

---

Differential Calculus  
of  
Several Variables

---

David Perkinson

ABSTRACT. These are notes for a one semester course in the differential calculus of several variables. The first two chapters are a quick introduction to the derivative as the best affine approximation to a function at a point, calculated via the Jacobian matrix. Chapters 3 and 4 add the details and rigor. Chapter 5 is the basic theory of optimization: the gradient, the extreme value theorem, quadratic forms, the Hessian matrix, and Lagrange multipliers. Studying quadratic forms also gives an excuse for presenting Taylor's theorem. Chapter 6 is an introduction to differential geometry. We start with a parametrization inducing a metric on its domain, but then show that a metric can be defined intrinsically via a first fundamental form. The chapter concludes with a discussion of geodesics. An appendix presents (without proof) three equivalent theorems: the inverse function theorem, the implicit function theorem, and a theorem about maps of constant rank.

Version: January 31, 2008

---

# Contents

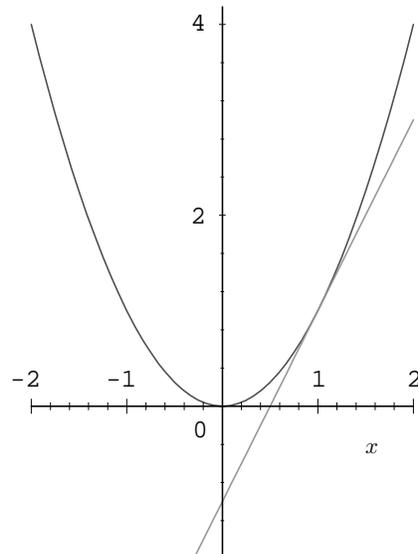
Chapter 1. Introduction	1
§1. A first example	1
§2. Calculating the derivative	6
§3. Conclusion	13
Chapter 2. Multivariate functions	17
§1. Function basics	17
§2. Interpreting functions	19
§3. Conclusion	31
Chapter 3. Linear algebra	35
§1. Linear structure	35
§2. Metric structure	37
§3. Linear subspaces	43
§4. Linear functions	49
§5. Conclusion	57
Chapter 4. The derivative	65
§1. Introduction	65
§2. Topology	65
§3. Limits, continuity	68
§4. The definition of the derivative	70
§5. The best affine approximation revisited	72
§6. The chain rule	73
§7. Partial derivatives	77
§8. The derivative is given by the Jacobian matrix	79
§9. Conclusion	83
Chapter 5. Optimization	89
§1. Introduction	89

---

§2. Directional derivatives and the gradient	89
§3. Taylor's theorem	92
§4. Maxima and minima	97
§5. The Hessian	106
Chapter 6. Some Differential Geometry	115
§1. Stretching	115
§2. First Fundamental Form	116
§3. Metrics	119
§4. Lengths of curves	120
§5. Geodesics	122
Appendix A. Set notation	133
Appendix B. Real numbers	137
§1. Field axioms	137
§2. Order axioms	138
§3. Least upper bound property	138
§4. Interval notation	139
Appendix C. Maps of Constant Rank	141
Appendix. Index	145

# Introduction

The main point of differential calculus is to replace curvy things with flat things: to approximate complicated functions with linear functions. For example, in one variable calculus, one approximates the graph of a function using a tangent line:



In the illustration above, the function  $g(x) = x^2$  is replaced by the simpler function  $\ell(x) = 2x - 1$ , a good approximation near the point  $x = 1$ . We begin these notes with an analogous example from multivariable calculus.

## 1. A first example

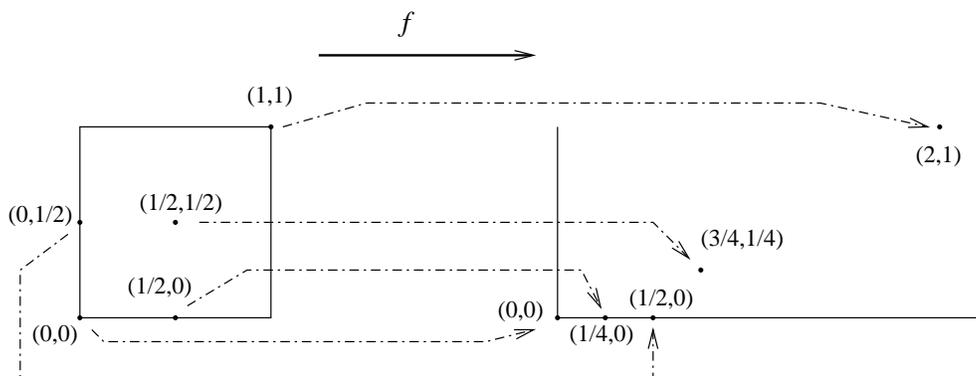
Consider the function  $f(u, v) = (u^2 + v, uv)$ . It takes a point  $(u, v)$  in the plane and hands back another point,  $(u^2 + v, uv)$ . For example,

$$f(2, 3) = (2^2 + 3, 2 \cdot 3) = (7, 6).$$

Our first task is to produce a picture of  $f$ . In one variable calculus, one usually pictures a function by drawing its graph. For example, one thinks of the function  $g(x) = x^2$  as a parabola by drawing points of the form  $(x, x^2)$ . If we try the same for our function, we

would need to draw points of the form  $(u, v, u^2 + v, uv)$ , which would require four dimensions. Further, in some sense, the graph of  $f$  is the simplest geometric realization of  $f$ : it contains exactly the information necessary to picture  $f$ , and no more. Thus, as usual in multivariate calculus, we are forced to picture an object that naturally exists in a space of dimension higher than three.

There are several ways of dealing with the problem of picturing objects involving too many dimensions, and in practice functions such as  $f$  arise in a context that suggests a particular approach. We will start with one important point of view. Suppose we want to picture  $f$  for  $u$  and  $v$  in the interval  $[0, 1]^*$ , so the points  $(u, v)$  lie in a unit square in the plane. Think of this unit square as a thin sheet of putty, and think of the function  $f$  as a set of instructions for stretching the putty into a new shape. For example, the point  $(1, 1)$  is stretched out to the point  $f(1, 1) = (1^2 + 1, 1 \cdot 1) = (2, 1)$ ; the point  $(\frac{1}{2}, \frac{1}{2})$  moves to  $f(\frac{1}{2}, \frac{1}{2}) = (\frac{3}{4}, \frac{1}{4})$ ; the origin,  $(0, 0)$ , remains fixed,  $f(0, 0) = (0, 0)$ . The motion of these points and a few others is shown below.



Our goal is to see where *every* point in the unit square is moved by  $f$ . To start, consider what happens to the boundary of the square:

**BOTTOM EDGE.** The bottom edge of the square consists of points of the form  $(u, 0)$  as  $u$  varies from 0 to 1. Applying  $f$  gives

$$f(u, 0) = (u^2 + 0, u \cdot 0) = (u^2, 0).$$

So as a point moves along the bottom edge at a constant unit speed from  $(0, 0)$  to  $(1, 0)$ , its image under  $f$  moves between the same two points, moving slowly at first, then more and more quickly, (velocity =  $2u$ ).

