$ is entirely contained in $\mathcal{W}$ then $\Delta t_j A'_{s_{i-1}}(t_{j-1})$ and $\Delta s_i B'_{t_{j-1}}(s_{i-1})$ are both vectors in the tangent plane to the surface at $f(s_{i-1}, t_{j-1})$.

If the mesh of the partition is small enough, the parallelogram formed by these two vectors in the tangent plane will be very close to the part of the surface covered by f on the rectangle $[s_{i-1}, s_i]\times[t_{j-1}, t_j]$.

This parallelogram has area $\Delta t_j \Delta s_i |A'_{s_{i-1}}(t_{j-1}) \times B'_{t_{j-1}}(s_{i-1})|$. Suppose h is a continuous and bounded function defined on $\mathcal{U}$. As before, we can imagine that $h(X,Y)$ represents charge density, this time at the spot $(X, Y, g(X,Y))$ on the surface. h will be nearly constant on the part of $\mathcal{U}$ corresponding to the shadow of the tiny parallelogram formed by $\Delta t_j A'_{s_{i-1}}(t_{j-1})$ and $\Delta s_i B'_{t_{j-1}}(s_{i-1})$ up on the tangent plane at $f(s_{i-1}, t_{j-1})$, so the total charge on this piece will be nearly

$$h(X(s_{i-1}, t_{j-1}), Y(s_{i-1}, t_{j-1}))\ |A'_{s_{i-1}}(t_{j-1}) \times B'_{t_{j-1}}(s_{i-1})|\ \Delta s_i\ \Delta t_j.$$

Adding together these numbers for all the partition rectangles entirely contained in $\mathcal{W}$ generates an integral

$$\int_{\mathcal{W}} h(X(s,t),Y(s,t))\ |A'_s(t)\times B'_t(s)|\ ds\ dt$$
$$= \int_{\mathcal{W}} h(X,Y)\sqrt{(Y_2Z_1-Y_1Z_2)^2+(X_2Z_1-X_1Z_2)^2+(X_2Y_1-X_1Y_2)^2}\ ds\ dt$$

where in the last integral X, Y and Z and their derivatives are to be evaluated at (s,t).

This integral **should**, if our thinking is correct, be the same as the surface integral

$$\int_{\mathcal{U}} h(X,Y)\sqrt{(g_1(X,Y))^2+(g_2(X,Y))^2+1}\ dX\ dY$$

and provide an alternate way of calculating that integral as an integral over $\mathcal{W}$ instead of an integral over $\mathcal{U}$. In this context the integral is called a **parametric form of the surface integral of h on this piece of the surface.**

As before, a density is sometimes given as a function k of the three coordinates (X,Y,Z) on the surface. Since $Z = g(X,Y)$ this gives the density $h(X,Y) = k(X,Y,g(X,Y))$ as a function of two variables which is how we have expressed our surface integral. In this case, the same integral is also called the **parametric form of the surface integral of k on this piece of the surface.**

In applications one often (even usually) knows the function g, which defines the surface, explicitly rather than simply inferring that such a g exists somehow. In that case the integral formula can above can be modified to include g and with g some more geometric content.

Substitute $Z_i = \langle g_1, g_2\rangle \cdot \langle X_i, Y_i\rangle$ into the formula

$$(Y_2Z_1 - Y_1Z_2)^2 + (X_2Z_1 - X_1Z_2)^2 + (X_2Y_1 - X_1Y_2)^2$$

(which appears under the radical in the integration formula above) and expand and combine like terms.

Then expand $(g_1^2 + g_2^2 + 1)(X_1Y_2 - Y_1X_2)^2$. You will find the result to be the same. This means that the parametric form of this surface integral can be calculated as

$$\int_{\mathcal{W}} h(X,Y)\sqrt{(g_1(X,Y))^2+(g_2(X,Y))^2+1}\ |X_1Y_2 - Y_1X_2|\ ds\ dt.$$

where in each case the functions X, Y, X_i and Y_i are to be evaluated at (r,s) in the integral.

In calculations it is convenient to note that the sign of $X_1Y_2 - Y_1X_2$ does not vary on $\mathcal{W}$.

Finally, we come to a notation issue. It is common to see the **equality of the surface integral and the parametric form of that integral** expressed as:

$$\int_{\mathcal{U}} h\ \sqrt{(\nabla g)^2+1}\ dX\ dY\ =\ \int_{\mathcal{W}} h(X,Y)\ \sqrt{(\nabla g)^2\circ f+1}\ |det\ f'|\ ds\ dt.$$

Sometimes the explicit composition with the functions f and (X, Y) on the right is suppressed and you see

$$\int_{\mathcal{W}} h\ \sqrt{(\nabla g)^2 + 1}\ \ |det\ f'|\ ds\ dt.$$

It is assumed that the user will simply know to do the right thing in context. Its only virtue is brevity.

38.1. ***Exercise***. *Verify the following statements:*

The area of the shadow in the XY plane of the parallelogram formed by $\Delta t A'_s(t)$ and $\Delta s B'_t(s)$ in the tangent plane at $f(s,t)$ on the surface is $|X_1 Y_2 - Y_1 X_2|$.

If θ is the angle between the tangent plane at $f(s,t)$ and the XY plane then $\dfrac{1}{cos(\theta)} = \sqrt{g_1^2 + g_2^2 + 1}$.

Use these geometrical facts to justify the last integration formula.

38.2. ***Exercise***. *Suppose that the surface $Z = g(X, Y)$ in the parametric form of the surface integral found above arises as a part of a level set $M(X, Y, Z) = \omega$ of continuously differentiable M with $D_3 M(P) \neq 0$ whenever $P = (X, Y, Z)$ is in the level set.*

Show that the angle θ between the surface at P and the XY plane satisfies

$$cos(\theta) = \frac{\nabla M(P)}{|\nabla M(P)|} \cdot \vec{k}.$$

This generates in this case an alternative expression for the surface integral:

$$\int_{\mathcal{W}} h(X, Y)\ \frac{|\nabla M(P)|}{D_3 M(P)} |X_1 Y_2 - Y_1 X_2|\ ds\ dt.$$

Every now and then, symmetry or other considerations will allow you to determine $cos(\theta)$ on the surface at each (s,t). But apart from these special circumstances and the picture provided by this representation, the advantage of the formula is limited due to the fact that to calculate $\frac{\nabla M(X,Y,Z)}{|\nabla M(X,Y,Z)|}$ usually requires that you know $Z = g(X, Y)$ so you could just as well have used the earlier formula.

38.3. ***Exercise***. *What do the formulas discussed in Exercises 38.1 and 38.2 look like when $Z = g(X, Y) = \omega$, a constant?*

What do these formulas look like when $f(s,t) = \langle s, t, 0\rangle$: that is, when f does, essentially, nothing?

What do these formulas look like when $f(s,t) = sV + tW$ where V and W are two nonzero constant vectors in $\mathbb{R}^3$ which are not multiples of each other?

38.4. ***Exercise***. *Consider the function* $f(r,\theta) = \left\langle rcos(\theta), rsin(\theta), \sqrt{1-r^2}\right\rangle$ *with domain* $\mathcal{W} = (0,1)\times(0,\frac{\pi}{2})$ *and with range equal to* $\mathcal{U}$, *that part of the unit sphere that sticks into the "all positive" octant of space, the graph of* $g(X,Y) = \sqrt{1-X^2-Y^2}$.

Show that f *is a "good parameterization" of* $\mathcal{U}$.

Calculate the area of this piece of the sphere by thinking of it in three ways.

First do the calculation as

$$\int_{\mathcal{W}} \sqrt{(Y_2Z_1-Y_1Z_2)^2+(X_2Z_1-X_1Z_2)^2+(X_2Y_1-X_1Y_2)^2}\ dr\ d\theta.$$

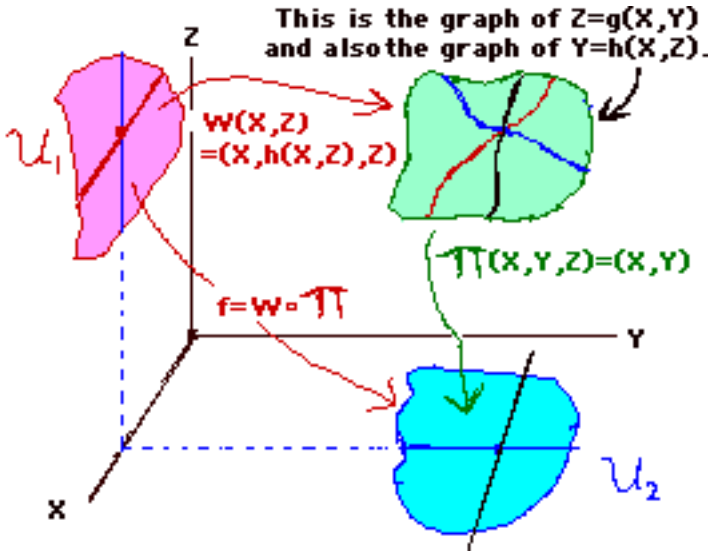
Second, use the formula

$$\int_{\mathcal{W}} \sqrt{(g_1(X,Y))^2+(g_2(X,Y))^2+1}\ \ |X_1Y_2-Y_1X_2|\ dr\ d\theta.$$

Third, determine the angle that the normal to the sphere makes with the XY *plane as a function of* r *and use Exercise 38.2.*

At this point we will tie up a loose end and demonstrate the formulas in action at the same time.

It is quite possible that the same piece of surface might be the graph of a function $Z = g(X,Y)$ over an open set $\mathcal{U}_2$ in the XY plane or $Y = h(X,Z)$ over an open set $\mathcal{U}_1$ in the XZ plane. It would be pleasant if the area of this surface (and other surface integrals too) calculated as an integral over $\mathcal{U}_1$ was the same as the area calculated as an integral over $\mathcal{U}_2$.



Note that $Y = h(X, g(X, Y))$. Let the function $\mathcal{Y}(X, Y)$ denote $h(X, g(X, Y))$. Differentiating and then evaluating at a specific (X, Y, Z) on the surface yields

$$\begin{aligned} D_1\mathcal{Y}(X,Y) &= 0 = D_1h(X,Z) + D_2h(X,Z)D_1g(X,Y) \\ D_2\mathcal{Y}(X,Y) &= 1 = D_2h(X,Z)D_2g(X,Y). \end{aligned}$$

Let W be the function from $\mathcal{U}_1$ to the surface defined by $W(X, Z) = (X, h(X, Z), Z)$ and let π be the function from $\mathbb{R}^3$ to $\mathbb{R}^2$ defined by $\pi(X, Y, Z) = (X, Y)$.

The function $f = \pi \circ W$ takes points from $\mathcal{U}_1$ and sends them to points in $\mathcal{U}_2$. The function f is one-to-one because both W and π are one-to-one, and every point in $\mathcal{U}_2$ is $f(X, Z)$ for some point (X, Z) in $\mathcal{U}_1$. Note that $f(X, Z) = (X, h(X, Z))$.

The grid curves $A_X(Z)$ and $B_Z(X)$ are differentiable and $A'_X(Z) = \langle 1, D_1h(X, Z)\rangle$ and $B'_Z(X) = \langle 0, D_2h(X, Z)\rangle$.

f will be a good parameterization of $\mathcal{U}_2$ provided that D_2h is never 0. But the calculation above shows that $1 = D_2h(X, Z)D_2g(X, Y)$ so D_2h can never be 0 in this situation.

Note also that

$$f'(X,Z) = \begin{pmatrix} 1 & 0 \\ D_1h(X,Z) & D_2h(X,Z) \end{pmatrix} \quad \text{and} \quad (det\ f')(X,Z) = D_2h(X,Z).$$

We would like to show that

$$\int_{\mathcal{U}_1} \sqrt{(\nabla h)^2 + 1}\ dX\ dZ\ = \int_{\mathcal{U}_2} \sqrt{(\nabla g)^2 + 1}\ dX\ dY.$$

The change of variable formula says that

$$\int_{\mathcal{U}_2} \sqrt{(\nabla g)^2 + 1}\ dX\ dY = \int_{\mathcal{U}_1} \sqrt{(\nabla g)^2 \circ f + 1}\ |det\ f'|\ dX\ dZ.$$

So we would like to show that the two integrands of the integrals over $\mathcal{U}_1$ are identical. Since they are positive integrands we can square both yielding the sufficient condition for equality:

$$(\nabla h)^2 + 1\ = (\ (\nabla g)^2 \circ f + 1\ )\ |det\ f'|^2.$$

Expanding this yields

$$\begin{aligned} &(\ D_1h(X,Z)\ )^2\ +\ (\ D_2h(X,Z)\ )^2\ +\ 1 \\ &\quad = (\ D_1g(X,Y)D_2h(X,Z)\ )^2\ +\ (\ D_2g(X,Y)D_2h(X,Z)\ )^2\ +\ (\ D_2h(X,Z)\ )^2 \end{aligned}$$

where (X, Y, Z) corresponds to a specific point on the surface.

From the earlier calculation we know that

$$D_2g(X,Y)D_2h(X,Z) = 1 \quad \text{and} \quad D_2h(X,Z)D_1g(X,Y) = -D_1h(X,Z)$$

and the result follows.

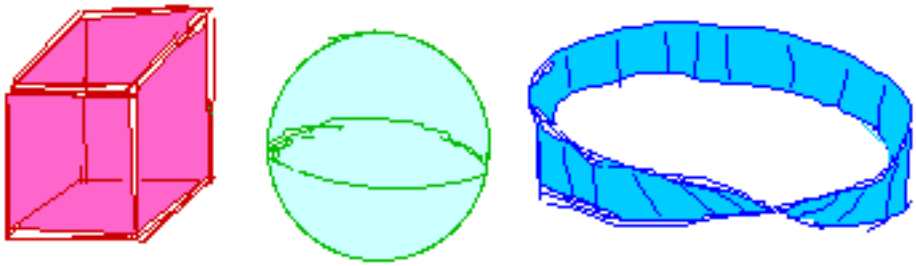
38.5. ***Exercise***. *Suppose we have a piece of a surface parameterized by*

$$f(s,t) = \langle X, Y, Z\rangle = \left\langle sin(t),\ e^t s,\ t\ ln(s)\right\rangle$$

for $0 < t < 1$ *and* $1 < t < 2$. *Verify that this is a good parameterization. Create an integral for the area and use a utility such as Maple to find an approximation to that area.*[42]

39. Surface Integrals Over Composite Surfaces

Often we are concerned with surface integrals over sets which do not obey our rather restrictive criterion: surfaces which can be parameterized by a single function in terms of two variables. The sphere, the torus, and the cube are common shapes which fail this criterion and we need to be able to work with them.



Another interesting shape is the **Möbius strip**, which you most likely have encountered at one time or another. It can be created by taking a narrow strip of paper and putting a half-twist along its long direction and then taping the opposite short sides together. It has the peculiar property of possessing only one side and one edge. It is not the graph of a function of two variables.

Each of these examples can be broken into a finite number of pieces bounded by piecewise good curves where each piece only touches another along these curves and where each piece, minus its bounding curve, *is* a surface according to our definition.

We will then **define the surface integral over the aggregate object** to be the sum of the integrals of all these non-overlapping pieces. This decomposition is not too hard to accomplish in many common cases.

We will now make this a little more precise, although the process is quite messy from a notational standpoint. Just keep in mind what we are trying to do. We make the following suppositions.

- Suppose $\mathcal{S}$ is a bounded set in space, the union of sets $\mathcal{S}_i$ for $i = 1, \ldots, n$.
- For each i there is a piecewise good loop $\mathcal{C}_i$ in the $s_i t_i$ plane surrounding an open set $\mathcal{O}_i$, so that $\mathcal{C}_i$ is the boundary of $\mathcal{O}_i$, and for which there is given an open set $\mathcal{U}_i$ containing both $\mathcal{C}_i$ and $\mathcal{O}_i$. Define $\overline{\mathcal{O}_i}$ to be the set of points from $\mathcal{O}_i$ together with its bounding loop $\mathcal{C}_i$.
- For each i there is a vector valued function function $f_i = \langle X_i, Y_i, Z_i\rangle$ defined on $\mathcal{U}_i$ which is a good parameterization of a surface in space.

- $\mathcal{S}_i$ is exactly the object formed as the set of all $f_i(s_i, t_i)$ where the points (s_i, t_i) are taken from $\overline{\mathcal{O}_i}$.
- When $i \neq j$ any point P in the overlap $\mathcal{S}_i \bigcap \mathcal{S}_j$ of two of the pieces must be of the form $P = f_i(s_i, t_i) = f_j(s_j, t_j)$ where (s_i, t_i) is in $\mathcal{C}_i$ and (s_j, t_j) is in $\mathcal{C}_j$. In other words, $\mathcal{S}_i$ and $\mathcal{S}_j$ can only touch on their bounding curves.
- h is a bounded real valued function defined, for every i, on all points $f_i(s_i, t_i)$ whenever (s_i, t_i) is in $\mathcal{O}_i$ and for which $\int_{\mathcal{O}_i} h \circ f_i \, |\det f_i'| \, ds_i \, dt_i$ exists for each i.

We will call any collection of sets and functions satisfying the first five items on the list an **admissible decomposition of $\mathcal{S}$**.

The part of $\mathcal{S}_i$ consisting of points $f_i(s_i, t_i)$ where (s_i, t_i) is taken from $\mathcal{O}_i$ is called a **member of the decomposition.** It consists of $\mathcal{S}_i$ "minus" its bounding curve.

If $\mathcal{S}$ has an admissible decomposition we will refer to it as a **composite surface.**

With all this setup and all these suppositions, we now define **the surface integral of h over $\mathcal{S}$** to be

$$\int_{\mathcal{S}} h \, d\mathcal{S} = \sum_{i=1}^{n} \int_{\mathcal{O}_i} h \circ f_i \, |\det f_i'| \, ds_i \, dt_i.$$

39.1. ***Exercise***. *A box is constructed in the shape of the surface of the standard unit cube in the all-positive octant in space with distances measured in meters. The density of the surface of this cube varies according to the function $h(X, Y, Z) = X + Y + Z$ kilograms per square meter. What is the mass of this cube?*

39.2. ***Exercise***. *Calculate the mass of the surface of the sphere centered at the origin of radius 1 meter if its density is $2Z$ kilograms per square meter at a point (X, Y, Z) with $Z > 0$ and $-3Z$ kilograms per square meter at a point (X, Y, Z) with $Z < 0$.*

39.3. ***Exercise***. *** Suppose $\mathcal{S}$ has two admissible decompositions. Whenever h is a function defined on all members of both decompositions and for which the surface integral can be calculated using one decomposition then it can be calculated using the other too. The value obtained using the second decomposition process will agree with that obtained from the first decomposition.*

40. Orientation of Surfaces and Integrals Involving Vector Fields

We saw in Section 24 that certain line integrals involving vector fields, such as those representing flow or work or flux, required specification of a continuous choice of unit tangent vector along the curve. We called a choice like that an orientation for the curve, and we made restrictions on our curves so that there were two possible orientations.

In this section we will do something similar for surfaces in space.

Let us start by describing our surface $\mathcal{S}$ as the graph of a continuously differentiable function g of two variables defined on an open set $\mathcal{O}$ in the plane.

At any point on this surface, there are exactly two unit normal vectors, and these two vectors depend on the geometry of the surface and not the parameterization.

For each point P of $\mathcal{S}$ let $\mathcal{N}(P)$ denote a choice of one of the two unit normals to $\mathcal{S}$ at P.

Such a selection constitutes a vector valued function $\mathcal{N}$, called a unit normal vector field, on $\mathcal{S}$.

We will call $\mathcal{N}$ **consistent** if $\mathcal{N}$ is a continuous vector valued function on $\mathcal{O}$.

An **orientation** for $\mathcal{S}$ is a choice of consistent unit normal vector field for $\mathcal{S}$.

An **oriented surface** is a surface $\mathcal{S}$ together with a choice $\mathcal{N}$ of an orientation.

Every surface obtained as the graph of differentiable g has an orientation. In fact,

$$\frac{\langle -g_1,\ -g_2,\ 1\rangle}{|\langle -g_1,\ -g_2,\ 1\rangle|} \quad \text{and} \quad \frac{\langle g_1,\ g_2,\ -1\rangle}{|\langle g_1,\ g_2,\ -1\rangle|}$$

are two different orientations for $\mathcal{S}$. The first is usually called the upward unit normal and the second the downward unit normal for g.

40.1. ***Exercise***. *For certain domains $\mathcal{O}$ there can be more than two orientations of $\mathcal{S}$. Describe an example of this. (*) How many orientations might there be?*

40.2. ***Exercise***. *For a good parameterization f with $f(s,t) = \langle X(s,t), Y(s,t), Z(s,t)\rangle$ of the surface $\mathcal{S}$ you can create a consistent unit normal as*

$$\mathcal{N}_f(f(s,t)) = \frac{\langle X_1(s,t), Y_1(s,t), Z_1(s,t)\rangle \times \langle X_2(s,t), Y_2(s,t), Z_2(s,t)\rangle}{|\langle X_1(s,t), Y_1(s,t), Z_1(s,t)\rangle \times \langle X_2(s,t), Y_2(s,t), Z_2(s,t)\rangle|}.$$

*(*) Show that any consistent unit normal $\mathcal{N}$ for $\mathcal{S}$ can be obtained as $\mathcal{N} = \mathcal{N}_f$ for **some** good parameterization f of the surface. Any such f is called **consistent** with $\mathcal{N}$.*

We will now use this concept to calculate the flux of a vector field through a surface.