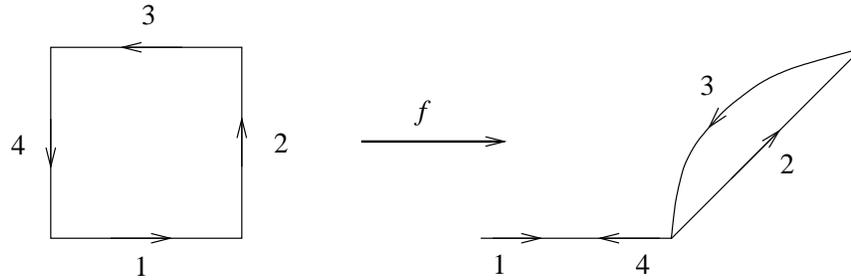
**RIGHT EDGE.** The right edge consists of points of the form  $(1, v)$ , and  $f(1, v) = (1 + v, v)$ ; the image is a line segment starting at  $(1, 0)$  when  $v = 0$  and ending at  $(2, 1)$  when  $v = 1$ .

**TOP EDGE.** Points along the top have the form  $(u, 1)$ , and  $f(u, 1) = (u^2 + 1, u)$ . Calling the first and second coordinates  $x$  and  $y$ , we have  $x = u^2 + 1$  and  $y = u$ . Thus, points in the image of the top edge satisfy the equation  $x = y^2 + 1$ , forming a parabolic arc from  $(2, 1)$  to  $(1, 0)$  as we travel from right to left along the top edge of the square.

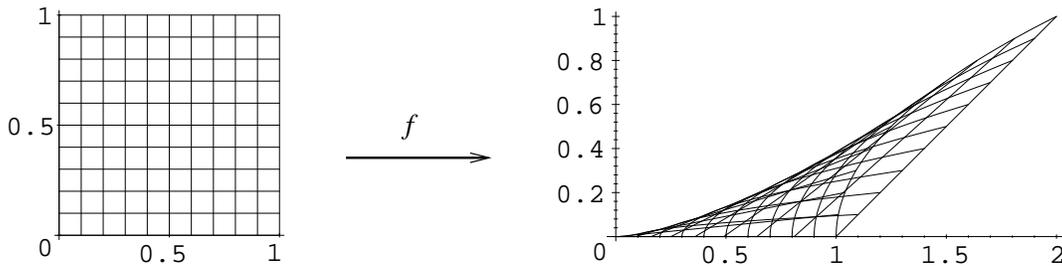
\* $[0, 1]$  denotes the real numbers between 0 and 1, including the endpoints. For a reminder on interval notation, see Appendix B.

LEFT EDGE. Points on the left edge have the form  $(0, v)$ , and  $f(0, v) = (v, 0)$ . Thus, we can picture  $f$  as flipping the left edge over a  $45^\circ$  line, placing it along the bottom edge.

The following picture summarizes what  $f$  does to the boundary of the unit square:



We now want to figure out what is happening to the interior of the square. Considering what happens to the boundary, it seems that  $f$  is taking our unit square of putty, folding it along a diagonal, more or less, and lining up the left edge with the bottom edge. To get more information, draw a square grid of lines on the unit square, and plot the images of each of these lines under  $f$ . A vertical line, a distance  $c$  from the origin, will consist of points of the form  $(c, v)$  with  $v$  varying, and  $f$  will send these points to points of the form  $f(c, v) = (c^2 + v, cv)$ . The image is a line passing through  $(c^2, 0)$  when  $v = 0$ , and  $(c^2 + 1, c)$  when  $v = 1$ . A horizontal line at height  $c$  will consist of points of the form  $(u, c)$  which are sent by  $f$  to points of the form  $f(u, c) = (u^2 + c, uc)$  lying on a parabolic arc connecting  $(c, 0)$  to  $(1 + c, c)$ . The following picture shows the image of a grid under  $f$ :



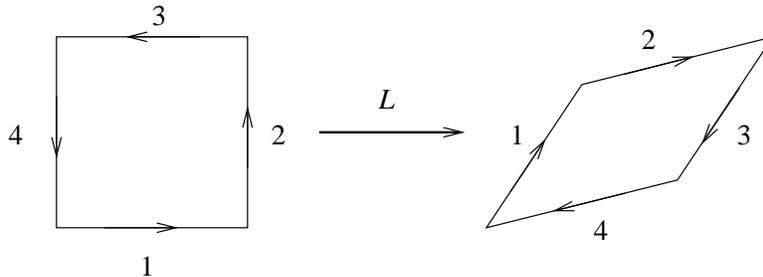
You should be starting to see that multivariate functions are more interesting than the functions one meets in one variable calculus. Even though  $f$  involves only a few very simple terms, its geometry is fairly complicated. Differential calculus provides one main tool for dealing with this complexity: it shows how to approximate a function with a simpler type of function, namely, a *linear* function. We will give a rigorous definition of a linear function later (cf. page 49). For now, it is enough to know that we can easily analyze a linear function; there is no mystery to its geometry. The image of a square grid under a linear function is always a parallelogram; the image of a grid will be a grid of parallelograms.

To be more precise, consider the function  $f$ . Given a point  $p$  in the unit square, differential calculus will give us a linear function that closely approximates  $f$  provided we stay near the point  $p$ . (Given a different point, calculus will provide a different linear function.) To illustrate the idea, take the point  $p = (\frac{1}{4}, \frac{3}{4})$ . The linear function provided by calculus turns out to be:

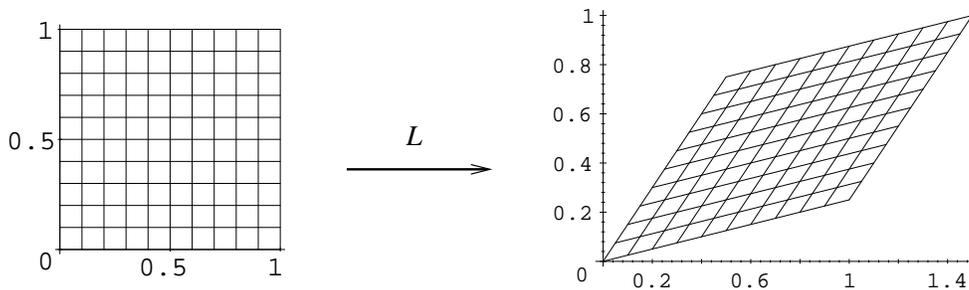
$$L(u, v) = (\frac{1}{2}u + v, \frac{3}{4}u + \frac{1}{4}v).$$

(To spoil a secret, read the footnote.<sup>†</sup>)

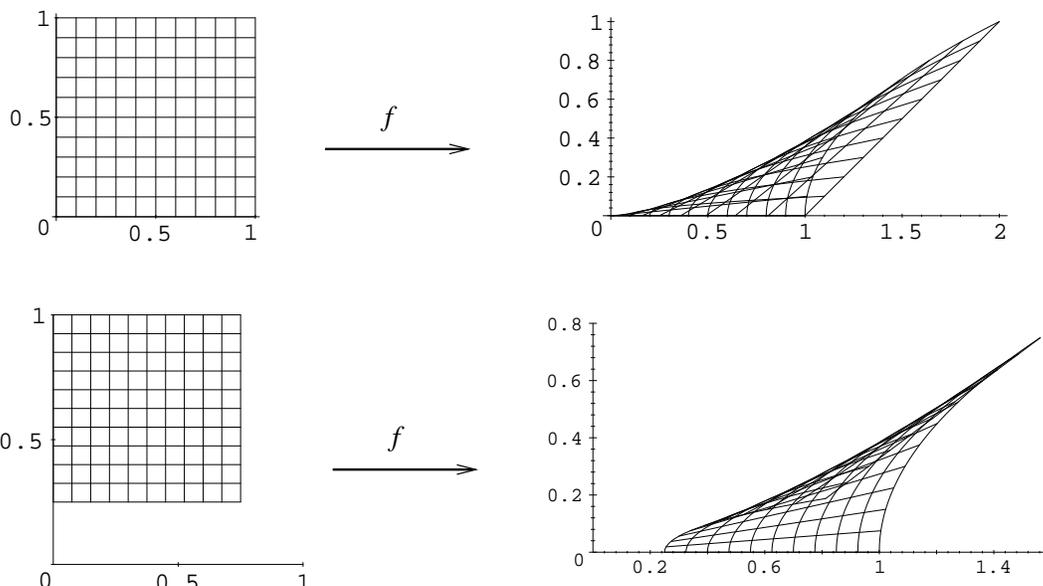
To picture  $L$ , consider the image of a unit square. The function  $L$  acts on the vertices as follows:  $L(0,0) = (0,0)$ ,  $L(1,0) = (\frac{1}{2}, \frac{3}{4})$ ,  $L(1,1) = (\frac{3}{2}, 1)$ , and  $L(0,1) = (1, \frac{1}{4})$ . It turns out that linear functions always send lines to lines; hence we can see how  $L$  acts on the boundary of the unit square:



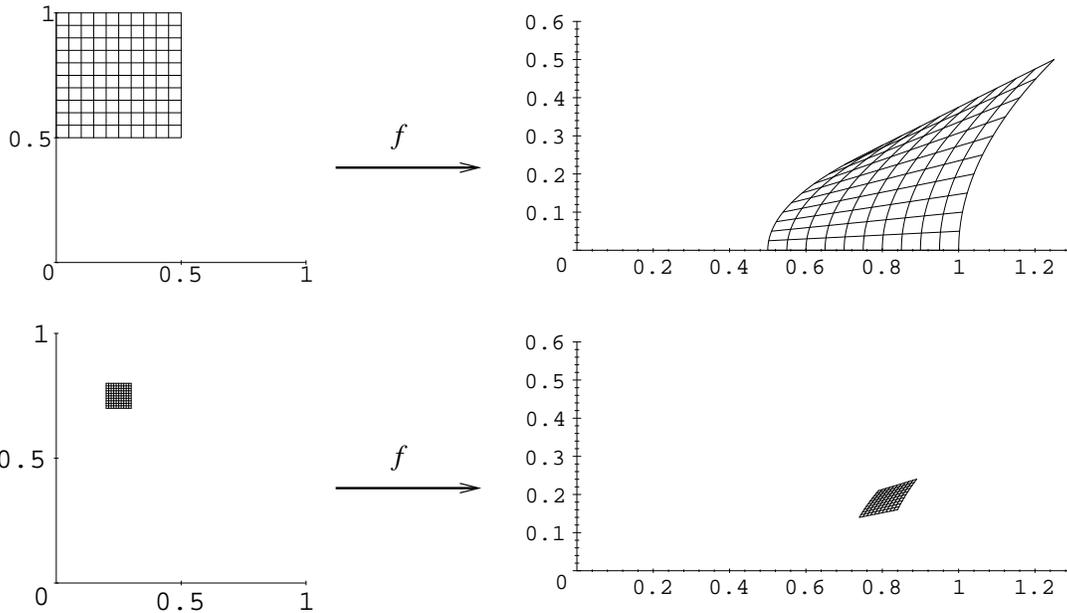
The function  $L$  stretches, skews, and flips the unit square. Here is what  $L$  does to a grid (changing scale a bit compared with the previous picture)



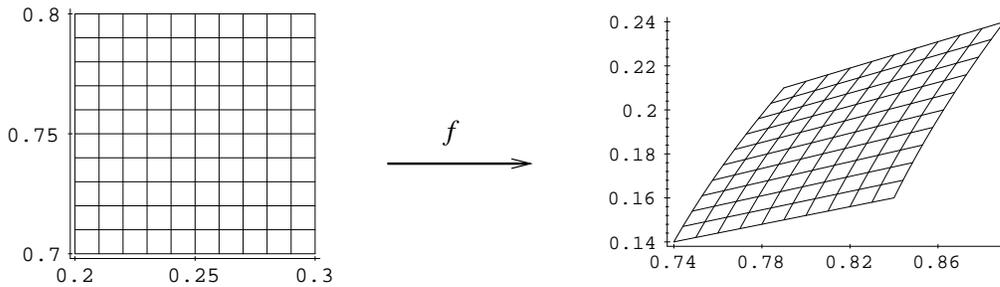
In what sense does  $L$  approximate  $f$ ? Certainly, the image of a grid under  $L$  looks much different from the earlier picture of the image of a grid under  $f$ . The idea is that  $L$  approximates  $f$  *locally*: it is only a good approximation near the chosen point,  $p = (\frac{1}{4}, \frac{3}{4})$ . Let us look at the images under  $f$  of successively smaller grids about  $p$ .



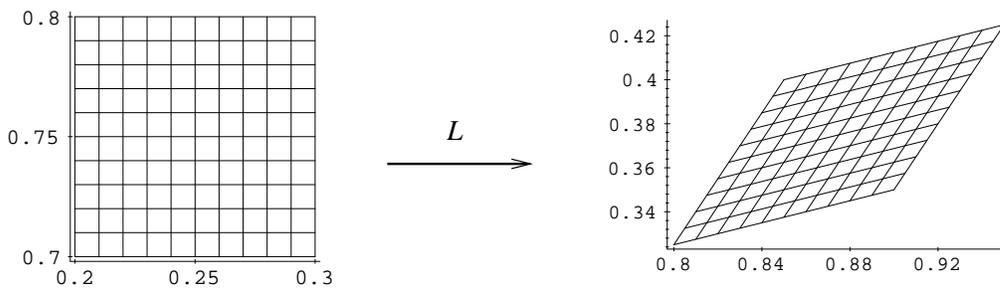
<sup>†</sup>The function  $L$  will later be denoted as  $Df_{(\frac{1}{4}, \frac{3}{4})}$  or as  $f'(\frac{1}{4}, \frac{3}{4})$ . It is the *derivative* of  $f$  at the point  $(\frac{1}{4}, \frac{3}{4})$ .



Here is a scaled-up version of the last picture (note the labeling of the axes):



Thus, if we look at the image under  $f$  of a *small* grid centered about  $(\frac{1}{4}, \frac{3}{4})$ , we get a slightly warped parallelogram (note that the lines are slightly curved). Compare this last picture with the earlier picture we had of the image of a grid under  $L$ . You should notice that the parallelogram there and the warped parallelogram above are very similar in shape. Here is a picture of the image under  $L$  of a small grid centered about  $p$ .



The **key thing** to see is that the parallelogram above is virtually the same as the warped parallelogram we had for  $f$  earlier, discounting a translation. **It has almost the same size and shape.** It turns out that the further we scale down, i.e., the closer we stay to the point  $(\frac{1}{4}, \frac{3}{4})$ , the better the match will be. That is what is meant when we say that  $L$  is a good linear approximation to  $f$  near  $(\frac{1}{4}, \frac{3}{4})$ .

## 2. Calculating the derivative

This section contains a crash course in calculating the derivative of a multivariate function. In that way, you can start to use calculus right away. Everything we do here will be covered in detail later in the notes, and as you read, your understanding will grow. The idea is that when we finally get around to making rigorous definitions and proving theorems, you will have enough experience to appreciate *what* is true and will be able to fully concentrate on *why* it is true.