Suppose you are given an oriented surface $\mathcal{S}$ in space with unit normal selection $\mathcal{N}$. Suppose also that we are given a continuous vector field $F = \langle M, N, P\rangle$ defined on an open set containing the $\mathcal{S}$ and a good parameterization $f = \langle X, Y, Z\rangle$ defined on an open set $\mathcal{W}$ which is consistent with the orientation.

The **flux of F through the oriented surface with orientation $\mathcal{N}$** is defined to be the surface integral of the function $F \cdot \mathcal{N}$ over the surface. Using the **consistent** parameterization f and letting f_1 denote $\langle X_1, Y_1, Z_1\rangle$ and f_2 denote $\langle X_2, Y_2, Z_2\rangle$ we find that the integrand is

$$\begin{aligned} F \cdot \mathcal{N}\, |f_1 \times f_2| =& F \cdot f_1 \times f_2 \\ =& \langle M, N, P\rangle \cdot \langle Y_1Z_2 - Z_1Y_2,\ Z_1X_2 - X_1Z_2,\ X_1Y_2 - Y_1X_2\rangle. \end{aligned}$$

This gives the flux as:

$$\textbf{Flux: } \int_{\mathcal{W}} M\ (Y_1Z_2 - Z_1Y_2) + N\ (Z_1X_2 - X_1Z_2) + P\ (X_1Y_2 - Y_1X_2)\ ds\ dt.$$

In case $Z = g(X, Y)$ the integrand becomes $\langle M, N, P\rangle \cdot \langle -g_1, -g_2, 1\rangle(X_1Y_2 - X_2Y_1)$ so

$$\textbf{Flux: } \int_{\mathcal{W}} \langle M, N, P\rangle \cdot \langle -g_1, -g_2, 1\rangle(X_1Y_2 - X_2Y_1)\ ds\ dt.$$

One frequently sees a notation similar to

$$\int_{\mathcal{S}} F(\mathcal{S}) \cdot \mathcal{N}(\mathcal{S})\ d\mathcal{S}$$

to denote the flux of F past the oriented surface $\mathcal{S}$. The orientation is built into $\mathcal{N}$. The purpose of this notation is to focus attention on the surface and normal and away from the parameterization. Sometimes you can even use this formula to calculate in simple flux situations.

We can interpret the flux to denote the net flow of a fluid past the surface in the direction indicated by $\mathcal{N}$. In that interpretation, a given vector value of F represents the velocity of a fluid. Its magnitude is the amount of fluid which would flow past a given place in a pipe of unit cross-sectional area placed parallel to F in the fluid. The discussion of Section 24 now applies with just a little modification.

Finally, we extend the idea of flux to composite surfaces.

An **oriented composite surface** is a composite surface together with an admissible decomposition and an orientation for each of the members of the decomposition. The orientation itself is denoted $\mathcal{N}$ and the orientation for the member of the decomposition corresponding to subscript i will be denoted, naturally, $\mathcal{N}_i$.

Suppose $\mathcal{S}$ is an oriented composite surface with n members and F is a vector field continuous on an open set containing $\mathcal{S}$. The **flux of F through the oriented composite surface $\mathcal{S}$** is

$$\textbf{Flux Past Composite } \mathcal{S}: \int_{\mathcal{S}} F(\mathcal{S}) \cdot \mathcal{N}(\mathcal{S})\ d\mathcal{S} = \sum_{i=1}^{n} \int_{\mathcal{S}_i} F(\mathcal{S}_i) \cdot \mathcal{N}_i(\mathcal{S}_i)\ d\mathcal{S}_i.$$

There may be many ways to decompose a given composite surface and many orientations for each decomposition. If there are n members

of a decomposition, there are 2^n different orientations. Different orientations are likely to give different results. It is up to the user to decide how these choices should be made in a particular case.[43]

40.3. ***Exercise***. *Let $\mathcal{S}$ denote the standard unit cube in the all-positive octant of space. Let F be the vector field $F = \langle Z, Y^2, XY\rangle$. Calculate the flux "out" of this composite surface: that is, choose unit normals pointing away from the interior of the cube at each point.*

---

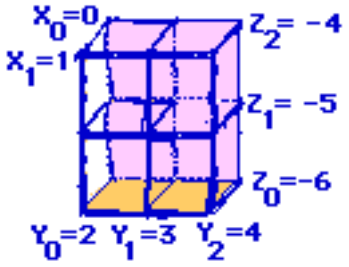
41. Volume

Suppose $[a_1, b_1]$, $[a_2, b_2]$ and $[a_3, b_3]$ are intervals. A **closed rectangular solid** formed from these intervals in $\mathbb{R}^3$ will be denoted $[a_1, b_1] \times [a_2, b_2] \times [a_3, b_3]$, and consist of those ordered triples (X, Y, Z) with $a_1 \leq X \leq b_1$ and $a_2 \leq Y \leq b_2$ and $a_3 \leq Z \leq b_3$.

A set $\mathcal{O}$ in space is called **bounded** when there is a closed rectangular solid which contains $\mathcal{O}$.

We are going to describe how to define and calculate a number which we will interpret as the volume of a bounded open set in space. We will also define integrals of bounded continuous functions defined on such sets and indicate how they might be calculated.

Suppose $a_1 = X_0 < \cdots < X_n = b_1$ is a partition of the interval $[a_1, b_1]$ and $a_2 = Y_0 < \cdots < Y_m = b_2$ is a partition of the interval $[a_2, b_2]$ and and $a_3 = Z_0 < \cdots < Z_p = b_3$ is a partition of the interval $[a_3, b_3]$. The collection P of rectangular solids formed from all the subintervals from consecutive partition members of these three intervals form what is called a **partition of the rectangular solid** $[a_1, b_1] \times [a_2, b_2] \times [a_3, b_3]$. There are mnp of these smaller rectangles. The **mesh** of this partition is the length of the longest edge of any rectangular solid in the partition.



A set of points C with members $C_{i,j,k}$ for $i = 1 \ldots n$ and $j = 1 \ldots m$ and $k = 1 \ldots p$ in the rectangular solid is called **subordinate to the partition** P if $C_{i,j,k}$ is in the subrectangle $[X_{i-1}, X_i] \times [Y_{j-1}, Y_j] \times [Z_{k-1}, Z_k]$ for each i, j and k.

We suppose h is a **bounded continuous real valued function** defined on $\mathcal{O}$.

Consider the sum $\sum_{\mathfrak{O}} h(C_{i,j,k})\Delta X_i \Delta Y_j \Delta Z_k$ where this notation indicates that the sum is over those subscripts corresponding to rectangular solids in P which are entirely inside $\mathfrak{O}$. Sums formed in this way are called **Riemann sums**, and depend on $h,\ C$ and P.

It is a fact[44] that under these conditions there is a number denoted

$$\int_{\mathfrak{O}} h(X,Y,Z)\ dX\ dY\ dZ$$

to which this sum is arbitrarily close provided only that the mesh of P is small enough. This number is called the **the triple integral of** h **over** $\mathfrak{O}$. It could also be called a **volume integral** or the **integral of h weighted by volume.**

For a positive integer n let $\mathbb{B}_n$ be the collection of all rectangular solids of the form $\left[\frac{i}{2^n}, \frac{i+1}{2^n}\right] \times \left[\frac{j}{2^n}, \frac{j+1}{2^n}\right] \times \left[\frac{k}{2^n}, \frac{k+1}{2^n}\right]$ where $i,\ j$ and k are any integers. Let $C^n_{i,j,k}$ denote the point $\left(\frac{i}{2^n}, \frac{j}{2^n}, \frac{k}{2^n}\right)$. So

$$\lim_{n\to\infty} \sum_{\mathbb{B}_n} \frac{f(C^n_{i,j,k})}{8^n} = \int_{\mathfrak{O}} h(X,Y)\ dX\ dY\ dZ$$

where the sum is over all $i,\ j$ and k for which the rectangle $\left[\frac{i}{2^n}, \frac{i+1}{2^n}\right] \times \left[\frac{j}{2^n}, \frac{j+1}{2^n}\right] \times \left[\frac{k}{2^n}, \frac{k+1}{2^n}\right]$ is entirely in $\mathfrak{O}$.

If f and g are continuous and bounded on bounded open $\mathfrak{O}$ and c is a real number then

$$\int_{\mathfrak{O}} f(X,Y,Z)\ dX\ dY\ +\ c\ \int_{\mathfrak{O}} g(X,Y,Z)\ dX\ dY\ dZ = \int_{\mathfrak{O}} f(X,Y,Z) + c\ g(X,Y,Z)\ dX\ dY\ dZ.$$

If $m \le f(P) \le M$ for all P in $\mathfrak{O}$ for constants m and M then

$$\int_{\mathfrak{O}} m\ dX\ dY\ dZ\ \le \int_{\mathfrak{O}} f(X,Y,Z)\ dX\ dY\ dZ\ \le \int_{\mathfrak{O}} M\ dX\ dY\ dZ.$$

Also, if $f \ge g$ and $f(Q) > g(Q)$ for even one point Q in $\mathfrak{O}$ then

$$\int_{\mathfrak{O}} f(X,Y,Z)\ dX\ dY\ dZ\ > \int_{\mathfrak{O}} g(X,Y,Z)\ dX\ dY\ dZ.$$

If f is nonnegative and the open sets $\mathfrak{O}$ and $\mathfrak{U}$ are both contained in the domain of f then the union of these two sets, $\mathfrak{O} \cup \mathfrak{U}$, is an open set and

$$\int_{\mathfrak{O}\cup\mathfrak{U}} f(X,Y,Z)\,dX\,dY\,dZ\ \le \int_{\mathfrak{O}} f(X,Y,Z)\,dX\,dY\,dZ + \int_{\mathfrak{U}} f(X,Y,Z)\,dX\,dY\,dZ.$$

If $\mathfrak{O} \cap \mathfrak{U} = \varnothing$ equality holds above, as in the case of integrals in the plane.

41.1. ***Exercise***. *** Try to prove the statements in the last paragraph.*

If $\mathfrak{I}$ is a sequence of open sets in space and $\mathfrak{I}_n \subset \mathfrak{I}_{n+1}$ for each n then the union of all these nested open sets is itself open. Let's call this set $\mathfrak{O}$. It is a fact that if f is a bounded continuous function on $\mathfrak{O}$ then $\lim_{n\to\infty} \int_{\mathfrak{I}_n} f(X,Y,Z)\ dX\ dY\ dZ = \int_{\mathfrak{O}} f(X,Y,Z)\ dX\ dY\ dZ$.

We would also like to think about integrals over some closed sets. You will recall that in the $2D$ case we considered a skinny set $\mathfrak{C}_\varepsilon$ around a bounding curve $\mathfrak{C}$ and saw that $\int_{\mathfrak{C}_\varepsilon} h(X,Y)\ dX\ dY < 4ML\varepsilon$ for constants M and L. We need a replacement for this condition in $3D$.

Sometimes a bounded closed set $\mathfrak{K}$ in $3D$ is trapped between sequences of bounded open sets

$$\mathfrak{I}_n \subset \mathfrak{I}_{n+1} \subset \mathfrak{K} \subset \mathfrak{O}_{n+1} \subset \mathfrak{O}_n$$

for positive integers n and with

$$\lim_{n\to\infty} \int_{\mathfrak{O}_n} 1\ dX\ dY\ dZ - \int_{\mathfrak{I}_n} 1\ dX\ dY\ dZ = 0.$$

This implies that for any function f bounded and continuous on at least one $\mathfrak{O}_n$

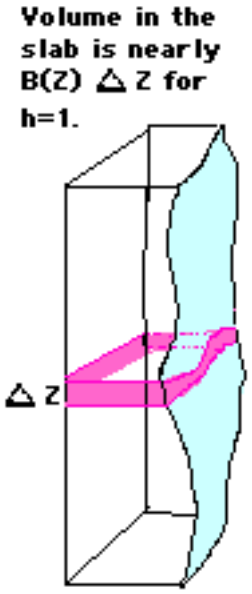
$$\begin{aligned}\lim_{n\to\infty} \int_{\mathfrak{I}_n} f(X,Y,Z)\ dX\ dY\ dZ &= \lim_{n\to\infty} \int_{\mathfrak{O}_n} f(X,Y,Z)\ dX\ dY\ dZ \\ &= \int_{\mathfrak{O}} f(X,Y,Z)\ dX\ dY\ dZ\end{aligned}$$

and we define $\int_{\mathfrak{K}} f(X,Y,Z)\ dX\ dY\ dZ$ to be this common limit.

So we have defined integrals on certain types of closed sets too: namely, closed sets which can be "approximated" by open sets from inside and outside and for bounded continuous functions defined on some open set containing this closed set. We will leave for later classes the precise characterization of which closed sets we might be talking about here.

However the closed set consisting of those points in space from the XY plane up to the graph of a nonnegative continuous function defined on the closed set in the plane surrounded by a good loop is a closed set of this type. Also, any set which can be broken up into a finite number of pieces of this type is, itself, of this type. This gives the most commonly found examples.

41.2. ***Exercise***. *** Try to prove the statements in the last paragraph. If that seems too hard work on special cases, such as the upper unit hemisphere.*



We now come to the issue of how one might actually calculate an integral of a bounded and continuous function on a bounded open set $\mathfrak{O}$.

For each Z in $[a_3, b_3]$ define S_Z to be the set of those ordered pairs (X, Y) in $[a_1, b_1] \times [a_2, b_2]$ for which (X, Y, Z) is in $\mathfrak{O}$. Each S_Z is an open set and when it is nonempty the function $\mathcal{A}_Z$ defined on S_Z by $\mathcal{A}_Z(X, Y) = h(X, Y, Z)$ is bounded and continuous.

Define $\mathcal{B}(Z) = \int_{S_Z} \mathcal{A}_Z(X, Y)\, dX\, dY$ for each Z in $[a_3, b_3]$, where if S_Z is empty this number is 0.

Let T be the set of those Z in $[a_3, b_3]$ for which (X, Y, Z) is in $\mathfrak{O}$ for any (X, Y). T is open and the function $\mathcal{B}$ is continuous and bounded on T.

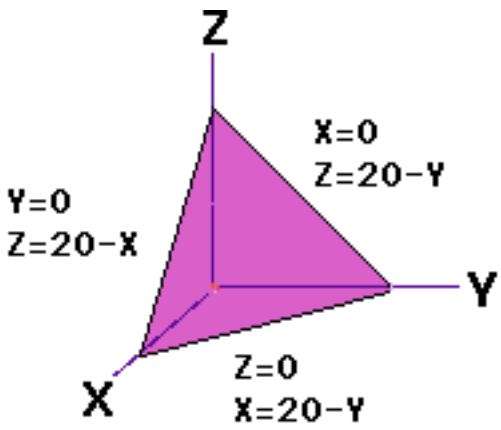
It is a fact[45] that under these conditions

$$\int_{\mathfrak{O}} h(X, Y, Z)\, dX\, dY\, dZ = \int_T \mathcal{B}(Z)\, dZ = \int_T \left(\int_{S_Z} \mathcal{A}_Z(X, Y)\, dX\, dY \right) dZ.$$

The last integral is called the **iterated integral of** h on $\mathfrak{O}$, and is the main tool used to actually calculate an integral on an open set in space, or a closed set which can be approximated by open ones as above. Under the conditions of this section, iterated integrals can be calculated in any order, by modifying the definition slightly to integrate first with respect to different pairs of variables. The theorem which identifies the triple integral with the various iterated integrals is called **Fubini's Theorem** and, as in two dimensions, is very important.

When $h = 1$, the constant function, this integral is to be interpreted as the **volume of** $\mathfrak{O}$. If h is nonnegative, one could interpret the integral as the **mass** of a lump of material in the shape of $\mathfrak{O}$ with density function h. For a general h the integral could represent total **charge** on the lump with charge density h.

41.3. ***Exercise***. *Consider the region in the all-positive octant under the plane* $Z = 20 - X - Y$.



Satisfy yourself that the region has volume given by

$$\int_{Y=0}^{Y=20} \int_{X=0}^{X=20-Y} \int_{Z=0}^{Z=20-X-Y} 1 \; dZ \; dX \; dY.$$

Evaluate this integral. Then change the order of integration in five different ways (there are six possible orders in which the integration can be performed) and calculate the volume of this region again.

41.4. ***Exercise***. ** Suppose f is defined on a bounded open set $\mathfrak{O}$ in space and the third order partial derivatives $D_{i,j,k}f$ and $D_{k,i,j}f$ of f exist and are continuous in $\mathfrak{O}$. What can you conclude about these partial derivatives?*

41.5. ***Exercise***. *Consider the region in the all-positive octant under the surface $Z = sin(X) + 3$ and with Y coordinate bounded above by $Y = X^2 + Z + 1$ and X bounded above by $X = 2\pi$.*

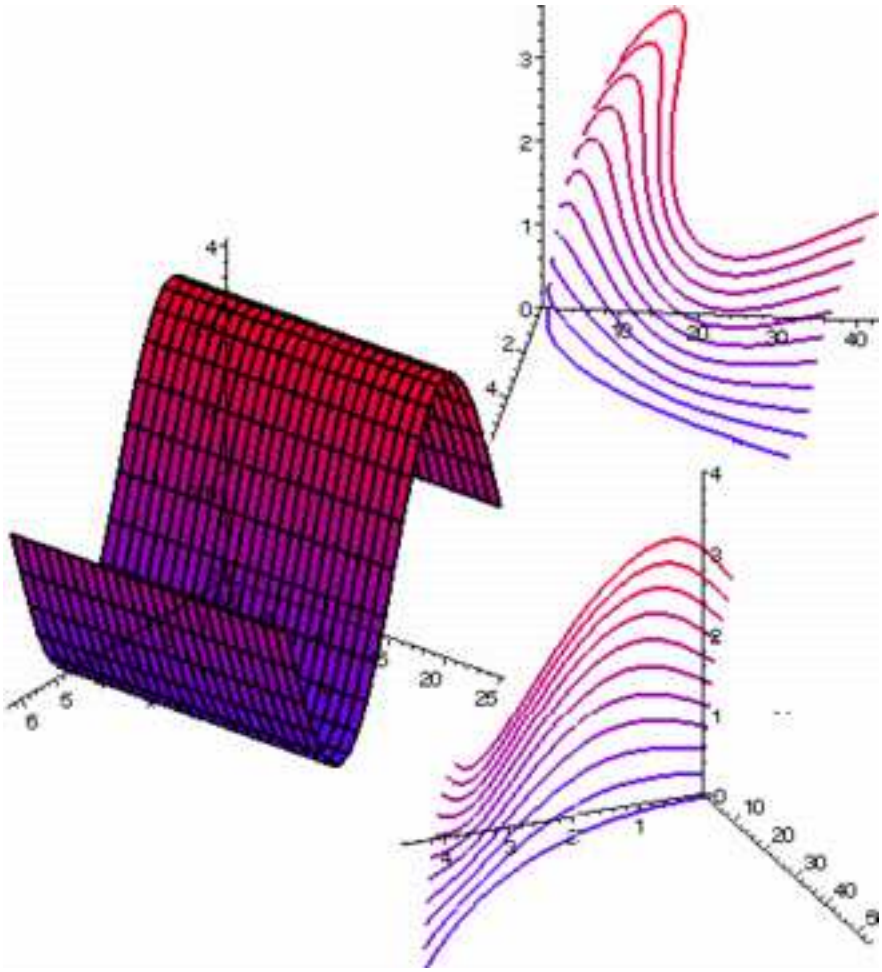
Satisfy yourself that the region has volume given by

$$\int_{X=0}^{X=2\pi} \int_{Z=0}^{Z=sin(X)+3} \int_{Y=0}^{Y=X^2+Z+1} 1 \; dZ \; dX \; dY.$$

Evaluate this integral.

It is possible to change the order of integration without much trouble in several ways, but others require that you break the region into several pieces and integrate over each piece separately. Investigate this issue.

*The graphs found below **might** help visualize the region. The picture on the left is a piece of the maximum-Z surface. The two pictures on the right are ten curves on the maximum-Y surface from $X = 0$ to $X = 2\pi$ with the topmost curve in the maximum-Z surface too. The viewpoint of the top perspective is from the larger-X side, while the viewpoint of the lower one is from the larger-Y side.*



42. Change of Variable for $3D$ Integrals

At this point we will explore parametric representations of open sets in space and the corresponding **change of variables** integration formula.

We will presume that $\boldsymbol{\mathcal{W}}$ is an open set in space and f is a vector valued function with domain $\boldsymbol{\mathcal{W}}$, $f(s,t,u) = \langle X(s,t,u), Y(s,t,u), Z(s,t,u)\rangle$.

As before, we use the notation X_i, Y_i and Z_i in place of D_iX, D_iY and D_iZ, this time for $i = 1, 2$ or 3.

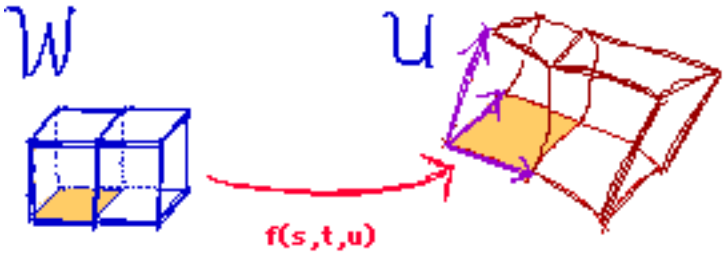
We adopt the notation $A_{t,u}(s) = B_{s,u}(t) = C_{s,t}(u) = f(s,t,u)$.

We will make a number of "niceness" presumptions[46] about f.

(i) We presume that f is one-to-one and that the collection of values $f(s,t,u)$ for (s,t,u) in $\boldsymbol{\mathcal{W}}$ constitute an open subset $\boldsymbol{\mathcal{U}}$ in space.

(ii) The functions $A_{t,u}$, $B_{s,u}$ and $C_{s,t}$ are continuously differentiable.