**2.1. Euclidean  $n$ -space.** The geometry in this course takes place in a generalization of the regular one, two, and three-dimensional spaces with which you are already familiar. It is called Euclidean  $n$ -space and denoted  $\mathbb{R}^n$ . Recall that once an origin and scale are fixed, the position along a line can be given by a single number. The position in a plane, given an origin and two axes, can be determined by an ordered list of two numbers: the first number in the list tells you how far to move along the first axis and the second tells you how far to move parallel to the second axis. Similarly, an ordered list of three numbers can be used to determine positions in three-dimensional space. For example, the point  $(1, 2, -3)$  means: move along the first axis a distance of one unit, then move in the direction of the second axis a distance of two units, then move in the direction opposite that of the third axis a distance of three units. Three-dimensional space can be thought of as the collection of all ordered triples of real numbers.

*Euclidean  $n$ -space* is defined to be the collection of all ordered lists of  $n$  real numbers. Thus,  $(1, 4.6, -0.3, \pi/2, 17)$  is a point in 5-space. The 5-tuple,  $(7.2, 6.3, -3, 5, 0)$  is another. The *origin* in  $\mathbb{R}^5$  is  $(0, 0, 0, 0, 0)$ . The collection of all such 5-tuples is Euclidean 5-space. The point  $(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)$  is a point in 10-space. We will typically denote an arbitrary point in Euclidean  $n$ -space by  $(x_1, x_2, \dots, x_n)$ . One-dimensional space is usually denoted by  $\mathbb{R}$  rather than  $\mathbb{R}^1$ , and a typical point such as  $(4)$  is just written as 4. In other words, one-dimensional space is just the set of real numbers.

We will call the elements of  $\mathbb{R}^n$  either *points* or *vectors* depending on the context. Later, we will add extra structure to  $n$ -space so that we can measure distances and angles, and scale and translate vectors.

**2.2. Multivariate functions.** A function from  $n$ -space to  $m$ -space is something which assigns to each point of  $\mathbb{R}^n$  a unique point in  $\mathbb{R}^m$ . We say that the *domain* of the function is  $\mathbb{R}^n$  and the *codomain* is  $\mathbb{R}^m$ . In general, the assignment made by the function can be completely arbitrary, but the functions with which multivariable calculus deals are of a much more restricted class. A typical example of the type we will see is the function defined by

$$f(w, x, y, z) = (w^2 + y + 7, x - yz, w - 5y^2, xy + z^2, w + 3z).$$

The function  $f$  “maps” 4-space into 5-space, i.e., it sends points in  $\mathbb{R}^4$  into points in  $\mathbb{R}^5$ . For instance, the point  $(0, 1, 2, 3) \in \mathbb{R}^4$  is mapped by  $f$  to the point

$$\begin{aligned} f(0, 1, 2, 3) &= (0^2 + 2 + 7, 1 - 2 \cdot 3, 0 - 5 \cdot 2^2, 1 \cdot 2 + 3^2, 0 + 3 \cdot 3) \\ &= (9, -5, -20, 11, 9) \in \mathbb{R}^5. \end{aligned}$$

Similarly  $f(0, 0, 0, 0) = (7, 0, 0, 0, 0)$  and  $f(1, 1, 1, 1) = (9, 0, -4, 2, 4)$ .

The function  $f$  is made by listing five real-valued functions in order:

$$\begin{aligned} f_1(w, x, y, z) &= w^2 + y + 7 \\ f_2(w, x, y, z) &= x - yz \\ f_3(w, x, y, z) &= w - 5y^2 \\ f_4(w, x, y, z) &= xy + z^2 \\ f_5(w, x, y, z) &= w + 3z. \end{aligned}$$

These functions are called the *component functions* of  $f$ .

Here are a few other examples of functions:

1.  $g(x, y) = (x^2 - 3y + 2, x^4, x^2 - y^2, 0)$ , a function from  $\mathbb{R}^2$  to  $\mathbb{R}^4$ , i.e., a function with domain  $\mathbb{R}^2$  and codomain  $\mathbb{R}^4$ ;
2.  $h(x_1, x_2, x_3, x_4, x_5, x_6) = (x_1x_2 - x_3x_4, x_6)$ , a function with domain  $\mathbb{R}^6$  and codomain  $\mathbb{R}^2$ ;
3.  $\ell(t) = \cos(t)$ , a function with domain and codomain  $\mathbb{R}$ .

**2.3. Partial derivatives.** Let  $f$  be any function from  $\mathbb{R}^n$  to  $\mathbb{R}^m$ , and let  $p$  be a point in  $\mathbb{R}^n$ . We are working towards defining the derivative of  $f$  at  $p$ . To start, we will perform the simpler task of computing *partial derivatives*. The basic idea, which we will formulate as a definition later, is to pretend all the variables but one are constant and take the ordinary derivative with respect to that one variable. For example, consider the function from  $\mathbb{R}^3$  to  $\mathbb{R}$  defined by

$$g(x, y, z) = x^2 + xy^3 + 2y^2z^4.$$

To take the partial derivative of  $g$  with respect to  $x$ , pretend that  $y$  and  $z$  are constants and take the ordinary derivative with respect to  $x$ :

$$\frac{\partial g}{\partial x} = 2x + y^3.$$

To take the partial with respect to  $y$ , pretend the remaining variables,  $x$  and  $z$ , are constant. Similarly for the partial with respect to  $z$ :

$$\frac{\partial g}{\partial y} = 3xy^2 + 4yz^4, \quad \frac{\partial g}{\partial z} = 8y^2z^3.$$

These partials can be evaluated at points. For example, the partials of  $g$  at the point  $(1, 2, 3)$  are

$$\frac{\partial g}{\partial x}(1, 2, 3) = 1 \cdot 2 + 2^3 = 10, \quad \frac{\partial g}{\partial y}(1, 2, 3) = 3 \cdot 1 \cdot 2^2 + 4 \cdot 2 \cdot 3^4 = 660,$$

$$\frac{\partial g}{\partial z}(1, 2, 3) = 8 \cdot 2^2 \cdot 3^3 = 864.$$

The geometric interpretation is clear from one variable calculus. Once all but one of the variables is fixed, we are left with a function of one variable, and the partial derivative gives the rate of change as that variable moves. The rate of change of  $g$  as  $x$  varies at the point  $(1, 2, 3)$  is 10. The rate of change in the direction of  $y$  at that point is 660 and in the direction of  $z$  is 864.

Some other examples:

1. If  $g(u, v) = u^2v - v^3 + 3$ , then  $\partial g/\partial u = 2uv$  and  $\partial g/\partial v = u^2 - 3v^2$ . In particular,  $\partial g/\partial u(2, 3) = 12$  and  $\partial g/\partial v(2, 3) = -23$ .

2. If  $g(t) = \cos(t)$ , then  $\partial g/\partial t = dg/dt = -\sin(t)$ .

So far, we have just considered partial derivatives of real-valued functions. To take partials of functions with more general codomains, simply take the partials of each of the component functions. For example, using our  $f$  from above:

$$f(w, x, y, z) = (w^2 + y + 7, x - yz, w - 5y^2, xy + z^2, w + 3z),$$

we have

$$\frac{\partial f}{\partial w} = \left( \frac{\partial f_1}{\partial w}, \frac{\partial f_2}{\partial w}, \frac{\partial f_3}{\partial w}, \frac{\partial f_4}{\partial w}, \frac{\partial f_5}{\partial w} \right) = (2w, 0, 1, 0, 1)$$

$$\frac{\partial f}{\partial x} = (0, 1, 0, y, 0)$$

$$\frac{\partial f}{\partial y} = (1, -z, -10y, x, 0)$$

$$\frac{\partial f}{\partial z} = (0, -y, 0, 2z, 3)$$

and, for instance,

$$\frac{\partial f}{\partial y}(1, 2, 3, 4) = (1, -4, -30, 2, 0).$$

**2.4. The Jacobian matrix.** We are now ready to define the *Jacobian matrix*. In a sense, the main task of multivariable calculus, both differential and integral, is to understand the geometry hidden in the Jacobian matrix. For instance, you will soon see how to read off the derivative of a function from it. So do not be surprised if it is somewhat complicated to write down.

An arbitrary function with domain  $\mathbb{R}^n$  and codomain  $\mathbb{R}^m$  has the form  $f(x_1, \dots, x_n) = (f_1, \dots, f_m)$  where each of the component functions,  $f_i$ , is a real-valued function with domain  $\mathbb{R}^n$ . To make the dependence on  $n$  variables explicit, we can write  $f_i(x_1, \dots, x_n)$  instead of the shorthand  $f_i$ , just as we do for  $f$ , itself. The *Jacobian matrix* for  $f$  is a rectangular box filled with partial derivatives. The entry in the  $i$ -th row and  $j$ -th column is  $\partial f_i/\partial x_j$ :

$$Jf := \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}$$

Continuing with our function from above,

$$f(w, x, y, z) = (w^2 + y + 7, x - yz, w - 5y^2, xy + z^2, w + 3z),$$

we get

$$Jf = \begin{pmatrix} 2w & 0 & 1 & 0 \\ 0 & 1 & -z & -y \\ 1 & 0 & -10y & 0 \\ 0 & y & x & 2z \\ 1 & 0 & 0 & 3 \end{pmatrix}$$

It is difficult to remember which way to arrange the partial derivatives in the Jacobian matrix. It helps, and is later important conceptually, to see that while each entry in the matrix is a partial derivative of a component function,  $f_i$ , the columns of the matrix are the partial derivatives of the function  $f$ , itself. For instance, recall from above that

$$\frac{\partial f}{\partial w} = (2w, 0, 1, 0, 1).$$

Note how this partial derivative corresponds to the first column of  $Jf$ . Similarly, check that the other partials of  $f$  correspond to the remaining columns. Thus, we could write for an arbitrary function  $f$  with domain  $\mathbb{R}^n$ ,

$$Jf = \left( \frac{\partial f}{\partial x_1} \quad \frac{\partial f}{\partial x_2} \quad \cdots \quad \frac{\partial f}{\partial x_n} \right),$$

remembering to think of each partial derivative of  $f$  as a column.

Notice that the Jacobian matrix itself is a function of the coordinates in the domain. Hence, it can be evaluated at various points in the domain. For instance, setting  $w = 1$ ,  $x = 2$ ,  $y = 3$ , and  $z = 4$  above gives the Jacobian of  $f$  evaluated at the point  $(1, 2, 3, 4)$ :

$$Jf(1, 2, 3, 4) = \begin{pmatrix} 2 & 0 & 1 & 0 \\ 0 & 1 & -4 & -3 \\ 1 & 0 & -30 & 0 \\ 0 & 3 & 2 & 8 \\ 1 & 0 & 0 & 3 \end{pmatrix}$$

Another example: let  $g(u, v) = (u, v, u^2 - v^2)$ . The partials of  $g$  are

$$\frac{\partial g}{\partial u} = (1, 0, 2u), \quad \frac{\partial g}{\partial v} = (0, 1, -2v).$$

Thus, the Jacobian matrix of  $g$  is

$$Jg = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 2u & -2v \end{pmatrix}.$$

Note the correspondence between the partial derivatives of  $g$  and the columns of the matrix. Again, we can evaluate the Jacobian, say at the point  $(1, 4)$ :

$$Jg(1, 4) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 2 & -8 \end{pmatrix}.$$

The case where the domain or codomain is  $\mathbb{R}$  is perhaps a little tricky. For instance, let  $h(x, y, z) = x^2 + xy^3 - 3yz^4 + 2$ . The Jacobian matrix is

$$Jh = ( 2x + y^3 \quad 3xy^2 - 3z^4 \quad -12yz^3 ),$$

a matrix with a single row. On the other hand, the Jacobian matrix of  $c(t) = (t, t^2, t^3)$  has a single column:

$$Jc = \begin{pmatrix} 1 \\ 2t \\ 3t^2 \end{pmatrix}.$$

The most special case of all is the case of one variable calculus, where both the domain *and* codomain are  $\mathbb{R}$ . For instance, the Jacobian matrix for  $g(x) = x^2$  is the matrix containing a single entry, namely  $g'$ , the usual one variable derivative of  $g$ :

$$Jg = \begin{pmatrix} 2x \end{pmatrix}.$$

**2.5. Matrices and linear functions.** To each matrix with  $m$  rows and  $n$  columns we can associate a function  $L$  from  $\mathbb{R}^n$  to  $\mathbb{R}^m$ . If the  $i$ -th row of the matrix consists of the numbers  $a_1, \dots, a_n$ , then the  $i$ -th component function will have the form

$$L_i(x_1, \dots, x_n) = a_1x_1 + \dots + a_nx_n.$$

For instance, the matrix

$$\begin{pmatrix} 2 & 1 & 3 \\ 0 & -5 & 6 \end{pmatrix}$$

corresponds to the function

$$L(x, y, z) = (2x + y + 3z, -5y + 6z).$$

Note how the coefficients of the components of  $L$  come from the rows of the matrix. The choice of the name of the function,  $L$ , and the names of the variables,  $x$ ,  $y$ , and  $z$ , are not important. A function is called *linear* precisely when it comes from a matrix in this way. We'll talk about linear functions in much more depth later in the notes. For now, to get the hang of this, try matching up the matrices appearing below on the left with their corresponding linear functions on the right. The answers appear at the bottom of the page.<sup>‡</sup>

$$(a) \quad \begin{pmatrix} 2 & 3 \\ 1 & 2 \\ 4 & 7 \end{pmatrix} \quad (1) \quad \ell(q) = (q, 0, -3q)$$

$$(b) \quad \begin{pmatrix} 2 & 1 & 4 \\ 3 & 2 & 7 \end{pmatrix} \quad (2) \quad L(x, y) = (0, 0)$$

$$(c) \quad (1 \ 0 \ -3) \quad (3) \quad r(u, v) = (2u + 3v, u + 2v, 4u + 7v)$$

$$(d) \quad \begin{pmatrix} 1 \\ 0 \\ -3 \end{pmatrix} \quad (4) \quad L(x, y, z) = (2x + y + 4z, 3x + 2y + 7z)$$

$$(e) \quad \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \quad (5) \quad d(t) = 7t$$

$$(f) \quad (7) \quad (6) \quad M(s, t, u) = s - 3u$$

We can go backwards, too. Given a linear function, we can construct the corresponding matrix. So from now on, we will think of matrices and linear functions as being essentially the same thing.

<sup>‡</sup>(a) $\leftrightarrow$ (3), (b) $\leftrightarrow$ (4), (c) $\leftrightarrow$ (6), (d) $\leftrightarrow$ (1), (e) $\leftrightarrow$ (2), (f) $\leftrightarrow$ (5).