(iii) The number $A'_{t,u}(s) \cdot B'_{s,u}(t) \times C'_{s,t}(u)$ is never 0.



We will refer to an f like this as a **good parameterization** of the open set $\mathcal{U}$.

The curves parameterized by $A_{t,u}$, $B_{s,u}$ and $C_{s,t}$ constitute a **coordinate grid** on $\mathcal{U}$, and each point in $\mathcal{U}$ can be located uniquely as the crossing place of three of these curves. These curves act as alternative coordinate lines on $\mathcal{U}$, and a mesh of closely spaced grid curves helps one navigate around on $\mathcal{U}$ just as the ordinary $3D$ rectangular grid would. Small pieces of $\mathcal{U}$ bounded by the grid curves will not be rectangular solids necessarily but will, when small enough, look like parallelepipeds.

The condition that $A'_{t,u}(s) \cdot B'_{s,u}(t) \times C'_{s,t}(u) \neq 0$ ensures that the parallelepipeds in the grid do not get too "skinny" or, at least, never go entirely "flat."

42.1. ***Exercise***. *Verify:*

The condition from (ii) implies that f is continuously differentiable.

In fact

$$A'_{t,u}(s) = \big\langle X_1(s,t,u), Y_1(s,t,u), Z_1(s,t,u) \big\rangle$$
$$\text{and } B'_{s,u}(s) = \big\langle X_2(s,t,u), Y_2(s,t,u), Z_2(s,t,u) \big\rangle$$
$$\text{and } C'_{s,t}(u) = \big\langle X_3(s,t,u), Y_3(s,t,u), Z_3(s,t,u) \big\rangle.$$

The volume of the parallelepiped formed by $A'_{t,u}(s)$, $B'_{s,u}(t)$ and $C'_{s,t}(u)$ is $|A'_{t,u}(s) \cdot B'_{s,u}(t) \times C'_{s,t}(u)|$ and this magnitude is a continuous function on $\mathcal{W}$ and is never 0. So the sign of $A'_{t,u}(s) \cdot B'_{s,u}(t) \times C'_{s,t}(u)$ is constant on $\mathcal{W}$, a fact that will be useful in calculations.

For those of you who know about determinants,

$$A'_{t,u}(s) \cdot B'_{s,u}(t) \times C'_{s,t}(u) = \det f'(s,t,u),$$

a fact that shortens notation considerably.

42.2. ***Exercise***. ** Show that if f is a good parameterization of $\mathcal{U}$ then $g = f^{-1}$ is a good parameterization of $\mathcal{W}$.*

By analogy with the earlier section on parametric descriptions of plane sets we will create an integral formula for volume involving the parameterization.

We presume that f is a good parameterization of $\mathcal{U}$ defined on $\mathcal{W}$. We will presume that the domain $\mathcal{W}$ of f is inside the rectangular solid $[a_1, b_1] \times [a_2, b_2] \times [a_3, b_3]$ in space. Suppose P is a partition of this rectangular solid composed of subrectangular solids $[s_{i-1}, s_i] \times [t_{j-1}, t_j] \times [u_{k-1}, u_k]$ for $i = 1 \dots n$ and $j = 1 \dots m$ and $k = 1 \dots p$.

Suppose $[s_{i-1}, s_i] \times [t_{j-1}, t_j] \times [u_{k-1}, u_k]$ is entirely contained in $\mathcal{W}$.

$\Delta s_i \; A'_{t_{j-1},u_{k-1}}(s_{i-1})$ and $\Delta t_i \; B'_{s_{i-1},u_{k-1}}(t_{j-1})$ and $\Delta u_k \; C'_{s_{i-1},t_{j-1}}(u_{k-1})$ are all vectors in space.

If the mesh of the partition is small enough, the parallelepiped formed by these three vectors will be very close to the small part of $\mathcal{U}$ covered by f on the rectangular solid $[s_{i-1}, s_i] \times [t_{j-1}, t_j] \times [u_{k-1}, u_k]$.

This parallelepiped has volume

$$\begin{aligned}&\Delta s_i \Delta t_j \Delta u_k \; |A'_{t_{j-1},u_{k-1}}(s_{i-1}) \cdot B'_{s_{i-1},u_{k-1}}(t_{j-1}) \times C'_{s_{i-1},t_{j-1}}(u_{k-1})| \\ =&\Delta s_i \Delta t_j \Delta u_k \; |\det f'(s_{i-1}, t_{j-1}, u_{k-1})|.\end{aligned}$$

Suppose h is a continuous and bounded function defined on $\mathcal{U}$. We can imagine that h represents charge or mass density, for example, at each point in $\mathcal{U}$. h will be nearly constant on this tiny piece of $\mathcal{U}$, so the total **charge** or **mass** on this piece will be nearly the volume of the parallelepiped multiplied by $h(f(s_{i-1}, t_{j-1}, u_{k-1}))$.

Adding together these numbers for all the partition rectangular solids entirely contained in $\mathcal{W}$ generates a Riemann sum which is an approximation to the total charge or mass on $\mathcal{U}$ and yields the integral

$$\int_{\mathcal{W}} h(f(s,t,u)) \; |\det f'(s,t,u)| \; ds \; dt \; du$$

which **should**, if our thinking is correct, be the same as

$$\int_{\mathcal{U}} h(X,Y) \; dX \; dY \; dZ.$$

In this context, this integral on $\mathcal{W}$ is called a **parametric form for the integral of h on $\mathcal{U}$**.

Though plausible, proof that the two integrals agree must be found elsewhere.

It is common to see the collapsed notation

$$\int_{\mathcal{W}} h \circ f \; |\det f'| \; ds \; dt \; du \; = \; \int_{\mathcal{U}} h \; dX \; dY \; dZ$$

to denote the equality of these two integrals.

When $h = 1$, the constant function, we have a representation of the **volume** of $\mathcal{U}$ as

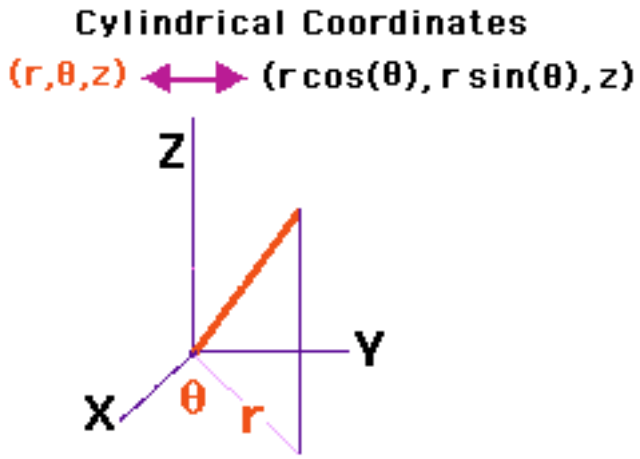
$$\int_{\mathcal{U}} 1 \; \; dX \; dY \; dZ \; = \; \int_{\mathcal{W}} |\det f'(s,t,u)| \; ds \; dt \; du.$$

42.3. ***Exercise***. ***Cylindrical Coordinates*** *are an alternative coordinate system defined as follows: Let* $\mathcal{W}$ *denote the set of triples* (r, θ, z) *with* $r > 0$ *and* $0 < \theta < 2\pi$ *and* z *any real number. Let* $\mathcal{U}$ *denote those points in space excluding that part of the* XZ *plane where* $X \geq 0$.

Define the function C *from* $\mathcal{W}$ *to* $\mathcal{U}$ *by*

$$C(r, \theta, z) = \langle r\ cos(\theta),\ r\ sin(\theta),\ z \rangle .$$

C is, essentially, ordinary polar coordinates with the third dimension tacked on.



(i) Prove that this is an orthogonal coordinate system.

(ii) Show that $\det\ C'(r, \theta, z) = r$.

(iii) Identify the part of $\mathcal{W}$ *that corresponds to the portion of the right circular cylinder with base on the* XY *plane whose axis is the* Z *axis with radius* 2 *and height* 7 *which is in* $\mathcal{U}$.

(iv) Find the volume of the cylinder described above in two ways by integrating with respect to each coordinate system.

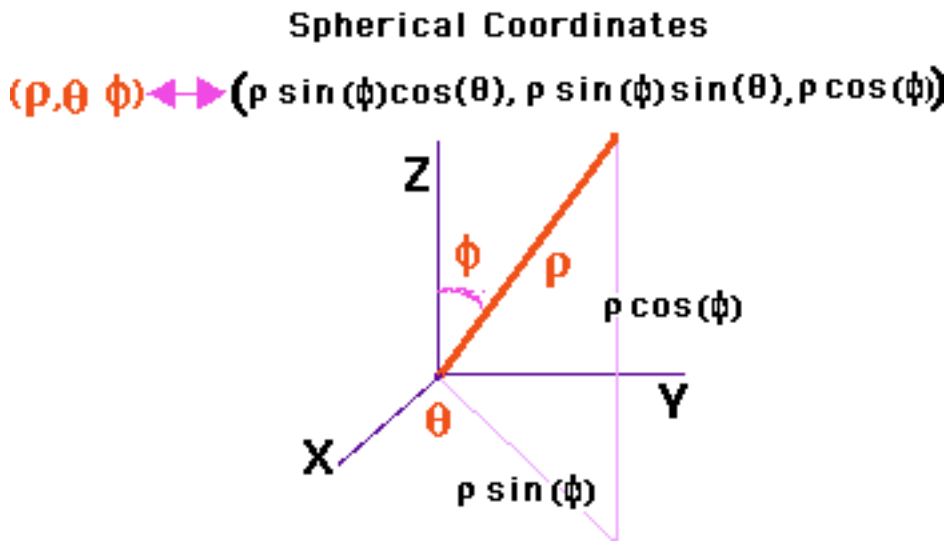
(v) Suppose the cylinder above is filled with a gas whose density decreases exponentially with height: the density at point (X, Y, Z) *is* e^{-Z}. *Calculate the mass in two ways as in (iv).*

42.4. ***Exercise***. ***Spherical Coordinates*** *are an alternative coordinate system defined as follows: Let* $\mathcal{W}$ *denote the set of triples* (ρ, θ, ϕ) *with* $\rho > 0$ *and* $0 < \theta < 2\pi$ *and* $0 < \phi < \pi$. *Let* $\mathcal{U}$ *denote those points in space excluding that part of the* XZ *plane where* $X \geq 0$.

Define the function S *from* $\mathcal{W}$ *to* $\mathcal{U}$ *by*

$$S(\rho, \theta, \phi) = \langle\ \rho\ sin(\phi)\ cos(\theta),\ \rho\ sin(\phi)\ sin(\theta),\ \rho\ cos(\phi)\ \rangle .$$

The r from cylindrical coordinates is $\rho\ sin(\phi)$ here, and ρ is the distance from the image point in $\mathcal{U}$ *to the origin.*



(i) Prove that spherical coordinates are an orthogonal coordinate system.

(ii) Show that $\det\ S'(\rho,\theta,\phi) = -\rho^2\ sin(\phi)$.

(iii) Identify the part of $\mathcal{W}$ *that corresponds to the portion of the sphere centered at the origin with radius* 5 *which is in* $\mathcal{U}$.

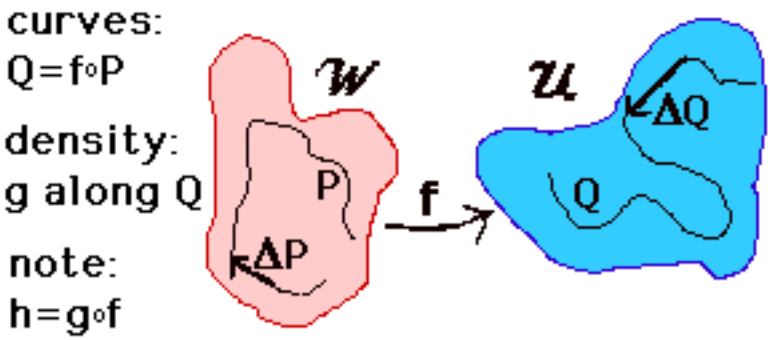
(iv) Find the volume of the sphere described above in two ways by integrating with respect to the spherical coordinate system and then the ordinary rectangular coordinate system.

(v) Suppose the sphere above is filled with a gas whose density decreases exponentially with height: the density at point (X,Y,Z) *is* e^{-Z}. *Calculate the mass in two ways as in (iv).*

42.5. ***Exercise***. *In this problem we investigate a change of variables formula for arclength and line integrals which is a little different in flavor from the one implied by Exercise 18.2. We delayed the discussion of this topic until coordinate changes in space had been considered.*

Let $\mathcal{W}$ *and* $\mathcal{U}$ *be open sets, both in the plane or both in space. Let* f *be a good parameterization of* $\mathcal{U}$ *with domain* $\mathcal{W}$. *Finally, suppose* P *is a good parameterization of a curve in* $\mathcal{W}$ *with domain* $[c,d]$ *and* $Q = f \circ P$ *is the corresponding good parameterization of a curve in* $\mathcal{U}$. *Then the arclength as calculated using* Q *is*

$$\int_c^d |Q'(t)|\ dt = \int_c^d |f'(P(t))P'(t)|\ dt = \int_c^d \left| f'(P(t))\frac{P'(t)}{|P'(t)|} \right|\ |P'(t)|\ dt.$$



More generally, suppose g is a function defined for points along Q. g represents, perhaps, a linear charge density present on the wire in $\mathcal{U}$ parameterized by Q. If $g \circ Q$ is continuous then $h = g \circ f$ is defined for points along P and $h \circ P$ is continuous. So

$$\int_c^d g(Q(t))\ |Q'(t)|\ dt = \int_c^d h(P(t))\ |f'(P(t))P'(t)|\ dt$$
$$= \int_c^d h(P(t))\ \left| f'(P(t))\frac{P'(t)}{|P'(t)|} \right|\ |P'(t)|\ dt.$$

Let's consider this integral from the standpoint of a witness in $\mathcal{W}$ who can see the three factors in the integrand and who knows that this integral corresponds to a line integral in $\mathcal{U}$.

The witness recognizes the number $h(P) = g(Q)$ to be the linear density as perceived by an observer in $\mathcal{U}$.

The factor $|P'|$ is the standard arclength weighting as measured along P in $\mathcal{W}$.

The factor $\left| f'(P(t))\frac{P'(t)}{|P'(t)|} \right|$ is the interesting one. It's function is to change the density on the curve P so that corresponding pieces ΔP and ΔQ for the same short time increment Δt "weigh" the same. If ΔP points in a direction which f magnifies greatly then ΔQ will be much longer than ΔP. The density on ΔP must be magnified by this same factor if ΔP is to weigh the same as ΔQ. Note that this magnification factor depends only on the direction of P' and not its magnitude. It is the "stretch factor" imparted by f to any vector pointing in the same direction as $P'(t)$ at P.

If we approximate both integrals with Riemann sums using the same partition of $[c, d]$, the necessity for this specific factor to change density becomes more obvious.

$$\sum_{i=1}^n g(Q_i)\ |\Delta Q_i| \approx \sum_{i=1}^n h(P_i)\ \left| f'(P_i)\ \frac{\Delta P_i}{|\Delta P_i|} \right|\ |\Delta P_i|.$$

42.6. ***Exercise***. *** In this problem we investigate for surface integrals a situation analogous to the one given above for line integrals.*

Let f be a good parameterization between open sets in space. Suppose P is a good parameterization of a surface in space with domain $\mathfrak{D}$ in the plane and with values in the domain of f.

Define $Q = f \circ P = \langle X, Y, Z \rangle$. So Q is a good parameterization of a surface too, with the same domain as P.

Define $Q_1 = \langle X_1, Y_1, Z_1 \rangle$ and $Q_2 = \langle X_2, Y_2, Z_2 \rangle$ and let $k(X, Y, Z)$ denote a density on the surface given by Q.

As we saw, the surface integral of k on this surface is

$$\int_{\mathfrak{D}} k \, |Q_1 \times Q_2| \, ds \, dt.$$

Create a surface integral on the surface given by P with the same value and interpret the meaning of the modified density function in the manner of the last exercise.

---

43. Divergence Theorem and Stokes' Theorem

Every book must end and you have arrived at the end of this one. In this section you will see definitions for a few of the cast of characters you will encounter should you decide to go further in mathematics. They are easy definitions, but we had no reason to look at them before now. We will do some calculations of the kind you already have seen in this and earlier chapters. We will observe that several of these calculations give the same answer, even though they seem at the outset to refer to very different integrals.

Elsewhere in this chapter we proved as many of the key results as possible given the knowledge base assumed of the readership. Here there will be very few proofs and quite a few vague appeals to intuition and philosophical ruminations. Nevertheless, I hope the patterns we look at strike some as intriguing and prompt further investigation.

In thinking about this section consider the following fact from basic Calculus.

If f is a continuous function defined on an interval $[a, b]$ and differentiable on (a, b) then

$$f(b) - f(a) = \int_a^b f'(x) \, dx \approx \sum_{i=1}^{n} f'(x_i) \, \Delta x_i.$$

(The approximation is good when the mesh of the partition is small enough.)

When the integral was defined there was an orientation involved. First, f' measures how fast f is changing when you move in a particular direction on the interval. Second, the number a is the boundary on the "downward end" of the interval, while b is the boundary on the "upward end" of the interval. That is why $f(b)$ is added while $f(a)$ is subtracted in the combination on the left.

Note that the interval itself is one dimensional, while the boundary is just a couple of points. You might call the boundary zero dimensional.

What we have, essentially, is two different types of oriented sums. One is an oriented combination of the values of f over the zero dimensional *boundary* of the interval. The other is a sum, over the one dimensional interval itself, of the *oriented rate of change* of the function.

The *accumulated oriented changes over the interval* combine to give an *oriented combination of values on the boundary.*

We emphasize that the orientation is critical here. $f(a) - f(b)$ would be the wrong choices for combining the values of f on the boundary. This combination must be matched with the orientation along the interval with respect to which f' is defined.

The **mantra** for this section, the thing to remember, is this: the accumulation of an oriented rate of change over a region can equal the (proper) oriented accumulation of the function itself on the boundary of the region.

There is another "one-zero" dimensional example of this which we saw in the context of potentials.

Suppose g is a differentiable function in the plane or in space and P is a piecewise good parameterization of a curve $\mathfrak{C}$ on the interval $[a, b]$ in the domain of g. Let $A = P(a)$ and $B = P(b)$. Let $\mathfrak{T}$ denote the unit tangent to the curve corresponding to the orientation of P. Let F denote the vector field ∇g.

The curve $\mathfrak{C}$ is one dimensional, oriented by P in a specific direction. The boundary of $\mathfrak{C}$ consists of the points A and B. This boundary is oriented, with A at "the start" and B at "the end."

F can be used to find the rate of change of g in various directions.

We saw that

$$g(B) - g(A) = \int_{\mathfrak{C}} F(s) \cdot \mathfrak{T}(s)\, ds.$$

Once again, the oriented combination of the values of g on the boundary of $\mathfrak{C}$ is an integral involving the oriented rate of change of g over $\mathfrak{C}$.

Our next step will be to increase the dimension a bit. But first we need to define a new player.

If $F = \langle M, N\rangle$ or $\langle M, N, P\rangle$ is a differentiable vector field in the plane or in space we define $\nabla \cdot \mathbf{F}$ to be

$$\nabla \cdot F = D_1 M + D_2 N \quad \text{or} \quad \nabla \cdot F = D_1 M + D_2 N + D_3 P$$

whichever is appropriate. This is called the **divergence of F** and the notation **div F** is a rather common alternative notation for the divergence of F.

This looks a lot like one of the notations for the gradient, but the gradient takes a function and gives you a vector field, while the divergence takes a vector field and generates a function. The "$\cdot$" distinguishes the two notationally.

At this point it is traditional to introduce the mnemonically useful "Del Operator" written, depending on dimension, $\nabla = \langle D_1, D_2\rangle$ or $\langle D_1, D_2, D_3\rangle$. We will not dwell on the potential meaning of ∇ standing alone, but instead think of it as a notational tool, helping us to remember that (in three dimensions)

$$div\ F = \nabla \cdot F = \langle D_1, D_2, D_3\rangle \cdot \langle M, N, P\rangle = D_1 M + D_2 N + D_3 P$$

and

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

Now let's consider the two dimensional case for a minute and try to understand the meaning of $\nabla \cdot F$ for continuously differentiable F, and what it means for it to be positive, negative, large or small.

We will think of F as a velocity field in the plane, representing the direction and rate of flow of a layer of some substance swirling around on the plane. It is quite possible for locations in the plane to be "sources" or "sinks" of whatever this material is: more of this material can leave a small region than enters that region, or conversely.

There are a couple of easy physical models for this. In one model, there is a chemical reaction going on which takes place at different rates and even goes in different directions due to variations in temperature or the presence of catalysts, reactants and so on distributed anisotropically around on the surface. F can be tracking one of the chemical participants.

In a second model you could imagine that we are dealing with a single substance but that the plane is perforated by millions of microscopic pinholes. Each one is attached on the other side of the plane to a microscopic pump or suction device, pulling material off the surface or pumping more on at various places.

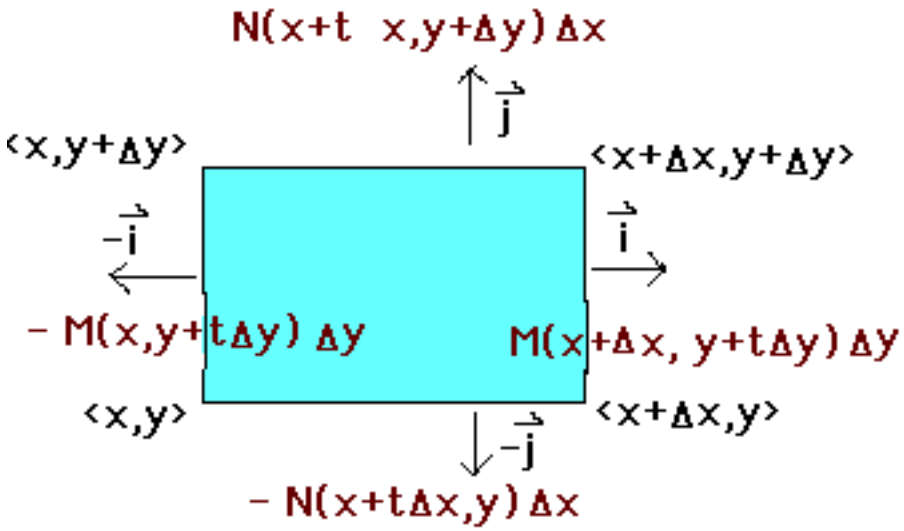
Let's parameterize a very tiny rectangle in the plane with corners at

$$\langle X,\ Y \rangle,\ \langle X + \Delta X,\ Y \rangle,\ \langle X,\ Y + \Delta Y \rangle \text{ and } \langle X + \Delta X,\ Y + \Delta Y \rangle$$

as shown in the diagram.

When F is continuously differentiable the flux of $F = \langle M, N \rangle$ past this piecewise good loop in the outward direction is

$$\begin{aligned}
&\int_0^1 (\ M(X + \Delta X,\ Y + t\ \Delta Y) - M(X,\ Y + t\ \Delta Y)\ )\ \Delta Y\ dt \\
&\quad + \int_0^1 (\ N(X + t\ \Delta X,\ Y + \Delta Y) - N(X + t\ \Delta X,\ Y)\ )\ \Delta X\ dt \\
&= \Delta X \Delta Y \int_0^1 \frac{M(X + \Delta X,\ Y + t\ \Delta Y) - M(X,\ Y + t\ \Delta Y)}{\Delta X}\ dt \\
&\quad + \Delta X \Delta Y \int_0^1 \frac{N(X + t\ \Delta X,\ Y + \Delta Y) - N(X + t\ \Delta X,\ Y)}{\Delta Y}\ dt.
\end{aligned}$$



For each t the integrand

$$\frac{M(X+\Delta X,\ Y+t\ \Delta Y)-M(X,\ Y+t\ \Delta Y)}{\Delta X}=D_1M(\overline{X},Y+t\ \Delta Y)$$

for some $\overline{X}$ between X and $X+\Delta X$ by the Mean Value Theorem.

Similarly, for each t there is $\overline{Y}$ between Y and $Y+\Delta Y$ with

$$D_2N(\overline{Y})=\frac{N(X+t\ \Delta X,\ Y+\Delta Y)-N(X+t\ \Delta X,\ Y)}{\Delta Y}.$$

We have, therefore, shown that

$$Min\ D_1M+Min\ D_2N\le\frac{\text{Flux Out of Rectangle}}{\text{Area of Rectangle}}\le Max\ D_1M+Max\ D_2N$$

where "Min" and "Max" denote the minimum and maximum vales of the indicated derivatives on the rectangle.

The continuity of these derivatives guarantees that the ratio converges to the divergence $D_1M(X,Y)+D_2N(X,Y)$ as the edge sizes both go to zero.

So the value of the divergence at a point in the plane can be interpreted as the "outward flux production rate per unit area" over very small areas near the point. Where it is positive, material is being produced or introduced. Where it is negative, material is being removed or destroyed.

Now suppose we find the double integral of $D_1M(X,Y)+D_2N(X,Y)$ over a bounded plane region $\mathcal{S}$ bounded by a piecewise good loop $\mathcal{C}$. We are calculating the total outward flux production over the region. If our thinking is correct, this must be the flux leaving the region. So if $\mathcal{N}$ is the outward normal vector for $\mathcal{C}$ we should have

$$\int_{\mathcal{S}}\nabla\cdot F\ dX\ dY=\int_{\mathcal{C}}F(s)\cdot\mathcal{N}(s)\ ds.$$

Our speculation is that integral of an oriented rate of change of a function over a two dimensional set should equal an oriented integral of the function itself

over the boundary—in this case an oriented one dimensional curve. This is another instance of the "theme" idea of this section. The statement specifying the condition under which the argument for equality suggested above can be solidified constitutes the **Normal Form of Green's Theorem** or the **Divergence Theorem in the Plane.**

43.1. ***Exercise***. *Let $\mathcal{S}$ denote the standard unit square in first quadrant of the XY plane with boundary $\mathcal{C}$ and outward normal $\mathcal{N}$. Let $F(X,Y) = \langle X^2, XY\rangle$. Show that*

$$\int_{\mathcal{S}} \nabla \cdot F \; dX \; dY = \int_{\mathcal{C}} F(s) \cdot \mathcal{N}(s) \; ds.$$

43.2. ***Exercise***. *Suppose we have a continuously differentiable vector field F over a solid region in space $\mathcal{W}$ completely surrounded by (that is, bounded by) an oriented composite surface $\mathcal{S}$ with outward normal $\mathcal{N}$. Using small rectangular solids, mimic the argument in $2D$ found above to conclude that $\nabla \cdot F$ represents the rate of outward flux production per unit volume. Conclude that*

$$\int_{\mathcal{W}} \nabla \cdot F \; dX \; dY \; dZ = \int_{\mathcal{S}} F(\mathcal{S}) \cdot \mathcal{N}(\mathcal{S}) \; d\mathcal{S}.$$

The staement which identifies a collection of conditions under which this formula is valid is called the ***Divergence Theorem.***

43.3. ***Exercise***. *Let $\mathcal{W}$ denote the solid standard unit cube in the all-positive octant of space. Let F be the vector field $F = \langle Z, Y^2, XY\rangle$. Calculate the integral of $\nabla \cdot F$ over this cube. Compare your result to your calculation from Exercise 40.3.*

In the next example we need to define the **curl** of a differentiable vector field $F = \langle M, N, P\rangle$ defined on an open set in space. This is a new vector field defined to be

$$\nabla \times \mathbf{F} = \langle D_2P - D_3N, \; D_3M - D_1P, \; D_1N - D_2M\rangle.$$

The reason for the "$\nabla\times$" notation is to extend the mnemonic device we saw before: if you treat the symbol ∇ *as if* it were a vector $\langle D_1, D_2, D_3\rangle$ then the curl of F "is" the cross product indicated by $\nabla \times F$.

You will sometimes see the notation **curl F** or **rot F** to denote the curl of F.

The curl involves sums of derivatives as did the divergence. We have an interpretation of divergence as representing a local rate of production of outward flux per unit volume. Prompted by the discussions above we should be inclined to regard curl as representing the local rate of production of something too ... but what?

Let's consider a small rectangle in space with one edge along the segment between (X,Y,Z) and $(X+\Delta X, Y, Z)$ and a second edge between (X,Y,Z) and

$(X, Y + \Delta Y, Z)$ and let's presume that the increments ΔX and ΔY are greater than 0. Suppose also that F is continuously differentiable on this rectangle.

The boundary is a piecewise good curve which we orient by a parameterization which traverses the segment from (X, Y, Z) to $(X + \Delta X, Y, Z)$, in that order, first.

Let's calculate the circulation of F around the boundary of this rectangle.

$$Q(t) = \begin{cases} \langle X + t\Delta X, Y, Z\rangle, & \text{if } 0 \le t < 1; \\ \langle X + \Delta X, Y + (t-1)\Delta Y, Z\rangle, & \text{if } 1 \le t < 2; \\ \langle X + (3-t)\Delta X, Y + \Delta Y, Z\rangle, & \text{if } 2 \le t < 3; \\ \langle X, Y + (4-t)\Delta Y, Z\rangle, & \text{if } 3 \le t < 4. \end{cases}$$

Form the integral $\int_1^4 F(Q(t)) \cdot Q'(t)\ dt$ as the sum of four integrals and change variables yielding

$$\Delta Y \int_0^1 N(X + \Delta X, Y + t\Delta Y, Z) - N(X, Y + t\Delta Y, Z)\ dt$$
$$-\Delta X \int_0^1 M(X + t\Delta X, Y + \Delta Y, Z) - M(X + t\Delta X, Y, Z)\ dt.$$

By the Mean Value Theorem, for each t there is a value $\overline{X}$ between X and $X + \Delta X$ and $\overline{Y}$ between Y and $Y + \Delta Y$ with

$$D_1 N(\overline{X}, Y + t\Delta Y, Z) = \frac{N(X + \Delta X, Y + t\Delta Y, Z) - N(X, Y + t\Delta Y, Z)}{\Delta X}$$

and

$$D_2 M(X + t\Delta X, \overline{Y}, Z) = \frac{M(X + t\Delta X, Y + \Delta Y, Z) - M(X + t\Delta X, Y, Z)}{\Delta Y}$$

.

By continuity of these derivatives as before we conclude that when both ΔX and ΔY are small

$$\frac{\int_1^4 F(Q(t)) \cdot Q'(t)\ dt}{\Delta X \Delta Y} \approx D_1 N(X, Y, Z) - D_2 M(X, Y, Z).$$

This is the Z component of the curl of F. We conclude that, at least when measured with rectangles with sides parallel to the X and Y axes and boundary oriented properly, the Z component of the curl of F can be interpreted as the rate of circulation production per unit area.

A similar interpretation holds for rectangles perpendicular to the other two axes. In fact, far more generally, if N is any unit vector $\nabla F \cdot N$ can be interpreted as the rate of circulation production per unit area over any surface (with a piecewise good boundary curve) perpendicular to N, which leads to the following speculation.

Suppose $\mathfrak{S}$ is an oriented surface with a unit normal $\mathfrak{N}$ and a piecewise good boundary curve $\mathfrak{C}$ with unit tangent $\mathfrak{T}$. Under what conditions, if any, will the following formula prove to be correct?

$$\int_{\mathfrak{S}} \nabla F(\mathfrak{S}) \cdot \mathfrak{N}(\mathfrak{S})\ d\mathfrak{S} = \int_{\mathfrak{C}} F(s) \cdot \mathfrak{T}(s)\ ds.$$

This equation does hold in many cases, and the precise statement of the result is called **Stokes' Theorem**. Once again, it says that an integral over a surface of an oriented change rate can be found by calculating an oriented integral over the boundary curve.

43.4. ***Exercise***. *Suppose* $F = \langle XY,\ Y,\ XZ\rangle$. *Calculate the surface integral of* $\nabla \times F \cdot \mathcal{N}$ *where* $\mathcal{N}$ *is the upward unit normal to the upper hemisphere of the standard unit sphere in space. Compare that to the circulation of* F *around the boundary circle in the* XY *plane.*

The curl is not defined for vector fields in two dimensions. However by changing the point of view we can do something with them. If $F(X,Y) = \langle M(X,Y),\ N(X,Y)\rangle$ is a vector field in the plane, define the vector field $\overline{F}$ in space by $\overline{F}(X,Y,Z) = \langle M(X,Y),\ N(X,Y),\ 0\rangle$. The field $\overline{F}$ looks just like F on the XY plane in space. And $\nabla \times \overline{F} = \langle 0,\ 0,\ D_1N - D_2M\rangle$.

Let's suppose that $\mathcal{S}$ is a region with piecewise good boundary $\mathcal{C}$ with unit tangent $\mathcal{T}$ in the plane and F is a continuously differentiable vector field in the plane.

By applying Stokes' Theorem to the plane region thought of as the XY plane in $3D$ we get, in many circumstances,

$$\int_{\mathcal{S}} \nabla\overline{F}(\mathcal{S}) \cdot \vec{k}\ d\mathcal{S} = \int_{\mathcal{S}} D_1N(\mathcal{S}) - D_2M(\mathcal{S})\ d\mathcal{S} = \int_{\mathcal{C}} F(s) \cdot \mathcal{T}(s)\ ds.$$

Conditions for equality of these last two integrals comprise the **Tangential Form of Green's Theorem** or **Stokes' Theorem Theorem in the Plane.**

43.5. ***Exercise***. *Calculate the integral of* $\overline{F} \cdot \vec{k}$ *over the unit circle for the field given by* $F(X,Y) = \langle -Y, X\rangle$. *Compare this result to the one you obtained in Exercise 24.2.*

In the early chapters we made reference to foundational material upon which various results, such as the intermediate value theorem for continuous functions, depend. In the earlier sections of this chapter, the equality of double and triple integrals with their brethren in parametric form was a notable hole, waiting to be filled, in the structure we have built. Now, in this concluding section we have named and discussed numerous results and proved virtually nothing, though the immediate utility of the results as stated here is enormous, in applications throughout Engineering and Physics.

In your further studies, as you come to fill in the holes we have left, you will touch bases with most branches of modern mathematics. Sometimes you will see several versions of these results, with different proofs, each illuminating different aspects of the structure which underlies it all.

As your understanding grows, that which is illuminated will depend largely on you—the facts you have seen before, your esthetic sense and the connections you

are able to make. Don't forget to step back, from time to time, to appreciate the beauty of your creation as you build.

Endnotes

Note 1, From Page 44:

The Maple 6 command (this and all subsequent Maple commands are to be typed on a command line after the command prompt) that will generate a (rotatable) graphic of this cone is:

$with(plots):\ implicitplot3d(\ x\hat{}2 + y\hat{}2 = z\hat{}2,$
$x = -2..2,\ y = -2..2,\ z = -2..2,\ grid = [30, 30, 30],\ axes = NORMAL,$
$shading = Z,\ style = PATCHCONTOUR);$

Note 2, From Page 45:

The Maple commands for the cylinder:

$with(plots):\ implicitplot3d(\ x\hat{}2 + y\hat{}2 = 1,$
$x = -2..2,\ y = -2..2,\ z = -2..2,\ grid = [30, 30, 30],\ axes = NORMAL,$
$shading = Z, style = PATCHCONTOUR);$

Note 3, From Page 45:

Maple will draw you a picture of a helicoid with:

$with(plots):\ implicitplot3d(\ y = z * x,$
$x = -2..2,\ y = -2..2,\ z = -2..2,\ grid = [30, 30, 30],\ axes = NORMAL,$
$shading = Z,\ style = PATCHCONTOUR);$

Note 4, From Page 45:

An "off-center bump" surface:

$with(plots):\ implicitplot3d(\ (x-1)\hat{}2 + 2 * (y+2)\hat{}2 = -z + 1,$
$x = -4..4,\ y = -4..4,\ z = -2..2,\ grid = [30, 30, 30],\ axes = NORMAL,$
$shading = Z,\ style = PATCHCONTOUR);$

Note 5, From Page 46:

The Maple command which will let you visualize the ellipsoid is:

$with(plots):\ implicitplot3d(\ x\hat{}2 + 2 * y\hat{}2 + 3 * z\hat{}2 = 1,$
$x = -1..1,\ y = -1..1,\ z = -1..1,\ grid = [30, 30, 30],\ axes = NORMAL,$
$shading = Z,\ style = PATCHCONTOUR);$

Note 6, From Page 46:

The Maple command which will let you work with the torus is:

$with(plots):\ implicitplot3d(\ (\ x\hat{}2 + y\hat{}2 + z\hat{}2\ -\ 4.25\ )$
$= 16 * (\ .25\ -\ z\hat{}2\ ),\ x = -3..3,\ y = -3..3,\ z = -1..1,$
$grid = [30, 30, 30],\ axes = NORMAL,$
$shading = Z,\ style = PATCHCONTOUR);$

Note 7, From Page 47:

Maple will generate a monkey saddle with:

$with(plots):\ implicitplot3d(\ x\hat{}3 - x * y\hat{}2 = z,$
$x = -1..1,\ y = -1..1,\ z = -1..1,\ grid = [30, 30, 30],\ axes = NORMAL,$
$shading = Z,\ style = PATCHCONTOUR);$

Note 8, From Page 47:

Maple will generate "fat axes" with:

$with(plots):\ implicitplot3d(\ x\hat{}2 * y\hat{}2 +\ x\hat{}2 * z\hat{}2\ +\ z\hat{}2 * y\hat{}2\ = 1,$
$x = -3..3,\ y = -3..3,\ z = -3..3,\ grid = [30, 30, 30],\ axes = NORMAL,$
$shading = Z,\ style = PATCHCONTOUR);$

Note 9, From Page 55:

Here are the Maple commands for the curve $Q(t) = \langle t, t^2, t^3 \rangle$:
$with(plots):\ spacecurve(\ [t, t\hat{}2, t\hat{}3\ ],$
$t = -1..2, axes = NORMAL, thickness = 2, color = GREEN);$

Note 10, From Page 55:

The following will get you the curve found above together with a helix $H(t) = \langle cos(t), sin(t), t \rangle$
$with(plots):\ spacecurve(\ \{\ [t, t\hat{}2, t\hat{}3\ , t = -1..2, color = GREEN],$
$[cos(t), sin(t), t, t = -3..8, color = RED\ ]\ \}, axes = NORMAL, thickness = 2);$

Note 11, From Page 60:

Linear plus Circular plus Circular: The numbers to modify are the first five in parentheses.
$with(plots):$
$(\ S,\ RadiusOne,\ FreqOne,\ RadiusTwo,\ FreqTwo\ ) := (\ 2 * Pi,\ 1,\ -1,\ 0,\ 0\ );$
$Q := evalm([S * t, 0]$
$+ [RadiusOne * cos(2 * Pi * FreqOne * t), RadiusOne * sin(2 * Pi * FreqOne * t)] +$
$[RadiusTwo * cos(2 * Pi * FreqTwo * t), RadiusTwo * sin(2 * Pi * FreqTwo * t)]);$
$plot([Q[1], Q[2], t = 0..2]);$

Note 12, From Page 62:

Here are the Maple instructions that will draw a graph of 101 data points which will recreate the picture in the text. The first part is to simply name 101 points. The interesting part is the next section where we define the function *linint* which **linearly interpolates** over the consecutive time intervals between positions. This could be replaced by a different interpolation scheme, such as one which smoothes out corners. The last part is just the plot instruction. There are several ways to modify this to plot your own points. If you want to change the number of points you must modify $NumOfPoints$. In the plot commands you must modify the range of t to whatever interval your times include. If you want to plot $3D$ points the modifications are a little more extensive. You must include a third coordinate in *linint* and the initial points will have four coordinates. You then must use the *spacecurve* command to plot your $3D$ $linint(t)$.

I would also like to point out that no attempt was made to optimize the calculation, as would surely be necessary if there were many points involved. In the instructions below I ask Maple to graph the sum of 101 different functions. The "characteristic function" denoted in the Maple instruction $charfcn$ wipes out 100 of them leaving only 1 of them nonzero on each time interval between consecutive data points. It should be possible to improve the speed of the calculation by at least a factor of 50 (that is $N/2$ where N is the number of data points.)