**2.6. The derivative.** We now have just enough machinery to describe the derivative of a multivariate function: it is the linear function associated with the function's Jacobian matrix. Let  $f$  be a function from  $\mathbb{R}^n$  to  $\mathbb{R}^m$ , and let  $p$  be a point of  $\mathbb{R}^n$ . The *derivative* of  $f$  at  $p$  is the linear function associated with  $Jf(p)$ ; it is denoted by  $Df_p$  and has domain  $\mathbb{R}^n$  and codomain  $\mathbb{R}^m$ . For example, let  $f(x, y) = (x^2 + 3y^2, 3x - y^4, x^4y)$  and let  $p = (1, 2)$ . First calculate the Jacobian matrix for  $f$ , then evaluate it at  $p$ :

$$Jf = \begin{pmatrix} 2x & 6y \\ 3 & -4y^3 \\ 4x^3y & x^4 \end{pmatrix}, \quad Jf(1, 2) = \begin{pmatrix} 2 & 12 \\ 3 & -32 \\ 8 & 1 \end{pmatrix}.$$

The derivative of  $f$  at  $p$  is the corresponding linear function

$$Df_{(1,2)}(x, y) = (2x + 12y, 3x - 32y, 8x + y).$$

Note that both  $f$  and its derivative at  $p$  have the same domain and codomain, which only makes sense since the point of taking the derivative is to get an uncomplicated function which can take the place of  $f$ , at least near  $p$ .

As another example, let  $g(t) = (t, t^2, t^3)$ . To calculate the derivative of  $g$  at  $t = 2$ , first calculate the Jacobian matrix of  $g$ , then evaluate it at 2:

$$Jg = \begin{pmatrix} 1 \\ 2t \\ 3t^2 \end{pmatrix} \quad Jg(2) = \begin{pmatrix} 1 \\ 4 \\ 12 \end{pmatrix}.$$

The derivative is the corresponding linear function

$$Dg_2(t) = (t, 4t, 12t).$$

Note that by convention we use the same names for the variables in the derivative as for the original function. Since  $g$  is a function of  $t$ , so is  $Dg_2$ . Also, be careful to evaluate the Jacobian matrix so that you have a matrix of numbers (not functions) before writing down the derivative.

In the case of an ordinary function from one variable calculus, one with domain and codomain both equal to  $\mathbb{R}$ , it seems that we now have two notions of the derivative. We used to think of the derivative as a number signifying a slope of a tangent line or a rate of change, and now we think of it as a linear function: on the one hand the derivative of  $g(x) = x^2$  at  $x = 1$  is  $g'(1) = 2$ , and on the other it is the linear function  $Dg_1(x) = 2x$ . The way to reconcile the difference is to remember that we have agreed to identify every linear function with a matrix; in this case,  $Dg_1$  is identified with the Jacobian matrix  $Jg(1) = (2)$ , and it is not such a big step to identify the matrix  $(2)$  with the number 2. The next chapter of these notes will help to reconcile the geometric meanings of the old and new notions of the derivative.

**2.7. The best affine approximation.** The statement that the derivative of a function is a good approximation of the function is itself only an approximation of the truth. For instance, the derivative of  $c(t) = (t, t^3)$  at  $t = 1$  is  $Dc_1(t) = (t, 3t)$ , and  $Dc_1$  is supposed to be a good approximation of  $c$  near the point  $t = 1$ . However  $c(1) = (1, 1)$ , and  $Dc_1(t)$  *never* equals  $(1, 1)$ . What is actually true is that the derivative is a good approximation if we are willing to add a translation and thus form what is known as the *best affine*<sup>§</sup> *approximation*. We now briefly explain how this is done. You are again asked to take a lot on faith. The

<sup>§</sup>The word "affine" is related to the word "affinity." The pronunciation is correct with the stress on either syllable.

immediate goal is to be able to blindly calculate the best affine approximation and get a vague idea of its purpose in preparation for later on.

First, a little algebra. *Addition* of points in  $\mathbb{R}^m$  is defined component-wise:

$$(x_1, x_2, \dots, x_m) + (y_1, y_2, \dots, y_m) := (x_1 + y_1, x_2 + y_2, \dots, x_m + y_m).$$

For example,  $(1, 2, 3) + (4, 5, 6) = (1 + 4, 2 + 5, 3 + 6) = (5, 7, 9)$ . We can also *scale* a point by a real number, again defined component-wise. If  $t \in \mathbb{R}$ , we scale a point by the factor of  $t$  as follows:

$$t(x_1, x_2, \dots, x_m) := (tx_1, tx_2, \dots, tx_m).$$

Hence, for example,  $3(1, 2, 3) = (3 \cdot 1, 3 \cdot 2, 3 \cdot 3) = (3, 6, 9)$ . The operations of addition and scaling can be combined:  $(1, 2) + 4(-1, 0) = (1, 2) + (-4, 0) = (-3, 2)$ . These algebraic operations will be studied in detail in Chapter 3<sup>¶</sup>.

Now, to form the best affine approximation of  $c$  at  $t = 1$ , take the derivative  $Dc_1(t) = (t, 3t)$  and make a new function, which we denote  $Ac_1$ , by adding the point of interest  $c(1) = (1, 1)$ :

$$Ac_1(t) := c(1) + Dc_1(t) = (1, 1) + (t, 3t) = (1 + t, 1 + 3t).$$

Note that  $Ac_1(0) = (1, 1)$ . It turns out that  $Ac_1$  near  $t = 0$  is a good approximation to  $c$  near  $t = 1$ .

Here is the general definition, using the shorthand  $x = (x_1, \dots, x_n)$  and  $p = (p_1, \dots, p_n)$ :

**Definition 2.1.** *Let  $f$  be a function from  $\mathbb{R}^n$  to  $\mathbb{R}^m$ , and let  $p$  be a point in  $\mathbb{R}^n$ . The best affine approximation to  $f$  at  $p$  is the function*

$$Af_p(x) := f(p) + Df_p(x).$$

One goal of these notes is to show that  $Af_p$  for values of  $x$  near the origin is a good approximation of  $f$  near  $p$ . For now, we are just trying to learn how to calculate  $Af_p$ .

As another example, let  $f(u, v) = (\cos(u), \sin(u), v^2)$ . We will calculate the best affine approximation to  $f$  at the point  $p = (0, 2)$ . The first step in just about every calculation in differential calculus is to find the Jacobian matrix:

$$Jf = \begin{pmatrix} -\sin(u) & 0 \\ \cos(u) & 0 \\ 0 & 2v \end{pmatrix}, \quad Jf(0, 2) = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 4 \end{pmatrix}$$

Hence, the derivative is

$$Df_{(0,2)}(u, v) = (0, u, 4v).$$

To get the best affine approximation, add  $f(p) = f(0, 2) = (1, 0, 4)$ :

$$Af_{(0,2)}(u, v) = f(0, 2) + Df_{(0,2)}(u, v) = (1, 0, 4) + (0, u, 4v) = (1, u, 4v + 4).$$

So  $Af_p(u, v) = (1, u, 4 + 4v)$ . The function  $Af_p$  evaluated near  $(0, 0)$  should be a good approximation to  $f$  evaluated near  $p = (0, 2)$ . We can at least see that  $Af_p(0, 0) = (1, 0, 4) = f(p)$ .

One slightly annoying property of  $Af_p$  is that its behavior near the *origin* is like  $f$ 's near  $p$ . By making a shift by  $-p$  in the domain before applying  $Af_p$ , we get a function

$$Tf_p(x) := Af_p(x - p) = f(p) + Df_p(x - p).$$

The behavior of  $Tf_p$  near  $p$  is like  $Af_p$ 's near the origin and hence like  $f$ 's near  $p$ . Thus, it is probably more accurate to say that  $Tf_p$ , rather than  $Df_p$  or  $Af_p$ , is a good approximation

<sup>¶</sup>Notation: We will often write  $A := B$  if  $A$  is defined by  $B$

to  $f$  at  $p$ . We will also call  $Tf_p$  the best affine approximation of  $f$  near  $p$ . Continuing the example from above, with  $f(u, v) = (\cos(u), \sin(u), v^2)$  and  $p = (0, 2)$ , we had  $Af_{(0,2)}(u, v) = (1, u, 4 + 4v)$ . Hence,

$$\begin{aligned} Tf_{(0,2)} &= Af_{(0,2)}((u, v) - (0, 2)) = Af_{(0,2)}(u, v - 2) \\ &= (1, u, 4 + 4(v - 2)) = (1, u, -4 + 4v). \end{aligned}$$

So  $Tf_p(u, v) = (1, u, -4 + 4v)$ . For instance, at the point  $p = (0, 2)$ , itself,  $Tf_p$  and  $f$  are an exact match:  $Tf_p(0, 2) = (1, 0, 4) = f(0, 2)$ . As a simpler example, the derivative of  $g(x) = x^2$  at  $x = 1$  is  $Dg_1(x) = 2x$ . Thus,  $Af_1(x) = g(1) + Dg_1(x) = 1 + 2x$ . Shifting, we get

$$Tf_1(x) = Af_1(x - 1) = 1 + 2(x - 1) = 2x - 1,$$

which is the tangent line with which we started this chapter.

We will continue to fudge and say that  $Df$  is a good approximation of  $f$  even though it is more accurate to refer to  $Af$  or to  $Tf$ .

### 3. Conclusion

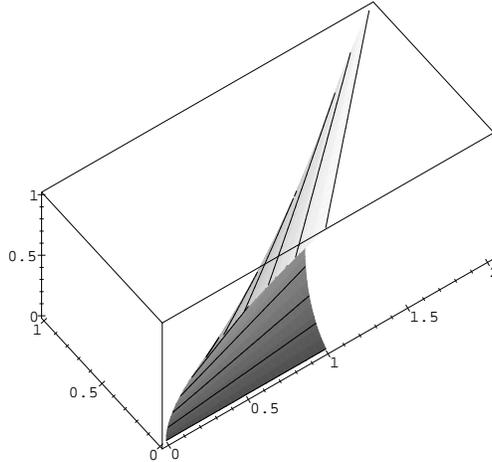
After reading this introductory chapter, you should be able to calculate the derivative and the best affine approximation of a multivariate function at a given point. From the example in the first section, you should have some idea of what is meant by an “approximation” of a function.

**3.1. To do.** It is clear that there is much to do before you can thoroughly understand what has already been presented. Here is a partial list of topics which we will need to cover later in the notes:

1. Give a precise definition of the derivative of a multivariate function.
2. From the definition of the derivative, explain exactly in what sense it provides a good approximation of a function. (This will actually follow fairly easily from the definition.)
3. Give the precise definition of a partial derivative.
4. Show that the derivative can be calculated as we have described here. (This will turn out to be a little tricky.)
5. Study algebra and geometry in  $\mathbb{R}^n$  including scaling, translations, and measurements of distances and angles.
6. Study general properties of linear functions.

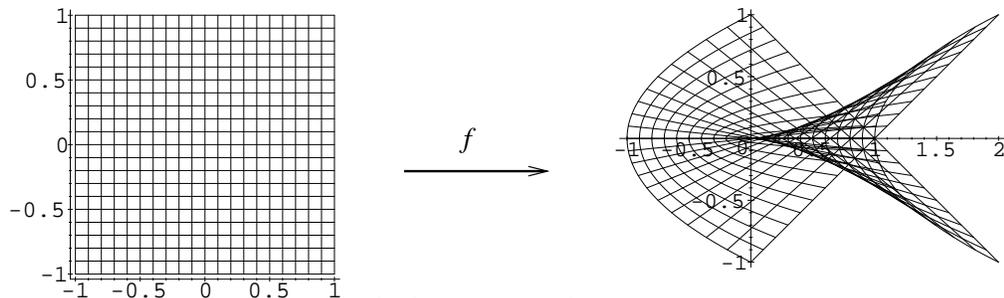
## EXERCISES

- (1) Consider the function  $\ell(u, v) = (u^2 + v, uv, u)$ . The image under  $\ell$  of the unit square with lower left corner at the origin is a surface in space:



How would you describe the geometric relation between  $\ell$  and the function from the first section of this chapter,  $f(u, v) = (u^2 + v, uv)$ ?

- (2) The function  $f(u, v) = (u^2 + v, uv)$ , from the first section, does not exactly fold the unit square along its diagonal. Which points of the unit square are sent by  $f$  to the upper boundary of the image?
- (3) Let  $f(u, v) = (u^2 + v, uv)$ , as before, but now let  $u$  and  $v$  vary between  $-1$  and  $1$ . Analyzing  $f$  as in the first section, we arrive at the following picture for  $f$ :



- (a) Describe the image under  $f$  of each side of the square.
- (b) Describe, roughly, what happens to the interior of the square.
- (4) The linear function given by differential calculus to approximate the function  $f(u, v) = (u^2 + v, uv)$  near the point  $(1/2, 1/2)$  is  $L(u, v) = (u + v, u/2 + v/2)$ .
- (a) Describe the image of a unit square under this new  $L$  as we did above when considering the point  $(1/4, 3/4)$  in the first section. What is strange?
- (b) What is it about  $f$  near the point  $(1/2, 1/2)$  that might account for the strange behavior just observed?
- (c) What linear function do you think will best approximate  $f$  near the origin, i.e., near  $(0, 0)$ ? (Try to do this geometrically, without formally calculating the derivative.)
- (5) For each of the following functions:
- calculate the Jacobian matrix;
  - evaluate the Jacobian matrix at the given point;

- 
- (iii) find the derivative at the given point;
  - (iv) find both versions of the best affine approximation, e.g.,  $Af_p$  and  $Tf_p$ , at the given point.
    - (a)  $f(u, v) = (u^2 + v, uv)$  at the point  $(\frac{1}{4}, \frac{3}{4})$ .
    - (b)  $f(x, y, z) = (3x^2 - 2y + z + 1, xy - 2z, 2x^2 - 5xz + 7, yz - 3z^3)$  at the point  $(-1, 2, 1)$ .
    - (c)  $p(r, \theta) = (r \cos(\theta), r \sin(\theta))$  at the point  $(1, \pi)$ .
    - (d)  $g(u, v, w) = uv + 5v^2w$  at the point  $(2, -3, 1)$ .
    - (e)  $r(t) = (\cos(t), \sin(t), t)$ , a helix, at the point  $t = 2\pi$ .
    - (f)  $f(x) = x^5$  at the point  $x = 1$ .



# Multivariate functions

Calculus is a tool for studying functions. In the first part of this chapter, we formally introduce the most basic vocabulary associated with functions. We then show how the functions with which we deal in these notes and their derivatives can be interpreted, i.e., what they mean and why they are useful. So at the beginning of the chapter, we do rigorous mathematics for the first time by giving a few precise definitions, but then quickly revert to the imprecise mode of the previous chapter. The hope is to give you a glimpse of a range of applications and provide a rough outline for several main results we must carefully consider later.