$restart : with(plots) : NumOfPoints := 101 :$
$P :=$ [[0.00 , 5.3 , 0.] , [0.01 , 5.35 , 1.11] ,
[0.02 , 5.14 , 2.13] , [0.03 , 4.73 , 2.97] , [0.04 , 4.20 , 3.60] ,
[0.05 , 3.64 , 4.04] , [0.06 , 3.10 , 4.34] , [0.07 , 2.55 , 4.55] ,
[0.08 , 2.00 , 4.70] , [0.09 , 1.39 , 4.78] , [0.10 , 0.77 , 4.76] ,
[0.11 , 0.04 , 4.56] , [0.12 , −.70 , 4.14] , [0.13 , −1.35 , 3.47] ,
[0.14 , −1.82 , 2.58] , [0.15 , −2.06 , 1.54] , [0.16 , −2.00 , 0.454] ,
[0.17 , −1.68 , −.600] , [0.18 , −1.15 , −1.53] , [0.19 , −.48 , −2.31] ,
[0.20 , 0.26 , −2.94] , [0.21 , 1.02 , −3.46] , [0.22 , 1.79 , −3.91] ,
[0.23 , 2.65 , −4.35] , [0.24 , 3.61 , −4.73] , [0.25 , 4.7 , −5.] ,
[0.26 , 5.91 , −5.09] , [0.27 , 7.17 , −4.93] , [0.28 , 8.39 , −4.49] ,
[0.29 , 9.46 , −3.82] , [0.30 , 10.3 , −2.94] , [0.31 , 11.0 , −1.95] ,
[0.32 , 11.3 , −.954] , [0.33 , 11.5 , −.028] , [0.34 , 11.5 , 0.806] ,
[0.35 , 11.5 , 1.54] , [0.36 , 11.3 , 2.22] , [0.37 , 11.2 , 2.89] ,
[0.38 , 10.9 , 3.56] , [0.39 , 10.4 , 4.20] , [0.40 , 9.83 , 4.76] ,
[0.41 , 9.09 , 5.14] , [0.42 , 8.19 , 5.28] , [0.43 , 7.27 , 5.13] ,
[0.44 , 6.42 , 4.70] , [0.45 , 5.76 , 4.04] , [0.46 , 5.32 , 3.24] ,
[0.47 , 5.09 , 2.39] , [0.48 , 5.04 , 1.55] , [0.49 , 5.13 , 0.759] ,
[0.50 , 5.3 , 0.] , [0.51 , 5.53 , −.759] , [0.52 , 5.84 , −1.55] ,
[0.53 , 6.29 , −2.39] , [0.54 , 6.92 , −3.24] , [0.55 , 7.76 , −4.04] ,
[0.56 , 8.82 , −4.70] , [0.57 , 10.1 , −5.13] , [0.58 , 11.4 , −5.28] ,
[0.59 , 12.7 , −5.14] , [0.60 , 13.8 , −4.76] , [0.61 , 14.8 , −4.20] ,
[0.62 , 15.7 , −3.56] , [0.63 , 16.4 , −2.89] , [0.64 , 16.9 , −2.22] ,
[0.65 , 17.5 , −1.54] , [0.66 , 17.9 , −.806] , [0.67 , 18.3 , 0.028] ,
[0.68 , 18.5 , 0.954] , [0.69 , 18.6 , 1.95] , [0.70 , 18.3 , 2.94] ,
[0.71 , 17.9 , 3.82] , [0.72 , 17.2 , 4.49] , [0.73 , 16.4 , 4.93] ,
[0.74 , 15.5 , 5.09] , [0.75 , 14.7 , 5.] , [0.76 , 14.0 , 4.73] ,
[0.77 , 13.4 , 4.35] , [0.78 , 13.0 , 3.91] , [0.79 , 12.6 , 3.46] ,
[0.80 , 12.3 , 2.94] , [0.81 , 11.9 , 2.31] , [0.82 , 11.6 , 1.53] ,

$[\ 0.83\ ,\ 11.5\ ,\ 0.600\ ]\ ,\ [\ 0.84\ ,\ 11.6\ ,\ -.454\ ]\ ,\ [\ 0.85\ ,\ 11.9\ ,\ -1.54\ ]\ ,$
$[\ 0.86\ ,\ 12.6\ ,\ -2.58\ ]\ ,\ [\ 0.87\ ,\ 13.4\ ,\ -3.47\ ]\ ,\ [\ 0.88\ ,\ 14.5\ ,\ -4.14\ ]\ ,$
$[\ 0.89\ ,\ 15.6\ ,\ -4.56\ ]\ ,\ [\ 0.90\ ,\ 16.8\ ,\ -4.76\ ]\ ,\ [\ 0.91\ ,\ 17.8\ ,\ -4.78\ ]\ ,$
$[\ 0.92\ ,\ 18.8\ ,\ -4.70\ ]\ ,\ [\ 0.93\ ,\ 19.8\ ,\ -4.55\ ]\ ,\ [\ 0.94\ ,\ 20.7\ ,\ -4.34\ ]\ ,$
$[\ 0.95\ ,\ 21.6\ ,\ -4.04\ ]\ ,\ [\ 0.96\ ,\ 22.6\ ,\ -3.60\ ]\ ,\ [\ 0.97\ ,\ 23.5\ ,\ -2.97\ ]\ ,$
$[\ 0.98\ ,\ 24.3\ ,\ -2.13\ ]\ ,\ [\ 0.99\ ,\ 25.0\ ,\ -1.11\ ]\ ,\ [\ 1.00\ ,\ 25.3\ ,\ 0.\ ]\ ]:$

$for\ i\ from\ 1\ to\ NumOfPoints\ do\ Q[i] := [P[i,2], P[i,3]]\ end\ do:$

$for\ i\ from\ 1\ to\ NumOfPoints\ do\ T[i] := P[i,1]\ end\ do:$

$for\ i\ from\ 1\ to\ NumOfPoints-1\ do\ deltaT[i] := P[i+1,1]-P[i,1]\ end\ do:$

$linint\ :=\ t->\ evalm(Q[1] * charfcn[T[1]](t) + add(\ (Q[k] + (t - T[k]) * (Q[k+1]-Q[k])/deltaT[k]) * (\ charfcn[T[k]..T[k+1]](t) - charfcn[T[k]](t)\ ),\ k = 1..NumOfPoints-1));$

$plot(\ [\ linint(\ t)\ [\ 1\ ]\ ,\ linint(\ t)\ [\ 2\ ]\ , t = 0..1\ ]\ )\ ;$

Note 13, From Page 62:

Type the following on a new command line on the same Maple worksheet as the above. We subtract off the apparent linear motion.

$plot(\ [\ linint(\ t)\ [\ 1\ ]\ -\ 20 * t\ ,\ linint(\ t)\ [\ 2\ ]\ , t = 0..1\ ]\ )\ ;$

Note 14, From Page 62:

Type the following on a new command line on the same Maple worksheet as the above. We now subtract off the apparent linear motion and a big circular motion too.

$plot(\ [\ linint(\ t)\ [\ 1\ ]\ -\ 20 * t\ -\ 5 * cos(6 * Pi * t)\ ,\ linint(\ t)\ [\ 2\ ]\ -\ 5 * sin(6 * Pi * t)\ , t = 0..1\ ]\ )\ ;$

Note 15, From Page 63:

Type the following on a new command line on the same Maple worksheet as the above. We now subtract off the apparent linear motion and both circular motions.

$plot(\ [\ linint(\ t)\ [\ 1\ ]\ -\ 20 * t\ -\ 5 * cos(6 * Pi * t)\ -\ 0.3 * cos(20 * Pi * t)\ ,\ linint(\ t)\ [\ 2\ ]\ -\ 5 * sin(6 * Pi * t)\ -\ 0.3 * sin(20 * Pi * t)\ , t = 0..1\ ]\ )\ ;$

Note 16, From Page 66:

The text presumes exposure to first year Calculus in this chapter.

Specifically, to read Section 16 completely you should know the meaning of the following words from Differential Calculus in one variable: limits, continuity, differentiability, one sided limits and one-sided continuity. You should know that sums, products and (where defined) compositions of continuous functions are continuous. You should know that sums, products and (where defined) compositions of differentiable functions are differentiable and how to calculate derivatives of compositions by the chain rule. You should understand the Mean Value Theorem. You should know how to calculate derivatives for polynomials, exponentials, logs, trig functions

and combinations of these and have worked through a large number of the usual Max-Min and applied problems.

To read Section 18 you should know about definite and indefinite integrals of continuous functions (including those mentioned above) and improper integrals defined as limits of definite integrals. You should understand how basic initial value problems are related to integrals. You should understand the techniques of integration by substitution and by parts. You should have plenty of experience going from Riemann sums to the associated integrals in applied problems.

Starting at Section 20 you will need some experience with sequences and series to follow the exercises and endnotes, although the text itself can be read without it for a ways farther. An outline of the basic facts concerning sequences and facts about derivatives and integrals too are presented here and there in the endnotes, but the discussion is likely too brief if you have never seen it before. Sequences are usually studied late in the first year of Calculus.

Note 17, From Page 66:

Depending on the choices of your previous math instructors you might have had some serious work with (and serviceable definition for) limits and continuity, or not. Proving statements in Calculus requires an understanding of these things, and that understanding usually takes several exposures to mature. If you have struggled with limits before that is all to the good.

In any case, here is the definition of **limit** for a vector function Q defined (at least) everywhere in some interval centered at c except, possibly, for c itself. We assume L is a vector of the same type as $Q(t)$.

$\lim_{t\to c} Q(t) = L$ exactly when for each $\varepsilon > 0$ there exists some $\delta > 0$ so that $|Q(t) - L| < \varepsilon$ whenever $0 < |c - t| < \delta$.

We say that the limit is L when the above condition can be shown to hold: that Q **converges** and, specifically, Q **converges to** L. What that means is that whenever you are required to put $Q(t)$ in a ball or circle of radius $\varepsilon > 0$ around L, you can accomplish this task by finding a $\delta > 0$ and picking t to be any number within a distance δ of c (c itself excluded.) And you must be able to accomplish this for ANY $\varepsilon > 0$. The definition doesn't explain how this might be done, nor does it tell you how to find L. It is merely an outline of what must be accomplished to show that a specific L is, in fact, the limit.

Suppose that the domain of Q is an interval (a, b) and c is in this interval. Recall that Q is continuous at c if the limit exists at c and is $Q(c)$, and Q is called continuous on a subinterval of (a, b) if it is continuous at each point in the subinterval.

There is an important fact about functions which are continuous on an interval which contains both endpoints. They have a property called **uniform continuity**:

Suppose Q is defined and continuous on the interval $[r, s]$. Q is called uniformly continuous on $[r, s]$ if for each $\varepsilon > 0$ there exists some $\delta > 0$ so that $|Q(t)-Q(u)| < \varepsilon$ whenever $0 < |t - u| < \delta$ and both t and u are in $[r, s]$.

In the definition of continuity on $[r, s]$, we merely require that such a δ exist for each ε at each point in $[r, s]$, but the δ for a given ε could vary from place to place. The statement above says that δ can be picked so that it "works" for a given ε everywhere on $[r, s]$. This is a pretty deep property of continuity and the real numbers.

That result is usually proved in the same class where they show the following two facts which are used in one form or another in many places in Calculus classes, but which are surprisingly hard to show. To prove them, and the fact about uniform continuity, you must go back to basics and do a better job of defining the real numbers, and that can take you pretty far afield!

The Intermediate Value Theorem
If a real function f is continuous on $[r, s]$ then for any real number Y between $f(r)$ and $f(s)$ there is some c in $[r, s]$ for which $f(c) = Y$. The theorem says that no values are "skipped" when you go from one height to another along the graph of a continuous function.

The Extreme Value Theorem
If a real function f is continuous on $[r, s]$ then there are numbers c and d in the interval for which $f(c) \leq f(x) \leq f(d)$ for every x in the interval. The theorem states that a continuous function on a closed interval actually attains both a maximum and a minimum value at numbers in that interval.

A real valued **sequence** is a function whose domain is the positive integers and whose range is a set of real numbers. If A is a sequence the value of A at domain member n is by custom denoted A_n.

A is said to be **monotone** if either $A_n \geq A_m$ for every pair of positive integers n and m with $n \geq m$, or if $A_n \leq A_m$ for every pair of positive integers n and m with $n \geq m$. The first case is called monotone increasing, while the second is called monotone decreasing.

A sequence is called **bounded** if there are real numbers r and s with $r \leq A_n \leq s$ for every positive integer n.

A sequence is said to **converge** if there is a number L such that for each $\varepsilon > 0$ an integer N can be found for which $|A_n - L| < \varepsilon$ for every integer $n \geq N$. The number L is called the limit of the sequence and denoted $lim_{n\to\infty} A_n$ when it exists.

It is a fact that a real valued function f defined on an interval (a, b) is continuous at c in (a, b) precisely when $\lim_{n\to\infty} f(A_n)$ exists and is $f(c)$ for **every** sequence A with range in (a, b) and which converges to c. This is an extremely useful observation.

A sequence can have at most one limit.

The Monotone Convergence Theorem for real valued sequences states that a bounded monotone sequence will converge.

The part that causes most of the work (in these more advanced classes where they discuss this in detail) lies not in showing that the the sequence values of a bounded monotone sequence are piling up on some place, but that there is actually a real number at that place to which they can converge. There are "no holes" in the real numbers.

The following definitions are related to this "no holes" fact:

Suppose A is any set of real numbers. A number U is called an **upper bound for** A if $x \leq U$ for every x in A. A number L is called a **lower bound for** A if $L \leq x$ for every x in A. It is a fact that every set with a lower bound has a **greatest lower bound**: a lower bound larger than all others. It is a fact that every set with an upper bound has a **least upper bound**: an upper bound smaller than all others.

A **subsequence** of a sequence A is another sequence B with the property that for each positive integer k there is a positive integer n_k with $B_k = A_{n_k}$ and **where $\mathbf{n_{k+1}}$ is always bigger than $\mathbf{n_k}$**.

Every subsequence of a convergent sequence converges, and to the same limit.

We will mention one final result about sequences.

If A is a sequence whose values lie in an interval $[\alpha, \beta]$ then there is a subsequence B of A which converges to a number in $[\alpha, \beta]$. The assumption that this interval contain its endpoints is important.

A **vector valued sequence** is a function whose domain is the positive integers and whose range is a set of vectors. The notation for vector valued sequences is similar to that for real valued sequence. A vector valued sequence is said to converge if the sequences at all coordinates converge. The vector formed from the limits of these coordinate sequences is called the limit of the vector valued sequence.

Note 18, From Page 68:

The Mean Value Theorem

Suppose f is a real valued function differentiable on (a, b) and continuous on $[a, b]$. Then there is a c in (a, b) for which $f'(c) = \frac{f(b)-f(a)}{b-a}$.

We break the proof into cases. The first case is if $f(a) = f(b)$. If f is constant on any subinterval of (a, b) then we can pick c to be any point in the subinterval and then $f'(c) = 0$. So we will further presume that f is not constant on any subinterval of (a, b). Since f is continuous on $[a, b]$ it attains both its maximum value M and its minimum value m at points in $[a, b]$. Since f is not constant one of these must be unequal to $f(a)$ and occur in the interval (a, b). We will suppose c is in (a, b) and $f(c) = M \neq f(a)$. The case where $f(c) = m \neq f(a)$ is left as an exercise.

Let Y be a sequence of numbers in $(f(a), M)$ with $\lim_{n\to\infty} Y_n = M$.

Define $\quad A_n = \{\, x \in (a, c) \mid f(x) = Y_n \,\}$ $\quad$ and $\quad B_n = \{\, x \in (c, b) \mid f(x) = Y_n \,\}$.

These sets are nonempty because f must pass through every value between $f(a)$ and M somewhere on both (a, c) and (c, b).

Let a_n be the least upper bound of A_n and b_n be the greatest lower bound of B_n. Continuity of f requires that $f(a_n) = f(b_n) = Y_n$. Also c is in (a_n, b_n) and all function values $f(x)$ for any x in (a_n, b_n) exceed Y_n. (Prove this!) So

$\lim_{n\to\infty} a_n = c$ and $\lim_{n\to\infty} b_n = c$.

$$\begin{aligned} 0 =& f(b_i) - f(a_i) = f(b_i) - f(c) + f(c) - f(a_i) \\ =& \frac{b_i - c}{b_i - a_i}\left(\frac{f(b_i) - f(c)}{b_i - c}\right) + \frac{c - a_i}{b_i - a_i}\left(\frac{f(c) - f(a_i)}{c - a_i}\right). \end{aligned}$$

Both ratios involving f converge to $f'(c)$ and the coefficient ratios add to 1 for each i so the sum converges to $f'(c)$ which is, therefore, 0.

The second case, where $f(b) \neq f(a)$, is an application of the first case to the function $g(x) = f(x) - Kx$ for the constant $K = \frac{f(b)-f(a)}{b-a}$.

The Cauchy Mean Value Theorem

This is a variant of the Mean Value Theorem with numerous applications. An example is found in an exercise below. We suppose that X and Y are two functions defined on an interval containing the interval (a, b). We suppose both X and Y are continuous on $[a, b]$ and differentiable on (a, b). Finally, we suppose that X' is never 0 on (a, b). We conclude that there is some c in (a, b) for which

$$\frac{Y'(c)}{X'(c)} = \frac{Y(b) - Y(a)}{X(b) - X(a)}.$$

Note first that if $X(b) - X(a) = 0$ then the mean value theorem would imply that $X'(t) = 0$ somewhere in (a, b) which, by assumption, cannot happen. So at least the two fractions above exist.

We define a function $K(t) = Y(t) - Y(a) - (X(t) - X(a))\frac{Y(b)-Y(a)}{X(b)-X(a)}$.

$K(b) = K(a) = 0$ and K satisfies the conditions of the Mean Value Theorem on the interval so there is a c in (a, b) with $K'(c) = 0$. The result follows.

---

*Exercise I. * If X and Y are differentiable on some interval $(a - \varepsilon, a + \varepsilon)$ around a and X' is never 0 on this interval and if Y' and X' are continuous at a then*

$$\lim_{h\to 0} \frac{Y(a + h) - Y(a)}{X(a + h) - X(a)} = \frac{Y'(a)}{X'(a)}.$$

---

*Exercise II. * If X and Y are differentiable on some interval $(a - \varepsilon, a + \varepsilon)$ around a and X' is never 0 on this interval except possibly at a then*

$$\lim_{t\to a} \frac{Y(t) - Y(a)}{X(t) - X(a)} = \lim_{t\to a} \frac{Y'(t)}{X'(t)}.$$

provided the second limit exists.

When $Y(a) = X(a) = 0$ this is one form of ***L'Hôpital's Rule.***

The Intermediate Value Theorem for Derivatives

Suppose f is differentiable on an interval containing $[a, b]$ and $f'(a) \neq f'(b)$ and Y is any number between $f'(a)$ and $f'(b)$. Then there is a c in (a, b) for which $f'(c) = Y$.

Note that we are not assuming that f' is continuous: only that it exists on the interval. This result implies that whatever discontinuities f' might have, they are not the kind that cause f' to jump past any values.

We suppose Y is any number between $f'(a)$ and $f'(b)$. By the Mean Value Theorem there is t in (a, b) with $f'(t) = \frac{f(b)-f(a)}{b-a}$. If $f'(t)$ happens to equal Y we are done. At least one, if not both, of the following statements are true: (i) Y is between $f'(t)$ and $f'(a)$ or (ii) Y is between $f'(t)$ and $f'(b)$.

Let us suppose that (ii) is true, and leave the other case as an exercise.

Define $H(x) = \frac{f(b)-f(x)}{b-x}$. H is continuous on $[a, b)$ and $\lim_{x\to b^-} H(x) = f'(b)$ so there is a number $s > t$ in (a, b) so that Y is between $H(a) = f'(t)$ and $H(s)$. Since H is continuous on (t, s) there is a number u in that interval with $H(u) = \frac{f(b)-f(u)}{b-u} = Y$. The Mean Value Theorem applies to this last fraction yielding c between t and u with $f'(c) = Y$.

Note 19, From Page 70:

Suppose a real valued function f is differentiable at c and $f'(c) = K > 0$. Since $\lim_{h\to 0} \frac{f(c+h)-f(c)}{h} = K$ there is a positive ε for which $\frac{K}{2} < \frac{f(c+h)-f(c)}{h}$ for any h with $-\varepsilon < h < \varepsilon$. When $0 < h < \varepsilon$ this gives $f(c + h) > f(c) + h\frac{K}{2}$. When $-\varepsilon < h < 0$ this gives $f(c+h) < f(c) + h\frac{K}{2}$. In words, $f(c)$ is strictly smaller than nearby function values to the right, and strictly bigger than nearby function values to the left.

A similar result holds if f' is always negative on (a, b).

The Mean Value Theorem also implies this, but without the interesting $h\frac{K}{2}$ term, which lets us think about how fast f must grow near c.

Note 20, From Page 70:

---

*Exercise III. * If X is differentiable on some interval containing $[a - \varepsilon, a + \varepsilon]$ around a and X' is never 0 on this interval the Intermediate Value Theorem for Derivatives implies that X' must have constant sign on this interval. So X is one-to-one on this interval. Since it is continuous, it cannot skip any values between $A = X(a-\varepsilon)$ and $B = X(a+\varepsilon)$. Each t in $[a-\varepsilon, a+\varepsilon]$ is associated with a unique X in $[A, B]$: that is, X has an inverse function. This inverse, which we denote t, is defined as follows: for u in $[A, B]$ let $t(u)$ be the unique member of $[a - \varepsilon, a + \varepsilon]$ for which $X(t(u)) = u$.*

Suppose X' is positive so $A < C = X(a) < B$. The case of X' negative is left as an exercise.

If Y is another function differentiable on an interval containing $[a - \varepsilon, a + \varepsilon]$ let W denote the function $W(u) = Y(t(u))$.

In Exercise I we saw that $\lim_{h\to 0} \frac{Y(a+h)-Y(a)}{X(a+h)-X(a)} = \frac{Y'(a)}{X'(a)}$ and that can now be rephrased as

$$W'(C) = \lim_{\Delta X\to 0} \frac{W(C + \Delta X) - W(C)}{\Delta X} = \frac{Y'(a)}{X'(a)}.$$

You sometimes see this written as

$$\frac{dY}{dX} = \frac{dY/dt}{dX/dt}.$$

Note 21, From Page 71:

This follows from the usual discussion of integration and Riemann sums which we apply to the coordinate functions and recall to mind here.

Suppose X is a continuous real valued function on the interval $[r, s]$. We let $r = t_0 < t_1 < \cdots < t_N = s$ and for each n between 1 and N we let c_n denote a selection of a point in the interval $[t_{n-1}, t_n]$. Let $\Delta t_n = t_n - t_{n-1}$ for each n between 1 and N. The selection of the t_n values is called a **partition** of $[r, s]$. The **mesh** of this partition is the largest of the N numbers Δt_n.

We will be interested in thinking about more than one partition and more than one possible choice of the c_n values so we will give them names. If we say T is a partition we mean a subscripted selection of members t_n of $[r, s]$ as above. If we say C is **subordinate** to T we mean that C is a subscripted selection of members c_n of $[t_{n-1}, t_n]$ for each n as above. If U is another partition, we say that U is a **refinement** of T if T is a subset of U. Any two partitions have what is referred to as a **common refinement**: form the union of the two partitions and label the members of the union in order.

The **Riemann sum** formed from partition T and subordinate C is the number

$$Riemann(T, C) = \sum_{n=1}^{N} X(c_n)\Delta t_n$$

.

X attains both a maximum and minimum value on each interval $[t_{n-1}, t_n]$ formed from a partition T. We denote these numbers $M_{T,n}$ and $m_{T,n}$ respectively. The **Upper and Lower Riemann sums** for a partition T are defined to be, respectively:

$$Upper(T) = \sum_{n=1}^{N} M_{T,n}\Delta t_n \quad \text{and} \quad Lower(T) = \sum_{n=1}^{N} m_{T,n}\Delta t_n.$$

These definitions were set up carefully to make the following facts easy to show.

(i) If T is any partition and C is subordinate to T then

$$Lower(T) \leq Riemann(T, C) \leq Upper(T).$$

(ii) If U is a refinement of T then

$$Lower(T) \leq Lower(U) \leq Upper(U) \leq Upper(T).$$

We now use the property of uniform continuity of the function X as discussed in the note above. In our context it states that for each $\varepsilon > 0$ we can find a $\delta > 0$ so that if the mesh of a partition T is less than δ then $M_{T,n} - m_{T,n} < \frac{\varepsilon}{s-r}$ for every n. The variation between the maximum and the minimum values of X on every $[t_{n-1}, t_n]$ cannot exceed $\frac{\varepsilon}{s-r}$.

(iii) Show that if the mesh of T is less than δ as in the paragraph above and C is subordinate to T then

$$\begin{aligned} Upper(T) - Lower(T) &< \varepsilon \\ \text{and so} \quad Upper(T) - Riemann(T,C) &< \varepsilon \\ \text{and} \quad Riemann(T,C) - Lower(T) &< \varepsilon. \end{aligned}$$

With these facts in mind we now turn to a specific collection of partitions obtained by repeatedly cutting $[r,s]$ in half.

For each integer $n \geq 1$ we let D^n be the partition containing the numbers $\frac{k}{2^{n-1}}(s-r)$ for $k = 0 \ldots 2^{n-1}$. So D^1 is the crudest possible partition, containing only the endpoints of the interval $[r,s]$ while D^2 cuts this interval into two equal pieces and in general D^{n+1} cuts each interval from D^n into two equally sized subintervals. Each D^{n+1} is a refinement of D^n. The mesh of D^n is $\frac{1}{2^{n-1}}(s-r)$.

(iv) The sequence $Upper(D^n)$ is monotone and bounded. So is the sequence $Lower(D^n)$. Therefore they both converge.

(v) If $\varepsilon > 0$ we can find δ as above so that whenever n satisfies $\frac{1}{2^{n-1}}(s-r) < \delta$ then $Upper(D^n) - Lower(D^n) < \varepsilon$. Since ε can be chosen to be arbitrarily small, this means that the sequences $Upper(D^n)$ and $Lower(D^n)$ converge to the same number, which we denote $\int_r^s X(t)dt$.

We now deal with the potential problem with other partitions. We want $Riemann(T,C)$ to be near to the number $\int_r^s X(t)dt$ no matter what the partition is and no matter the choice of C provided only that the mesh of T is small.

Specifically, for any $\varepsilon > 0$ we want to be able to guarantee that

$$\left| Riemann(T,C) - \int_r^s X(t)dt \right| < \varepsilon$$

requiring only that the mesh of T be small enough.

Select δ to be a positive number so small that whenever u and v are in $[r,s]$ and $|u-v| < \delta$ then $|X(u) - X(v)| < \frac{\varepsilon}{4(s-r)}$. Select n to be an integer with $\frac{1}{2^{n-1}}(s-r) < \delta$. Suppose T is any partition with mesh not exceeding δ and C is subordinate to T. Let U be the common refinement of T and D^n.

(vi) It follows that:

$$\begin{aligned} & |Riemann(T,C) - Upper(T)| < \frac{\varepsilon}{4} \\ \text{and} \quad & |Upper(T) - Upper(U)| < \frac{\varepsilon}{4} \\ \text{and} \quad & |Upper(U) - Upper(D^n)| < \frac{\varepsilon}{4} \\ \text{and} \quad & |Upper(D^n) - \int_r^s X(t)dt| < \frac{\varepsilon}{4}. \end{aligned}$$

(vii) The result we wanted now follows by application of the triangle inequality:

$$\left| Riemann(T,C) - \int_r^s X(t)dt \right|$$

cannot exceed the sum of the four left hand sides above.

Note 22, From Page 81:

Here is Bezunit.

$restart\ :\ with(plots)\ :$
$A\ :=\ [\ 2,\ 7\ ]\ ;\ B\ :=\ [\ 3,6\ ]\ ;VA\ :=\ [\ 6,-3\ ]\ ;\ VB\ :=\ [\ -1,6\ ]\ ;$

$bezcurveunit\ :=\ t\ -\ >\ (\ 1-t\ )\ \hat{}\ 3*A\ +$
$3*(t)*(1-t)\ \hat{}\ 2*(A+VA/3)\ +3*t\hat{}\ 2*(1-t)*(B-VB/3)+t\hat{}\ 3*B;$

$plot(\ [\ bezcurveunit(\ t\ )\ [\ 1\ ]\ ,\ bezcurveunit(\ t\ )\ [\ 2\ ]\ ,\ t=0..1\ ]\ )\ ;$

Note 23, From Page 82:

Here is Bezspeed

$A\ :=\ [\ 2,\ 7\ ]\ ;\ B\ :=\ [\ 3\ ,\ 6\ ]\ ;\ VA\ :=\ [\ 6,\ -3\ ]\ ;\ VB\ :=\ [\ -1,\ 6\ ]\ ;\ TA\ :=\ 6\ ;\ TB\ :=\ 9\ ;$

$bezcurvespeedchange\ :=\ t\ -\ >\ (\ (\ TB-t\ )\hat{}3*A\ +$
$3*(\ t-TA\ )*(\ TB-t\ )\hat{}2*(\ A+(\ TB-TA\ )*VA/3\ )+3*(\ t-TA\ )\hat{}2*$
$(\ TB-t\ )*(\ B-(\ TB-TA\ )*VB/3\ )+$
$(\ t-TA\ )\hat{}3*B\ )/(\ TB-TA\ )\hat{}3\ ;$

$plot(\ [\ bezcurvespeedchange(t)\ [1]\ ,\ bezcurvespeedchange(t)[\ 2]\ ,\ t=6..9\ ]\ )\ ;$

Note 24, From Page 83:

Here is Bezpatch

$A\ :=\ [\ 2,\ 7\ ]\ ;\ B\ :=\ [\ 3,\ 6\ ]\ ;\ C\ :=\ [\ 5,\ 2\ ]\ ;\ VA\ :=\ [\ 6,\ -3\ ]\ ;\ VB\ :=\ [\ -1,\ 6\ ]\ ;\ VC\ :=\ [\ 8,\ -1\ ]\ ;\ TA\ :=\ 6\ ;\ TB\ :=\ 9\ ;\ TC\ :=\ 11\ ;$

$bezpatch\ :=\ t\ -\ >$
$(\ (\ TB-t\ )\hat{}3*A\ +\ 3*(\ t-TA\ )*(\ TB-t\ )\hat{}2*(\ A+(\ TB-TA\ )*VA/3\ )+$
$3*(\ t-TA\ )\hat{}2*(\ TB-t\ )*(\ B-(\ TB-TA\ )*VB/3\ )+$
$(\ t-TA\ )\hat{}3*B\ )*charfcn[\ TA..TB\ ](t)/(\ TB-TA\ )\hat{}3+$
$(\ (\ TC-t\ )\hat{}3*B\ +\ 3*(\ t-TB\ )*(\ TC-t\ )\hat{}2*(\ B+(\ TC-TB\ )*VB/3\ )+$
$3*(\ t-TB\ )\hat{}2*(\ TC-t\ )*(\ C-(\ TC-TB\ )*VC/3\ )+(\ t-TB\ )\hat{}3*C\ )*$
$(\ charfcn[\ TB..TC\ ](t)-charfcn[\ TB\ ]\ (t)\ )/(\ TC-TB\ )\hat{}3\ ;$

$plot(\ [\ bezpatch(t)\ [\ 1\ ]\ ,\ bezpatch(t)[\ 2]\ ,t=6..11]\ );$

Note 25, From Page 83:

Type this on the same Maple worksheet as the function which linearly interpolates the 101 data points above. That way this function will have access to the data and vectors already defined. This function uses Bezpatch to interpolate smoothly between the data points. The process is called **Bezier interpolation.**

$for\ i\ from\ 2\ to\ NumOfPoints-1$
$do\ V[\ i\ ]:=(\ Q[\ i+1\ ]\ -Q[\ i-1\ ]\ )/(\ T[\ i+1]\ -T[\ i-1\ ]\ )\ end\ do:$
$V[\ 1\ ]\ :=\ (Q[\ 2\ ]\ -Q[\ 1]\ )/deltaT[\ 1]\ :$

```
V[ NumOfPoints ] :=
(Q[ NumOfPoints ] - Q[ NumOfPoints - 1 ] )/deltaT[ NumOfPoints - 1 ] :

    bezint := t -> 
(   (T[ 2] - t )^3 * Q[ 1 ] +
3 * (t - T[1]) * (T[2] - t )^2 * ( Q[1] + deltaT[1]/3 * V[1])+
3 * (t - T[1])^2 * (T[2] - t) * (Q[2] - deltaT[1]/3 * V[2])+
(t - T[1])^3 * Q[2]   ) * charfcn[ T[1]..T[2] ](t)/deltaT[1]^3+
add(   ( (T[k+1] - t )^3 * Q[k] +
3 * (t - T[k]) * (T[k+1] - t )^2 * (Q[k] + deltaT[k]/3 * V[k]) +
3 * (t - T[k])^2 * (T[k+1] - t ) * (Q[k+1] - deltaT[k]/3 * V[k+1]) +
(t - T[k])^3 * Q[k+1]   )*
( charfcn[ T[k]..T[k+1] ](t) - charfcn[ T[k] ](t) )/deltaT[k]^(3)
, k = 2..NumOfPoints - 2   ) +
( (T[NumOfPoints] - t )^3 * Q[NumOfPoints - 1] +
3 * (t - T[NumOfPoints - 1]) * (T[NumOfPoints] - t )^2*
( Q[NumOfPoints-1] + deltaT[NumOfPoints-1]/3*V[NumOfPoints-1] )+
3 * (t - T[NumOfPoints - 1])^2 * (T[NumOfPoints] - t )*
(Q[NumOfPoints] - deltaT[NumOfPoints - 1]/3 * V[NumOfPoints] ) +
(t - T[NumOfPoints - 1])^3 * Q[NumOfPoints] )*
( charfcn[ T[ NumOfPoints - 1 ]..T[ NumOfPoints ] ](t)-
charfcn[ T[ NumOfPoints - 1 ] ](t) )/
deltaT[ NumOfPoints - 1 ]^3


    plot([bezint(t)[1] - 20 * t - 5 * cos(6 * Pi * t) - 0.3 * cos(20 * Pi * t) ,
bezint(t)[2] - 5 * sin(6 * Pi * t) - 0.3 * sin(20 * Pi * t) , t = 0.01..0.99 ] );
```

Note 26, From Page 85:

Suppose Q is a continuous parameterization of a curve as in the text and T is the partition $t_0, \ldots, t_N$ of the interval $[c, d]$ in the domain of Q with $c = t_0 < t_1 < \cdots < t_N = d$. We will always enumerate a partition such as T in the standard way, with bigger subscripts corresponding to bigger times. Define $Arc_T = \sum_{i=1}^{N} |Q(t_i) - Q(t_{i-1})|$.

If we pick one of the terms $|Q(t_i) - Q(t_{i-1})|$ in this sum and any s with $t_i > s > t_{i-1}$ the triangle inequality tells us that $|Q(t_i) - Q(t_{i-1})| \leq |Q(t_i) - Q(s)| + |Q(s) - Q(t_{i-1})|$.

From this we conclude that if U is any refinement of T then $Arc_U \geq Arc_T$: that is, if you take a polygonal path that touches the curve at segment endpoints and chop one or more partition subintervals in pieces the sum of the lengths along the new path cannot be smaller than before.

We recall that if S and T are any two partitions they possess a common refinement $S \cup T$.

So either the numbers we obtain as Arc_T can be chosen (for various T) to be large without bound or not. In the first case we might say the curve has infinite

length. In the second case we can find a sequence of partitions T_n for which the numbers Arc_{T_n} constitute a monotone and bounded sequence, which therefore converges to a number L, and we can choose this sequence so that $L \geq Arc_U$ for any partition U. We would then think of L as the **length of the curve:** the arclength. We can require that each T_{n+1} be a refinement of T_n and that the mesh of the partitions converges to 0.

In case Q is continuously differentiable on $[c, d]$ the first case is impossible, and this number L is the same as the arclength defined as the integral of the speed as defined later in the section. To see this we note that for any relevant r and s,

$$|Q(r) - Q(s)| \leq |X(r) - X(s)| + |Y(r) - Y(s)| + \dots$$

Since $|X'|$ is continuous on closed $[c, d]$ it attains a maximum value, and the same is true for $|Y'|$ and, if present, $|Z'|$. Let the number K stand for the largest magnitude of these derivatives on $[c, d]$. By the Mean Value Theorem, $|X(r)-X(s)| \leq K|r-s|$ and a similar inequality holds for Y' and, if present, Z'. So $|Q(r)-Q(s)| \leq 3K|r-s|$ for any r and s. So for any partition as above

$$\sum_{i=1}^{N} |Q(t_i) - Q(t_{i-1})| \leq \sum_{i=1}^{N} 3K|t_i - t_{i-1}| = 3K(d - c).$$

So the sums are bounded.

Note 27, From Page 86:

Suppose Q is a parameterization of a curve as in the text and $t_0, \dots, t_N$ is a partition of the interval $[c, d]$ in the domain of Q. We want to know that $\sum_{i=1}^{N} |\Delta Q_i|$ is close to $\sum_{i=1}^{N} |Q'(t_i)|\ \Delta t_i$ when the mesh of the partition is small.

Suppose $\varepsilon > 0$. Since the coordinate functions of $Q' = \langle X', Y' \dots \rangle$ are continuous on $[c, d]$ there is a $\delta > 0$ so that if $|u - v| < \delta$ then $|X'(u) - X'(v)| < \varepsilon$, and the same for Y' and (if there is a Z coordinate) Z'.

Suppose that the mesh of the partition is smaller than this δ. By the Mean Value Theorem, for each i there is s_i in $[t_{i-1}.t_i]$ so that $\frac{X(t_i)-X(t_{i-1})}{t_i-t_{i-1}} = X'(s_i)$. This means that $\left|\frac{X(t_i)-X(t_{i-1})}{t_i-t_{i-1}} - X'(t_i)\right| = |X'(s_i) - X'(t_i)| < \varepsilon$. A similar inequality holds for the Y coordinate (and Z if there is one.)

The result now follows from repeated applications of the triangle inequality:

$$
\begin{aligned}
\left| \sum_{i=1}^{N} |\Delta Q_i| - \sum_{i=1}^{N} |\ Q'(t_i)|\ \Delta t_i \right| &= \left| \sum_{i=1}^{N} |\Delta Q_i| - |\ Q'(t_i)|\ \Delta t_i \right| \\
&\leq \sum_{i=1}^{N} \left| |\Delta Q_i| - |\ Q'(t_i)|\ \Delta t_i \right| \leq \sum_{i=1}^{N} |\Delta Q_i - \ Q'(t_i)\ \Delta t_i\ | \\
&= \sum_{i=1}^{N} \left| \left(\left| \frac{\Delta Q_i}{\Delta t_i} - \ Q'(t_i) \right| \right)\ \Delta t_i \right| \\
&\leq \sum_{i=1}^{N} \left| \left(\left| \frac{\Delta X_i}{\Delta t_i} - \ X'(t_i) \right| \right)\ \Delta t_i \right| \ + \text{ similar terms with } Y \text{ and possibly } Z \\
&< \sum_{i=1}^{N} \varepsilon\ \Delta t_i \ \ + \text{ similar terms with } Y \text{ and possibly } Z \\
&= \ \varepsilon\ (d-c)\ + \text{ similar terms with } Y \text{ and possibly } Z \\
&\leq 3\varepsilon(d-c).
\end{aligned}
$$

This can be made as small as you wish by choosing ε small enough.

Note 28, From Page 91:

The problem is named after a student in one of my classes who suggested it.

Note 29, From Page 104:

A set of real numbers A is called **open** if for each point a in A there is a number $\varepsilon > 0$ so that the entire interval $(a-\varepsilon, a+\varepsilon)$ is in A. A set K of real numbers is called **closed** if its complement (that is the set of real numbers **not** in K) is open. A set B of real numbers is called **bounded** if it is contained in an interval of the form $[a, b]$ for real numbers a and b.

Note 30, From Page 114:

In elementary Calculus the second derivative test can be used to decide if a function has a local maximum or minimum at a critical point. Specifically, if f is a differentiable real valued functioned defined around c and $f'(c) = 0$ then f might have a local extreme value at c. The second derivative test says that if $f''(c)$ exists and $f''(c) > 0$ then f has a local minimum at c and if $f''(c) < 0$ then f has a local maximum at c. If $f''(c) = 0$ then the test is inconclusive.

A result with stronger conditions on f and more informative conclusions is contained in the following exercise.

---

*Exercise IV. * If f is a twice continuously differentiable real valued functioned defined on the interval $(-\varepsilon, \varepsilon)$ and $f'(0) = 0$ and $f''(t) \geq \alpha > 0$ for all t then $f(t) \geq \frac{1}{2}\alpha t^2 + f(0)$ for all t. In particular f has a local minimum at $t = 0$.*

If $f'(0) = 0$ and $f''(t) \leq -\alpha < 0$ for all t then $f(t) \leq -\frac{1}{2}\alpha t^2 + f(0)$ for all t. In particular f has a local maximum at $t = 0$.

A similar situation holds for surfaces defined as the graph of g on an open set $\mathfrak{O}$ in the plane. We will presume that g is twice continuously differentiable: that is g is continuously differentiable on $\mathfrak{O}$ as are the components of ∇g.

We will suppose that $\nabla g(a,b) = 0$ so g has a horizontal tangent plane at (a,b) so g, potentially, has a local extreme value at (a,b). We will derive a **second derivative test** to decide the issue in many cases.

Every straight line through $\langle a,b,0\rangle$ in the XY plane except the line parallel to the Y axis can be parameterized by $\left\langle \frac{t}{\sqrt{1+m^2}} + a, \frac{mt}{\sqrt{1+m^2}} + b, 0 \right\rangle$. This line has "slope" m in the XY plane and the square root factor has been placed there so that the parameterization is being traversed at unit speed: $|t|$ is always the distance from $\left(\frac{t}{\sqrt{1+m^2}} + a, \frac{mt}{\sqrt{1+m^2}} + b, 0\right)$ to $(a,b,0)$.

Up on the surface, the curve with this shadow is

$$T_m(t) = \left\langle \frac{t}{\sqrt{1+m^2}} + a, \frac{mt}{\sqrt{1+m^2}} + b,\ g\left(\frac{t}{\sqrt{1+m^2}} + a, \frac{mt}{\sqrt{1+m^2}} + b\right) \right\rangle.$$

We (temporarily) let $E = \frac{1}{\sqrt{1+m^2}}$ and, in the equations below, evaluate the derivatives of g at $(tE + a, mtE + b)$. The derivative of T_m is then

$$T_m'(t) = \langle E, mE,\ ED_1 g + mED_2 g\rangle$$

and the second derivative is

$$T_m''(t) = \left\langle 0, 0,\ E^2 D_{1,1} g + 2mE^2 D_{1,2} g + m^2 E^2 D_{2,2} g \right\rangle.$$

Evaluating this at $t = 0$ we have

$$T_m''(0) = \left\langle 0, 0,\ E^2 D_{1,1} g\,(a,b) + 2mE^2 D_{1,2} g\,(a,b) + m^2 E^2 D_{2,2} g\,(a,b) \right\rangle.$$

Letting $A = D_{1,1} g\,(a,b)$, $B = D_{1,2} g\,(a,b)$ and $C = D_{2,2} g\,(a,b)$ and replacing E this unpalatable mess at the third coordinate becomes

$$S(m) = \frac{1}{1+m^2}(A + 2mB + m^2 C).$$

$S(m)$ is the second derivative at $t = 0$ of the parameterized curve we formed by slicing through the surface at $(a,b,g(a,b))$ with the vertical plane through the line with "slope"m in the XY plane.

If this quantity can change sign for different m values then $(a,b,g(a,b))$ cannot be a local maximum or minimum on the surface, because (as in Exercise IV) one of the curves will be above

$$\left\langle \frac{t}{\sqrt{1+m_1^2}} + a, \frac{m_1 t}{\sqrt{1+m_1^2}} + b,\ g(a,b) + \alpha t^2 \right\rangle$$

for some $\alpha > 0$ and for a small t interval around 0 while for a different "slope" the curve will lie beneath

$$\left\langle \frac{t}{\sqrt{1+m_2^2}} + a, \frac{m_2 t}{\sqrt{1+m_2^2}} + b,\ g(a,b) - \alpha t^2 \right\rangle.$$

Places like this on g are called **saddles.** This will happen when the quadratic formula gives two roots for $S(m)$, which will happen when

$$B^2 - AC = \Big(D_{1,2}g(a,b)\Big)^2 - \Big(D_{1,1}g(a,b)\Big)\Big(D_{2,2}g(a,b)\Big) > 0.$$

When $B^2 - AC = 0$ the test is inconclusive.

When $B^2 - AC < 0$ we will show that there is a local extreme value at $(a, b, g(a, b))$. When $B^2 - AC < 0$ it must be that A and C are both nonzero and have the same sign. The extreme value will be a maximum if A (and hence C too) is negative, and a minimum if both are positive.

Exercise V. Create examples to show that if $B^2 - AC = 0$ the test is inconclusive.

---

*Exercise VI. ** We suppose $B^2 - AC < 0$ and $A < 0$ (so $C < 0$ too.)*

(i) $S(m) = \frac{1}{1+m^2}(A + 2mB + m^2C)$ must always be negative since $S(0) = A < 0$. Also, $\lim_{m\to\pm\infty} S(m) = C < 0$. Show that there is a positive number α so that $S(m) < -\alpha < 0$ for all m.

(ii) Show that the only line through $(a, b, 0)$ which we did not consider, the parameterized curve $(a, t + b, 0)$, generates a parameterized curve on the surface corresponding to an ordinary differentiable function whose second derivative is C, which cannot exceed $-\alpha$ either, at $t = 0$.

(iii) Show that $\left|\frac{1}{1+m^2}\right| < 1$, $\left|\frac{2m}{1+m^2}\right| < 2$ and $\left|\frac{m^2}{1+m^2}\right| < 1$ for all m. From this we can conclude that a variation of A by ΔA and B by ΔB and C by ΔC can cause $S(m)$ to change by no more than $|\Delta A| + 2|\Delta B| + |\Delta C|$.

(iv) Since the second partial derivatives are all continuous, we can choose a distance ε so small that $|\Delta A| + 2|\Delta B| + |\Delta C| < \frac{\alpha}{2}$ where $\Delta A = D_{1,1}g(X, Y) - D_{1,1}g(a, b)$, $\Delta B = D_{1,2}g(X, Y) - D_{1,2}g(a, b)$ and $\Delta C = D_{2,2}g(X, Y) - D_{2,2}g(a, b)$ for any (X, Y) in a disk of radius ε around (a, b).

(v) In the parameterizations above $|t|$ was always the distance from the shadow of $(a, b, g(a, b))$ to the shadow of the point on the curve at t. So on every curve we discussed, the second derivative all along the parameterized curve is always negative and in fact never exceeds $-\frac{\alpha}{2}$ for any t in $(-\varepsilon, \varepsilon)$. The important point here is that the interval $(-\varepsilon, \varepsilon)$ does not vary from curve to curve: it is the same interval for all of them.

(vi) Use this to conclude that $g(X, Y) \le g(a, b) - \frac{\alpha}{4}\Big((X - a)^2 + (Y - b)^2\Big)$ whenever (X, Y) is in a disk of radius ε around (a, b). So $g(a, b)$ is a local maximum value.

(vii) Replace g by $-g$ to conclude that if $B^2 - AC < 0$ and A or C is positive then we have a local minimum.

Note 31, From Page 115:

Taylor Polynomials

We will suppose that f is a real valued function which is n times differentiable at a point a in its domain. Let

$$P_n(t) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(t-a)^k.$$

This is the n–th **Taylor Polynomial for f at a**. The values of f and P_n as well as the first n derivatives of these two functions agree at a. We would like to conclude that f and P_n are close to each other away from a but this need not be true in any useful way. For example define a function

$$f(t) = \begin{cases} 0, & \text{if } t = 0; \\ \mathrm{e}^{-\frac{1}{t^2}} & \text{if } t \neq 0. \end{cases}$$

This function is **very** flat at 0. All of its derivatives are 0 there, so the n–th Taylor Polynomial at 0 for any n is the zero polynomial, not a good approximation to f away from 0.

We now make additional assumptions which will allow us to conclude when and if P_n is close to f. We suppose that f has n continuous derivatives on an interval containing $[a,t]$ and that $f^{(n+1)}$ exists on (a,t).

$$\text{Define} \quad H(x) = f(x) - P_n(x) - \frac{f(t)-P_n(t)}{(t-a)^{n+1}}(x-a)^{n+1}.$$

H has been defined so that $H(a) = H(t) = 0$ and H has n continuous derivatives (with respect to x) on an interval containing $[a,t]$ and $H^{(k)}(a) = 0$ for $k = 0, \ldots, n$ and is $n+1$ times differentiable on (a,t).

The mean value theorem then implies that there is a c_1 in (a,t) for which $H'(c_1) = 0$. So there is a c_2 in (a,c_1) for which $H''(c_2) = 0$. This process can continue, yielding in the end c_{n+1} in (a,c_n) for which $H^{(n+1)}(c_{n+1}) = 0$. Let $c = c_{n+1}$ and note that c is in (a,t).

If you calculate $H^{(n+1)}(c_{n+1})$ you will find that

$$0 = H^{(n+1)}(c_{n+1}) = f^{(n+1)}(c) - \frac{f(t)-P_n(t)}{(t-a)^{n+1}}(n+1)!$$

and this gives

$$f(t) = P_n(t) + \frac{f^{(n+1)}(c)}{(n+1)!}(t-a)^{n+1} \quad \text{for some } c \text{ in } (a,t).$$

The last term on the right is called the remainder term $R_n(t)$ and if it is small then $P_n(t)$ is a good approximation to $f(t)$.

Show that the same result holds if $t < a$: that is if all this takes place on an interval of the form $[t,a]$.

Note 32, From Page 118:

There are important properties about continuous functions on closed and bounded subsets of their domain which we mention here. They are analogous to similar facts we have already used in the case of continuous functions defined on intervals. We

suppose that f is a continuous function defined on an open subset $\mathfrak{O}$ of $\mathbb{R}^N$ where N is $1, 2$ or 3. We presume that $\mathfrak{K}$ is a closed and bounded subset of $\mathfrak{O}$.

(i) The set of real numbers $\{\, f(P) \mid P \in \mathfrak{K} \,\}$ is bounded. Even more, there are points P and Q in $\mathfrak{K}$ where f actually attains a maximum and minimum value on $\mathfrak{K}$. Specifically, there are points P and Q in $\mathfrak{K}$ for which $f(P) \le f(S) \le f(Q)$ for all $S \in \mathfrak{K}$.

(ii) f is uniformly continuous on $\mathfrak{K}$. We take that to mean: for each $\varepsilon > 0$ there is a $\delta > 0$ so that for any $P \in \mathfrak{K}$ if $|P-Q| < \delta$ then $Q \in \mathfrak{O}$ and $|f(Q)-f(P)| < \varepsilon$. The important point here is that δ can be chosen to be the same for all points in $\mathfrak{K}$.

Note 33, From Page 131:

Suppose in the vicinity of a point P in the intersection, the level set for g is the graph of $Z(X,Y)$. Recall that $D_i Z(X,Y) = \frac{-(D_i g)(X,Y,Z(X,Y))}{(D_3 g)(X,Y,Z(X,Y))}$ for $i = 1, 2$.

Consider the function $U = (U_1, U_2, U_3)$ defined by $U(X,Y) = (X, Y, Z(X,Y))$ and define $W(X,Y)$ to be $h \circ U(X,Y)$ for (X,Y) in an open set on the plane. If $h(P) = c$, the shadow of the intersection of the two surfaces in the vicinity of P onto the XY plane will be the level set $W(X,Y) = c$.

$$
\begin{aligned}
W'(X,Y) = & \\
&= h'(U(X,Y))U'(X,Y) \\
&= \begin{pmatrix} D_1 h & D_2 h & D_3 h \end{pmatrix} \begin{pmatrix} D_1U_1 & D_2U_1 \\ D_1U_2 & D_2U_2 \\ D_1U_3 & D_2U_3 \end{pmatrix} \\
&= \begin{pmatrix} D_1 h & D_2 h & D_3 h \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \frac{-D_1 g}{D_3 g} & \frac{-D_2 g}{D_3 g} \end{pmatrix} \\
&= \begin{pmatrix} D_1 h - \frac{D_1 g D_3 h}{D_3 g} & D_2 h - \frac{D_2 g D_3 h}{D_3 g} \end{pmatrix}.
\end{aligned}
$$

This can only be the zero matrix at points where

$$\begin{pmatrix} D_1 h & D_2 h \end{pmatrix} = \frac{D_3 h}{D_3 g} \begin{pmatrix} D_1 g & D_2 g \end{pmatrix}$$

which would imply that $\nabla h = \nabla g$. If this is not the case at P then one or both of the coordinates of $W'(P_1, P_2)$ is nonzero. If the second is nonzero, then for points around (P_1, P_2) on the level set $W(X,Y) = c$ we can write Y as a function of X.

The parameterization $H(X) = \langle X, Y(X) \rangle$ traces out the graph of Y, and

$$H'(X) = \left\langle 1, \frac{-D_1 W}{D_2 W} \right\rangle = \left\langle 1, \frac{D_1 g D_3 h - D_1 h D_3 g}{D_2 h D_3 g - D_2 g D_3 h} \right\rangle.$$

For X in an interval around P_1 the intersection we wanted is parameterized explicitly by $Q(X) = U \circ H(X) = (X, Y(X), Z(X, Y(X)))$. Note that

$$Q'(X) = U'(H(X))H'(X) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \frac{-D_1 g}{D_3 g} & \frac{-D_2 g}{D_3 g} \end{pmatrix} \begin{pmatrix} 1 \\ \frac{D_1 g D_3 h - D_1 h D_3 g}{D_2 h D_3 g - D_2 g D_3 h} \end{pmatrix}$$
$$= \frac{1}{D_2 g D_3 h - D_3 g D_2 h} \begin{pmatrix} D_2 g D_3 h - D_3 g D_2 h \\ D_3 g D_1 h - D_1 g D_3 h \\ D_1 g D_2 h - D_2 g D_1 h \end{pmatrix}.$$

This tangent vector is an explicit multiple of $\nabla g \times \nabla h$ as we speculated it must be on the intersection.

So we come the the following sufficient condition to guarantee that the intersection of two level sets of differentiable functions g and h in space will be a curve: First, at each point on the intersection one component of (the same component of) g' and h' should be nonzero (though that component could change from place to place on the intersection) and ∇h should never be a multiple of ∇g on the intersection.

Note 34, From Page 138:

After wrestling with a couple of different (lengthy) approaches to filling in the details in the construction and calculation of integrals in the plane and in space it seems to me that expanding the generality in our results might be better left until after a student has learned about the **Lebesgue Integral**. Students who really need to have more generality will need to know about the Lebesgue Integral anyway for other reasons. Those who won't go that far will probably be satisfied with my assurance that the plausible arguments and pictures presented in this text can be generalized extensively.

After studying the Lebesgue Integral you will learn that there is essentially only one way of creating integrals consistent with our intuition about ordinary area in the plane and ordinary volume in space, and that both the Riemann sum construction of multiple integrals and the Fubini Theorem iterated integral approach agree with this Lebesgue Integral where they are all defined, which includes at least the constant functions on rectangles. This, together with a "continuity" condition implies that they agree with the Lebesgue Integral (and hence with each other) whenever multiple and iterated integrals are both defined.

So it is OK to use iterated integrals (and all the techniques of basic Integral Calculus which apply to them) to come up with the numbers practical applications require in situations of far greater generality than we suppose in this text.

Other topics, such as the change of variable formulas, are really awkward to prove sans Linear Algebra and also require a bit more of the Topology you will learn about in an Advanced Calculus course than I want to deal with here.

Note 35, From Page 139:

Consider the sum $M_P = \sum_{\mathfrak{O}} \Delta X_i \Delta Y_j$ where this notation indicates that the sum is over those subscripts corresponding to rectangles in P which are entirely inside $\mathfrak{O}$. This is an approximation to the area of $\mathfrak{O}$.

Suppose P and Q are partitions of our rectangle. We will suppose n and k are positive integers and the mesh of P is less than $\frac{1}{n}$ while the mesh of Q is smaller: less than $\frac{1}{nk}$.

If N is an integer larger than both $b-a$ and $d-c$ the partition P generates no more than $2nN$ vertical or horizontal gridlines. Each gridline has length no more than N so can cut through the inside of no more than knN rectangles from Q. The total area of all these rectangles from Q for all the gridlines cannot exceed $(2nN)(knN)\frac{1}{k^2n^2} = \frac{2N^2}{k}$.

The only way any part of a rectangle from M_P could be left out of the area calculation M_Q is if this lost area corresponded to a rectangle from Q which extended past a P gridline and subsequently extended outside of $\mathfrak{O}$. We have calculated the maximum total area of all the rectangles from Q which could do that, or in fact which cross any gridline from P at all.

Therefore $M_P - M_Q < \frac{2N^2}{k}$, a fact we will use in a moment.

You will note that by throwing extra gridlines into a partition we do not diminish M_P. In fact, the areas of all the old rectangles will be included in the new sum (possibly broken into smaller pieces) plus some new rectangles could be added to the sum, formed by breaking up larger rectangles that formerly were too big to fit entirely inside $\mathfrak{O}$. A partition Q formed by adding gridlines to a partition P is called a **refinement of** P. So for any partition P we can find a partition Q of arbitrarily small mesh and $M_Q \geq M_P$.

The set of all sums of the form M_P for all possible partitions is a set of real numbers bounded below by 0 and bounded above by N^2. So this set of sums has a least upper bound which we will denote A.

Suppose P is a partition and $A - M_P < \varepsilon$. From the remarks above we may presume that the mesh of P is as small as we want: say less than $\frac{1}{n}$ for a positive integer n.

Choose an integer k so big that $\frac{2N^2}{k} < \varepsilon$. So if Q is any partition whose mesh is less than $\frac{1}{nk}$ we know from the calculations above that

$$A - M_Q < A - M_P + M_P - M_Q < \varepsilon + \varepsilon = 2\varepsilon.$$

We have just shown that there is a number to which all the sums M_Q are arbitrarily close provided only that the mesh of Q is small enough. This number, A from above, is called the **area** of $\mathfrak{O}$.

Suppose that h is a bounded and continuous function defined on $\mathfrak{O}$. You will recall that the function h is bounded on $\mathfrak{O}$ provided that there is a positive number M for which $-M < h < M$ everywhere on $\mathfrak{O}$.

A set of points C with members $C_{i,j}$ for $i = 1 \ldots n$ and $j = 1 \ldots m$ in the plane is called **subordinate to the partition** P if $C_{i,j}$ is in the subrectangle $[X_{i-1}, X_i] \times [Y_{j-1}, Y_j]$ for each i and j.

Consider the sum $\sum_{\mathfrak{O}} h(C_{i,j})\Delta X_i \Delta Y_j$ where this notation indicates that the sum is over those subscripts corresponding to rectangles in P which are entirely inside $\mathfrak{O}$. A sum of this kind, which depends on h, C and P, is called a **Riemann sum**. We will denote this sum $Riemann(C, P)$ in some calculations below.

Select integer n so large that if the mesh of P is less than $\frac{1}{n}$ then $A - M_P < \varepsilon$. Now select k so large that if the mesh of Q is less than $\frac{1}{kn}$ then $M_P - M_Q < \varepsilon$.

It is now necessary to appeal to a fact we don't prove. A continuous function defined on a closed set is uniformly continuous on that closed set. In our situation we will use that fact as follows: let $\mathfrak{K}$ denote the closed set formed from all the closed rectangles from P which are entirely inside $\mathfrak{O}$. Now let k be (possibly) even larger than before: let k be so large that if G and H are any two points from $\mathfrak{K}$ and the distance between these two points is less than $\frac{1}{nk}$ then $|h(G) - h(H)| < \varepsilon$.

If Q^1 and Q^2 are partitions the **common refinement** of Q^1 and Q^2 is the partition obtained by combining all the gridlines from both partitions into a third partition, the "coarsest" partition which is a refinement of both Q^1 and Q^2.

Suppose Q^1 and Q^2 are partitions with mesh less than $\frac{1}{2nk}$ with common refinement Q and suppose $C^1_{i,j}$ is subordinate to Q^1 and $C^2_{i,j}$ is subordinate to Q^2.

We are going to select two points for each rectangle in Q. These selections will not be subordinate to Q, but they will be within a distance of $\frac{1}{2nk}$ of a point in the rectangle and hence within $\frac{1}{nk}$ of each other. Each rectangle in Q is a piece of exactly one rectangle from Q^1 and also exactly one rectangle from Q^2. If a rectangle $[X_{i-1}, X_i] \times [Y_{j-1}, Y_j]$ from Q is a part of the rectangle from Q^1 containing the point $C^1_{l,m}$ we define $C_{i,j} = C^1_{l,m}$. If a rectangle $[X_{i-1}, X_i] \times [Y_{j-1}, Y_j]$ from Q is a part of the rectangle from Q^2 containing the point $C^2_{u,w}$ we define $B_{i,j} = C^2_{u,w}$.

$$
\begin{aligned}
&|Riemann(C^1, Q^1) - Riemann(C^2, Q^2)| \\
&\qquad \leq \sum_{\mathfrak{O}} |h(C_{i,j}) - h(B_{i,j})| \ \Delta X_i \Delta Y_j \\
&\qquad = \sum_{\substack{\text{those rectangles} \\ \text{entirely contained} \\ \text{in } \mathfrak{K}}} |h(C_{i,j}) - h(B_{i,j})| \ \Delta X_i \Delta Y_j \\
&\qquad\qquad + \sum_{\substack{\text{those rectangles} \\ \text{which extend} \\ \text{beyond } \mathfrak{K}}} |h(C_{i,j}) - h(B_{i,j})| \ \Delta X_i \Delta Y_j \\
&\qquad < A\varepsilon + 2M2\varepsilon \\
&\qquad = (A + 4M)\varepsilon
\end{aligned}
$$

We have just shown that under these conditions there is a number denoted

$$\int_{\mathfrak{O}} h(X, Y) \ dX \ dY$$

to which every Riemann sum is arbitrarily close provided only that the mesh of P is small enough.

We now enshrine in exercises certain facts which are often useful.

Exercise VII. Show that if f is uniformly continuous on a bounded set in $\mathbb{R}^2$ then it is bounded.

Exercise VIII. Show that if f is bounded and continuous on a bounded open set in $\mathbb{R}^2$ and $\varepsilon > 0$ then there is a finite list of closed squares $\overline{S_i}$ with $i = 1 \ldots n$ inside $\mathfrak{O}$ with the following properties: First, the squares do not touch each other except possibly on their boundaries. Second, the part of $\mathfrak{O}$ outside of all these squares has area less than ε. Third, if S_i is the open square inside of and with the same edges as $\overline{S_i}$ for each i then

$$\left| \int_{\mathfrak{O}} f(X,Y)\ dXdY - \sum_{i=1}^{n} \int_{S_i} f(X,Y)\ dXdY \right| < \varepsilon.$$

Note 36, From Page 141:

We will prove Fubini's Theorem in the following case: when the set $\mathfrak{O}$ is the bounded open rectangle $(a,b) \times (c,d)$ and for functions h that are uniformly continuous on $(a,b) \times (c,d)$.

This is sufficient (for many applications) in light of the exercises from the last note.

Following the text, $S_Y = (a,b)$ for each Y in (c,d) so $\mathfrak{B}(Y) = \int_a^b h(X,Y)\ dX$.

Suppose $\varepsilon > 0$ and let n be an integer so big that if G and H are points in the rectangle and $|G - H| < \frac{1}{n}$ then $|h(G) - h(H)| < \varepsilon$. If numbers Y_1 and Y_2 are in (c,d) and $|Y_1 - Y_2| < \frac{1}{n}$ then

$$\left| \int_a^b h(X,Y_1)dX - \int_a^b h(X,Y_2)dX \right| \leq \int_a^b |h(X,Y_1) - h(X,Y_2)|dX < \varepsilon(b-a).$$

So $\mathfrak{B}$ is uniformly continuous on (c,d).

Exercise IX. Show that if the mesh of partitions P and Q of (a,b) are both less than $\frac{1}{2n}$ then the difference between any two Riemann sum approximations to $\mathfrak{B}(Y)$ using partitions P and Q cannot exceed $(b-a)\varepsilon$ in magnitude. This implies both must be within $(b-a)\varepsilon$ of $\mathfrak{B}(Y)$.

Further, we choose n big enough so that whenever the mesh of a partition P of (c,d) is less than $\frac{1}{n}$ and for any selection of points C subordinate to P then the Riemann sum $Riemann(C,P)$ formed from $\mathfrak{B}$, C and P is close to $\int_c^d \mathfrak{B}(Y)dY$.

Specifically:

$$\left| \int_c^d \boldsymbol{\mathcal{B}}(Y)dY - Riemann(C,P) \right| < \varepsilon.$$

Finally select integer k so big that both $\frac{b-a}{2^k} < \frac{1}{2n}$ and $\frac{d-c}{2^k} < \frac{1}{2n}$ and also so big that

$$\left| \int_{\mathfrak{O}} h(X,Y)dXdY - \mathbb{W} \right| < \varepsilon$$

where $\mathbb{W}$ is the regularly spaced Riemann sum approximation:

$$\mathbb{W} = \frac{(b-a)(d-c)}{4^k} \sum_{i,j=1}^{2^k-1} h\left(a + \frac{i(b-a)}{2^k},\ c + \frac{j(d-c)}{2^k} \right).$$

All the elements are in place to show that $\int_{\mathfrak{O}} h(X,Y)dXdY$ is close to $\int_c^d \boldsymbol{\mathcal{B}}(Y)dY$, but we are facing a **notational debacle** here if we are not careful, so we will proceed in steps. First let

$$\mathbb{W}_j = \frac{(b-a)}{2^k} \sum_{i=1}^{2^k-1} h\left(a + \frac{i(b-a)}{2^k},\ c + \frac{j(d-c)}{2^k} \right).$$

Define

$$\boldsymbol{\mathcal{B}}_j = \boldsymbol{\mathcal{B}}\left(c + \frac{j(d-c)}{2^k} \right).$$

From this we have two facts:

$$\mathbb{W} = \frac{(d-c)}{2^k} \sum_{j=1}^{2^k-1} \mathbb{W}_j \qquad \text{and} \qquad |B_j - W_j| < \varepsilon.$$

We also know that

$$\left| \int_c^d \boldsymbol{\mathcal{B}}(Y)dY - \frac{(d-c)}{2^k} \sum_{j=1}^{2^k-1} \boldsymbol{\mathcal{B}}_j \right| < \varepsilon.$$

Finally by the triangle inequality:

$$\left| \mathbb{W} - \frac{(d-c)}{2^k} \sum_{j=1}^{2^k-1} \boldsymbol{\mathcal{B}}_j \right| \leq \frac{(d-c)}{2^k} \sum_{j=1}^{2^k-1} \left| \mathbb{W}_j - \boldsymbol{\mathcal{B}}_j \right| < (d-c)\varepsilon.$$

With these facts in hand we have, once again by the triangle inequality:

$$\left| \int_{\mathfrak{O}} h(X,Y)dXdY - \int_c^d \mathcal{B}(Y)dY \right|$$

$$\leq \left| \int_{\mathfrak{O}} h(X,Y)dXdY - \mathbb{W} \right|$$

$$+ \left| \mathbb{W} - \frac{(d-c)}{2^k} \sum_{j=1}^{2^k-1} \mathcal{B}_j \right|$$

$$+ \left| \int_c^d \mathcal{B}(Y)dY - \frac{(d-c)}{2^k} \sum_{j=1}^{2^k-1} \mathcal{B}_j \right|$$

$$< \varepsilon + (b-c)\varepsilon + \varepsilon.$$

Since ε can be chosen as small as we wish, we have shown that the iterated and double integrals are equal for uniformly continuous functions on bounded rectangles.

Note 37, From Page 146:

Maple commands (to be typed after the command prompt) to explore this function are:

$with(plots)$;

$plot3d(\ cos(x) * cos(y)\ +\ 3\ ,\ x = 0..4 * Pi\ ,\ y = 0..4 * Pi\ )$;

$int(\ int(\ sqrt(\ sin(x)\hat{}2\ *\ cos(y)\hat{}2\ +\ sin(y)\hat{}2\ *\ cos(x)\hat{}2\ +\ 1)\ ,\ x = 0..4 * Pi\ )\ ,\ y = 0..4 * Pi\ )$;

$evalf(\ int(\ int(\ sqrt(\ sin(x)\hat{}2 * cos(y)\hat{}2 + sin(y)\hat{}2 * cos(x)\hat{}2 + 1)\ ,\ x = 0..4 * Pi\ )\ ,\ y = 0..4 * Pi\ )\ )$;

The first two lines create the graph. The third shows you the integral setup without actually trying to calculate it. The fourth line gives a floating point numerical estimate.

Note 38, From Page 147:

It is a fact that under these conditions it is not necessary to assume that $\mathcal{U}$ is open: it **must** be open.

Specifically, if f is one-to-one and continuous from an open set $\mathcal{W}$ in $\mathbb{R}^2$ with values in $\mathbb{R}^2$ then the collection of values $f(s,t)$ for (s,t) in $\mathcal{W}$ constitute an open subset in the plane.

This is called an **invariance of domain** result and is quite hard to show. It is easier to show if you assume also that f' is continuous and nonzero.

Note 39, From Page 152:

Here are Maple commands to create four longer gridlines for this coordinate system as well as four short pieces which surround a roughly polygonal piece of the XY plane.

$plot(\ \{$
$[\ f(u,.2)[1],\ f(u,.2)[2],\ u=.2..2\ ]\ ,$
$[\ f(.2,u)[1],\ f(.2,u)[2]\ ,\ u=.2..3]\ ,$
$[\ f(u,3)[1],\ f(u,3)[2],\ u=.2..2]\ ,$
$[f(2,u)[1]\ ,\ f(2,u)[2]\ ,\ u=.2..3]\ ,$
$[\ f(u,1)[1]\ ,\ f(u,1)[2]\ ,\ u=1.5..1.6]\ ,$
$[\ f(1.5,u)[1]\ ,\ f(1.5,u)[2]\ ,\ u=1..1.1]\ ,$
$[\ f(u,1.1)[1]\ ,\ f(u,1.1)[2]\ ,\ u=1.5..1.6]\ ,$
$[\ f(1.6,u)[1]\ ,\ f(1.6,u)[2]\ ,\ u=1..1.1]\ \}\ )\ ;$

Note 40, From Page 152:

Here are Maple commands for gridlines (nearly) surrounding the region of interest. Following that are two integrals which are germane to the problem.

$plot(\ \{$
$[f(u,.01)[1]\ ,\ f(u,.01)[2]\ ,\ u=.01..ln(4)]\ ,$
$[f(.01,u)[1]\ ,\ f(.01,u)[2]\ ,\ u=.01..3.14]\ ,$
$[f(u,3.14)[1],\ f(u,3.14)[2]\ ,\ u=.01..ln(4)]\ ,$
$[f(ln(4),u)[1]\ ,\ f(ln(4),u)[2]\ ,\ u=.01..3.14]\ \}\ )\ ;$

$int(\ int(\ 1\ ,\ y=0..sqrt(\ (15/8)\hat{}2-(15/17)\hat{}2*x\hat{}2\ )\ )\ ,\ x=-17/8..17/8\ )\ ;$

$int(\ int(\ (sin(t))\hat{}2\ +\ (sinh(s))\hat{}2\ ,\ s=0..ln(4)\ )\ ,\ t=0..Pi\ )\ ;$

Note 41, From Page 152:

If you replace the condition

> (iii) The vector $A'_s(t)\times B'_t(s)$ is never 0.

by

> $(iii)'$ $\vec{k}\cdot A'_s(t)\times B'_t(s)$ is never 0.

then you can replace

> (i) We presume that f is one-to-one and that the values $f(s,t)$ for (s,t) in $\mathbf{\mathcal{W}}$ constitute part of the graph of a differentiable function g restricted to an open subset $\mathbf{\mathcal{U}}$ of its domain $\mathbf{\mathcal{O}}$.

by

> $(i)'$ We presume that f is one-to-one.

In other words, if the tangent plane never goes vertical then the collection of values $f(s,t)$ for (s,t) in $\mathbf{\mathcal{W}}$ **must** be the graph of a function g defined on an open set $\mathbf{\mathcal{U}}$ in the plane.

From the last note we know that if $f=\langle X,Y,Z\rangle$ is one-to-one then so is $\overline{f}=\langle X,Y\rangle$. Since f is differentiable so is $\overline{f}$ and so the collection of all $\overline{f}(s,t)$ is

an open set $\mathcal{U}$ in the plane by invariance of domain. $\overline{f}$ has an inverse function which is itself differentiable. So $g = Z \circ \overline{f}^{\,-1}$ is the function required in the original condition (i).

Note 42, From Page 159:

Maple commands which might help to understand this problem:

$plot3d(\ [\ sin(t)\ ,\ exp(t) * s\ ,\ t * ln(s)\ ]\ ,\ s = 1..2\ ,\ t = 0..1\ ,\ axes = NORMAL\ ,\ thickness = 2\ ,\ shading = Z\ )\ ;$

$evalf(\ Int(\ Int(\ sqrt(\ exp(2 * t) * (t\ \hat{}\ 2\ +\ (ln(s))\ \hat{}\ 2\ -\ 2 * t * ln(s)\ +\ (cos(t))\ \hat{}\ 2\ )\ +\ t\ \hat{}\ 2\ *(cos(t))\ \hat{}\ 2\ /s\ \hat{}\ 2\ )\ ,\ s = 1..2\ )\ ,\ t = 0..1\ )\ )\ ;$

Note 43, From Page 163:

In this note we will discuss orientability on two dimensional manifolds in space. You will recall that in Section 30 we defined these to be sets $\mathcal{M}$ in space which can be formed from **overlapping patches**, each of which is the graph of one variable as a continuously differentiable function of the other two, where the free variables are drawn from an **open set**.

At any point on a surface, there are exactly two unit normal vectors, and these two vectors depend on the geometry of the surface and not the parameterization. So at any point on a two dimensional differentiable manifold there are exactly two unit normal vectors.

Let $\mathcal{M}$ be a two dimensional manifold. For each point P of $\mathcal{M}$ let $\mathcal{N}_P$ denote a choice of one of the two unit normals to $\mathcal{M}$ at P.

$\mathcal{N}$ is called a unit normal vector field on $\mathcal{M}$.

We will call $\mathcal{N}$ **consistent** if $\mathcal{N}$ is a continuous vector valued function on the domain of every patch used to create $\mathcal{M}$.

We will call $\mathcal{M}$ **orientable** if there is a consistent unit normal vector field. A choice of one of these is called an **orientation** for $\mathcal{M}$. An **oriented manifold** is a manifold $\mathcal{M}$ together with a choice $\mathcal{N}$ of an orientation.

We are in no position to create general conditions to decide when or if a manifold is orientable. Unlike orientability on curves or surfaces that are graphs, the situation here is subtle and hard.

The Möbius strip is an example of a manifold which has no orientation but most easily constructed manifolds, such as the sphere, are orientable.

But if we *are* on an orientable manifold then there is an orientation which will be consistent with *any* admissible decomposition of $\mathcal{M}$, which can then be used to define flux through a surface in a way that does not depend on the decomposition used to calculate it.

Note 44, From Page 164:

In this note you construct triple integrals:

Exercise X. Modify the discussion in the note above within which we discuss the

construction of the double integral to handle the $3D$ *case.*

Note 45, From Page 166:

In this note you think about the $3D$ Fubini Theorem:

Exercise XI. Modify the discussion in the note above within which we discuss Fubini's Theorem for double integrals to handle the triple integral case.

Note 46, From Page 168:

Once again, the $3D$ version of **invariance of domain** holds: if f is one-to-one and continuous from an open set $\mathcal{W}$ in $\mathbb{R}^3$ with values in $\mathbb{R}^3$ then the set of those values **must be** an open set in $\mathbb{R}^3$.

Index

(a, b), 11
$\langle a, b \rangle$, 11
0, 3
$1D$, 75
$2D$, 26
$3D$, 26
A^T, 128
$D_1 g,\ D_X g,\ \frac{\partial g}{\partial X}$, 107
$F(\mathcal{S}) \cdot \mathcal{N}(\mathcal{S})\ d\mathcal{S}$, 162
$L_{Q,c}(t)$, 76
$L_{g,P}(Q)$, 109, 119
$Proj_{\ W}(V) = \left(\frac{V \cdot W}{W \cdot W}\right) W$, 22
$Q'(c)$ or $\frac{dQ}{dt}(c)$ or $\frac{d}{dt}Q(c)$, 67
$Q(t) = P + tV$, 18
Q'' or Q''' or $Q^{(n)}$ or $\frac{d^n Q}{dt^n}$, 67
$V \cdot W$, 13, 27
$V \times W =$
$\langle v_2 w_3 - v_3 w_2\ ,\ v_3 w_1 - v_1 w_3\ ,\ v_1 w_2 - v_2 w_1 \rangle$, 34
$V\ = |V| \left\langle \frac{v_1}{|V|}, \frac{v_2}{|V|} \right\rangle\ = |V| \langle cos(\theta), sin(\theta) \rangle$, 11
$V \cdot W\ = |V||W| cos(\theta)$, 14
X_i, Y_i or Z_i (Chapter V), 147
$[a, b] \times [c, d]$, 138
$[a_1, b_1] \times [a_2, b_2] \times [a_3, b_3]$, 163
$\mathfrak{T}$, 88
$\det A$, 14, 34
$\det(P, V, W)$, 34
$\det(V, W)$, 14
$\int_C^D g(Q)\ |dQ|$, 87
$\int_{\mathfrak{C}} g(s)\ ds$, 87
$\int_{\mathfrak{C}} F(s) \cdot \mathfrak{T}(s)\ ds$, 90
$\int_{\mathfrak{O}} h(X, Y)\ dX\ dY$, 139, 206
$\int_{\mathfrak{O}} h(X, Y, Z)\ dX\ dY\ dZ$, 164
$\int_{\mathcal{S}}\ h\ d\mathcal{S}$, 160
$\int_r^s Q(t) dt$, 71
$(Q - P) \cdot N = 0$, 48
$\lim_{Q \to P} g(Q)$, 104, 118
$\lim_{t \to c} Q(t)$, 66
$\mathbb{R}$, 66
$\mathbb{R}^n$, 66
$\nabla \cdot F$, 175
$\nabla_V g(P)$, 112, 120
∇g, 106, 119
∇, 175
$\nabla \times F$, 178
$\sum h(C_{i,j,k}) \Delta X_i \Delta Y_j \Delta Z_k$, 164
$\sum_{\mathfrak{O}} h(C_{i,j}) \Delta X_i \Delta Y_j$, 138, 206
$\vec{i}$, 12, 27, 75
$\vec{j}$, 12, 27
$\vec{k}$, 27
$curl\ F$, 178
$div\ F$, 175
g', 129
g_1 and g_2 (Chapter V), 143
$grad\ g$, 106, 119
$rot\ F$, 178
$|V \times W| = |V||W|\ sin(\theta)$, 36
$|V|$, 11, 26

acceleration, 68
admissible decomposition, 160
alien
  from Arcturus, 59
  weird, 112
angle
  between two planes in $3D$, 39
  between two vectors in $2D$, 14
  between two vectors in higher dimensions, 32
antisymmetric, 35
approximation, 66
  derivative, 67
  integral, 71
arclength, 85, 198
  weighted by, 87
area, 205
  of a bounded open set in $2D$, 141
  of a parallelogram in $2D$, 15
  of a parallelogram in $3D$, 36
  of a shadow, 39
  of a surface, 145, 149
  polar coordinates, 99, 150

arrow, 2

bearings, 15
Bezier curves, 80
boundary
  of a plane set, 139
bounded
  function, 205
  in $\mathbb{R}^2$, 138
  in $\mathbb{R}^3$, 163
  sequence, 190
  set of real numbers, 199
bug, 27, 95

Cauchy Mean Value Theorem, 192
central force, 75
chain rule, 68, 111, 120, 129
change of variables, 71, 148, 154, 168, 172, 173
charge, 87, 141, 145, 149, 166, 170
circulation
  around a loop, 89
closed set
  in space, 118, 165
  in the plane, 104, 140
  of real numbers, 199
  rectangle, 138
  rectangular solid, 163
coefficient of static friction, 24
collision, 58
common refinement, 206
complement, 104
components, 11
composite surface, 160
  oriented, 162
cone, 44
confined to a line or plane, 72
conservative vector field, 132
consistent
  choice of unit normal, 211
  choice of unit normal on a surface, 161
  parameterization, 161
constant velocity motion, 6
constraint, 23, 131
continuous, 66, 107, 119, 189
  in space, 118
  in the plane, 105
  on an interval, 66
  uniformly, 189, 203
continuously differentiable, 71
control, 25
converge
  sequence, 190
  vector function, 189
coordinate
  grid, 148, 153, 169
  plane, 26
coordinates
  cylindrical, 171
  hyperbolic-elliptic, 152
  parabolic, 151
  polar, 95, 130, 150
  rectangular, 26
  spherical, 171
cross product, 34
curl, 178
curve, 55
  good, 85
  in a surface, 109
  oriented, 88
  piecewise good, 90
cusp, 79, 89
cycloid, 60
cylinder, 45
cylindrical coordinates, 171

debacle
  notational, 208
decomposition, 8
  admissible, 160
Del operator, 175
derivative, 67, 129
determinant, 14, 35
developable
  surface, 44
differentiable, 67, 106, 119
  continuously, 71
  on an interval, 67
direction, 2
  cosines, 28
  opposite, 3
  same, 3
  vector, 12, 27
directional derivative, 112, 120
displacement, 4
divergence, 175
Divergence Theorem, 178
  in the Plane, 178
dot product, 13, 27
double integral
  over a set in $\mathbb{R}^2$, 139
  parametric form, 149

eliminate the parameter, 18, 28
ellipsoid, 46
Extreme Value Theorem, 190

feature, 27, 95
feedback, 25
Fido, 37
flow
  along a curve, 89
flux
  past an oriented curve in $2D$, 92
  through an oriented surface, 162
force, 4
frequency, 60
friction, 24

Fubini's Theorem, 141, 166
function
  defined implicitly, 121
  real or real valued, 55
  vector or vector valued, 55

geometrical track, 19
good
  curve, 85
  loop, 85
good parameterization
  of a curve, 85
    piecewise, 90
  of a surface, 153
  of an open set in $\mathbb{R}^2$, 147
  of an open set in $\mathbb{R}^3$, 169
gradient, 106, 119
Greatest Lower Bound, 191
Green's Theorem
  Normal Form, 178
  Tangential Form, 180

helicoid, 45
helix, 55
hyperbolic-elliptic coordinates, 152

identity matrix, 129
implicit function, 121
improper integral, 141, 150
inclined plane, 23
instances, 2
integral, 71
  double, 139
    parametric form, 149
  improper, 141, 150
  iterated, 141, 166
  Lebesgue, 204
  line, 87
  surface, 145
    over a composite surface, 160
    parametric form, 155
  triple, 164
    parametric form, 170
  volume, 164
  weighted by arclength, 87
  weighted by surface area, 145
  weighted by volume, 164
integration
  by parts, 71
  by substitution, 71
Intermediate Value Theorem, 190
  for Derivatives, 192
interpolation
  Bezier, 83, 196
  linear, 62, 187
invariance of domain, 209, 212
inverse function derivative, 130
is at
    an object is at a vector, 17

iterated integral
  in $2D$, 141
  in $3D$, 166

Jacobian matrix, 129

L'Hôpital's Rule, 192
Lagrange Multiplier, 130
Least Upper Bound, 191
Lebesgue Integral, 204
level set, 125
lies in
    a vector lies in a line or a plane, 17
limit, 66, 105, 118
  vector function, 189
line integral, 87
linear combination, 3
linearization, 76, 109, 119
local
  extreme values, 114, 121
  maximum value, 114, 120
  minimum value, 114, 121
loop, 84
  good, 85
  piecewise good, 90
lower bound, 191

Möbius strip, 159, 211
magnitude, 2, 11, 26
manifold, 126, 211
mantra, 175
Maple, 44
mass, 87, 141, 145, 149, 166, 170
Mathematica, 44
matrix, 127
  addition, 128
  multiplication, 127
Mean Value Theorem, 68, 191
  Cauchy, 192
meaning, 37
members of an admissible decomposition, 160
mesh, 71, 138, 163, 194
mixed partial derivatives, 108
monkey saddle, 47
Monotone Convergence Theorem for Sequences, 190
monotone sequence, 190
Mr. Bacon's Train, 91
multiplication
  matrix, 128
  scalar, 3

normal, 14
  form
    for a line in $2D$, 19
    for a line in $3D$, 29
    for a plane in $3D$, 48
  vector for an oriented curve in $2D$, 92
Normal Form of Green's Theorem, 178

objective function, 131
octant, 26
one-to-one, 69
open set
  in space, 118
  in the plane, 104
  on the line, 199
order of a partial derivative, 109
orientable
  manifold or surface, 211
orientation
  for a composite surface, 162
  for a curve, 88
  for a manifold, 211
  for a piecewise good curve, 90
  for a surface, 161
oriented
  composite surface, 162
  curve, 88
  manifold or surface, 211
  piecewise good curve, 90
  surface, 161
origin, 6
orthogonal, 14
  coordinate system, 150

parabolic coordinates, 151
paraboloid, 46
parallelepiped, 36
parallelogram, 15, 36
parameter, 6
parameterized, 6
parametric
  form
    of a surface integral, 155
    of an integral in $2D$, 149
    of an integral in $3D$, 170
  vector equation
    for a line, 7, 18
    for a plane, 52
partial derivative, 107, 119
  mixed, 108
  order of, 109
  second and higher order, 108, 119
partition, 71, 194
  of a rectangle, 138
  of a rectangular solid, 163
path, 55
  connected, 132
  connecting two points, 132
perpendicular, 8, 14
Pete's World, 115, 132
piecewise good
  curve, 90
    oriented, 90
  loop, 90
  parameterization, 90
plane, 43
points at
    a vector points at a place in the plane or space, 17
polar coordinates, 95, 130, 150
position vector, 6
potential, 133
product rule, 68
projection
  scalar, 22
  vector, 22

quotient rule, 68

radial
  component, 98
  direction vector, 98
real
  valued
    function of one variable, 55
    function of three variables, 118
    function of two variables, 104
    sequence, 190
rectangle, 138
rectangular
  coordinates, 26
  solid, 163
refinement, 194, 205
  common, 194, 206
relative
  angle, 59
  position, 58
  speed, 59
  velocity, 58
resultant, 3
Riemann sum, 93, 138, 164, 173, 194, 206

saddle, 201
scalar
  multiplication, 3, 128
  product, triple, 34
  projection, 22
second
  derivative test, 114, 200
  partial derivatives, 108, 119
sequence
  real, 190
  vector, 191
shadow, 39
shift, 77
slime, 6
slug, 6
speed, 4, 68
  change, 77
spherical coordinates, 171
standard position, 6, 10
Stokes' Theorem, 180
  in the Plane, 180
subordinate, 138, 163, 194, 205
subsequence, 191

sum of two vectors, 2
surface, 42, 105
  area, 145
    weighted by, 145
  composite, 160
  developable, 44
  integral, 145
    over a composite surface, 160
    parametric form, 155
Susan's Hill, 113
symmetric, 13
synchronized, 58

tangent
  line, 79
  plane, 109
  vector for an oriented curve, 88
tangential
  component, 98
  direction vector, 98
Tangential Form of Green's Theorem, 180
Taylor Polynomial, 115, 121, 201
tension, 16
tensor, 13, 35
tetrahedron, 38
three dimensions, 26
time, 6
torus, 46, 126
translate among coordinate systems, 54
transpose, 128
transposition of a matrix, 128
triangle inequality, 34
triple
  integral
    over a set in $\mathbb{R}^3$, 164
    parametric form, 170
  scalar product, 34
two dimensions, 26

uniform continuity, 189, 203
unit
  circle, 12
  normal vector for an oriented curve in $2D$, 92
  sphere, 27
  tangent vector for an oriented curve, 88
  vector, 12
upper bound, 191

vector, 2
  addition, 3
  direction, 12, 27
  field, 107, 119
    conservative, 132
  projection, 22
  sequence, 191
  unit, 12
  valued function, 55
    in the plane, 107
    in the space, 119
velocity, 4, 68
volume
  integral, 164
    parametric form, 170
  of a parallelepiped, 37
  of an open set in $3D$, 166
  under a surface, 141
  weighted by, 164

weighted
  by arclength, 87
  by surface area, 145
  by volume, 164
Weird Alien, 112
work, 22, 89

zero vector, 3