## 1. Function basics

We start with two sets,  $S$  and  $T$ . For now, these can be sets of anything—beanie babies<sup>TM</sup>, juggling balls—not necessarily numbers. We think of a function from  $S$  to  $T$  as a way to assign a *unique* element of  $T$  to *every* element of  $S$ . To be precise about this idea, we first define the *Cartesian product*,  $S \times T$ , of the sets  $S$  and  $T$  to be the collection of all ordered pairs  $(s, t)$  where  $s$  is an element of  $S$  and  $t$  is an element of  $T$ :

$$S \times T := \{(s, t) \mid s \in S, t \in T\}.$$

Let's pause to talk about notation. Whenever we write something of the form  $A := B$  in these notes, using a colon and an equals sign, it will mean that  $A$  is *defined to be*  $B$ . This is different from a statement such as  $2(3 + 5) = 2 \cdot 3 + 2 \cdot 5$  which asserts that the object on the left is the same as the object on the right by consequence of previous definitions or axioms, in this case the distributive law for integers. So the symbol “:=” means “is defined to be.” We'll sometimes use it in reverse:  $A =: B$  means  $B$  is defined to be  $A$ .

Thus, back to the definition of the Cartesian product, we have defined  $S \times T$  to be  $\{(s, t) \mid s \in S, t \in T\}$ . This means that  $S \times T$  is the set of all objects of the form  $(s, t)$ , where  $s$  is an element of  $S$  and  $t$  is an element of  $T$ . The bar “|” can be translated as “such that” and the symbol “ $\in$ ” means “is an element of.” You will see many constructions of this form, namely,  $\{A \mid B\}$ , which can always be read as “the set of all objects of the form  $A$  such that the  $B$  holds” ( $B$  will be some list of restrictions). For a quick review of set notation, please see Appendix A.

The object  $(s, t)$  is an *ordered pair*, i.e., a list consisting of  $s$  and  $t$  in which order matters:  $(s, t)$  not the same as  $(t, s)$  unless  $s = t$ . An example: if  $S := \{1, 2, 3\}$  and  $T = \{\clubsuit, \heartsuit\}$ , then  $S \times T = \{(1, \clubsuit), (2, \clubsuit), (3, \clubsuit), (1, \heartsuit), (2, \heartsuit), (3, \heartsuit)\}$ . We can define the Cartesian product of more than two sets; given three sets,  $S$ ,  $T$ , and  $U$ , we define  $S \times T \times U$  to be ordered lists of three elements, the first from  $S$ , the second from  $T$ , and the last from  $U$ , and so on. For instance, Euclidean  $n$ -space,  $\mathbb{R}^n$ , is the Cartesian product of  $\mathbb{R}$  with itself  $n$ -times.

We are now ready to give the formal definition of a function.

**Definition 1.1.** *Let  $S$  and  $T$  be sets. A function  $f$  with domain  $S$  and codomain  $T$  is a subset  $\Gamma_f$  of  $S \times T$  such that for each  $s \in S$  there is a unique  $t \in T$  such that  $(s, t) \in \Gamma_f$ . If  $(s, t) \in \Gamma_f$ , then we write  $f(s) = t$ .*

We write  $f: S \rightarrow T$  to denote a function with domain  $S$  and codomain  $T$  and say that  $f$  is a function from  $S$  to  $T$ . If  $f(s) = t$ , we say that  $f$  sends  $s$  to  $t$ . Almost all of the functions with which we deal will be given by some specific rule such as  $f(x) = x^2$ , and in that case, it is easier to say “consider the function  $f(x) = x^2$ ” than “consider the function  $f$  consisting of pairs  $(x, x^2)$ .” In fact, normally we will not think of a function in terms of this formal definition at all; rather, we will think of a function as parametrizing a surface or as being a vector field, etc. In order to draw attention to the set  $\Gamma_f$ , we call  $\Gamma_f$  the *graph* of  $f$ , although, precisely speaking, it *is* the function  $f$ .

The following illustrates a common way of defining a function:

$$\begin{aligned} f: \mathbb{R}_{\geq 0} \times \mathbb{R} &\rightarrow \mathbb{R}^2 \\ (x, y) &\mapsto (x + y^2, xy) \end{aligned}$$

This defines  $f$  to be a function with domain  $\mathbb{R}_{\geq 0} \times \mathbb{R}$  and codomain  $\mathbb{R}^2$ . The symbol  $\mathbb{R}_{\geq 0}$  denotes the set of non-negative real numbers, and the symbol  $\mapsto$  indicates what  $f$  does to a typical point. Hence,  $f(x, y) = (x + y^2, xy)$ , and  $x$  is restricted to be a non-negative real number.

For future reference, the following definition summarizes some of the most basic vocabulary dealing with functions.

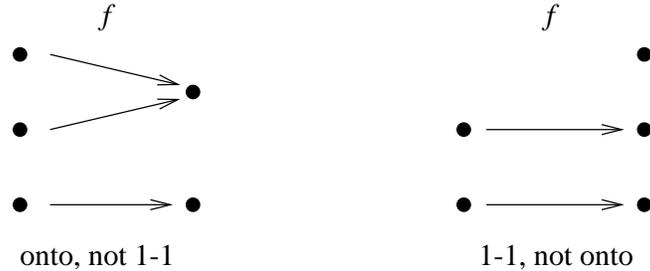
**Definition 1.2.** *Let  $f: S \rightarrow T$  be a function.*

1. *The function  $f$  is one-to-one (1–1) or injective if  $f(x) = f(y)$  only when  $x = y$ .*
2. *The image or range of  $f$  is  $\{f(s) \in T \mid s \in S\}$ . The image will be denoted by  $\text{im}(f)$  or  $f(S)$ .*
3. *The function  $f$  is onto or surjective if  $\text{im}(f) = T$ .*
4. *If  $f$  is both 1–1 and onto, then  $f$  is called a 1–1 correspondence between  $S$  and  $T$  or a bijection.*
5. *The inverse image of  $t \in T$  is  $f^{-1}(t) := \{s \in S \mid f(s) = t\}$ . If  $X \subseteq T$ , the inverse image of  $X$  is  $f^{-1}(X) := \{s \in S \mid f(s) \in X\}$ .*
6. *If the image of one function is contained in the domain of another, then we can compose the functions. To be precise, if  $g: T' \rightarrow U$  is a function such that  $\text{im}(f) \subseteq T'$ , then the composition of  $f$  and  $g$  is the function  $g \circ f: S \rightarrow U$  given by  $(g \circ f)(s) := g(f(s))$ .\**

---

\*Notation: if  $A$  and  $B$  are sets, then  $A \subseteq B$  means that  $A$  is a subset of  $B$ , i.e., every element of  $A$  is an element of  $B$ ; we write  $A \subset B$  if  $A$  is *proper subset* of  $B$ , i.e.,  $A$  is a subset of  $B$ , not equal to  $B$ . Outside of these notes, you may see the symbol “ $\subset$ ” used to mean what we reserve for “ $\subseteq$ .”

Thus,  $f$  is 1-1 if no two elements of  $S$  are sent by  $f$  to the same element of  $T$ ; the image of  $f$  is the set of all points in  $T$  that are hit by  $f$ ;  $f$  is onto if every element of  $T$  is hit by  $f$ ; and the inverse image of  $t$  is the set of all elements of  $S$  that are sent to  $t$ .



The function  $f: \mathbb{R} \rightarrow \mathbb{R}$  defined by  $f(x) = x^2$  is not 1-1 since, for instance,  $f(1) = f(-1) = 1$ ; it is not onto since no negative numbers are in the image; the inverse image of 4 is  $f^{-1}(4) = \{2, -2\}$ . On the other hand, the function  $g: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$  defined by  $g(x) = x^2$  is 1-1 and onto, i.e., a bijection. The function  $g$  differs from  $f$  because it has a different domain and codomain. The inverse of 4 under  $g$  is  $g^{-1}(4) = \{2\}$ .

## 2. Interpreting functions

From now on, we will deal almost exclusively with functions for which the domain and codomain are both subsets of Euclidean spaces. We may not explicitly specify the domain and codomain when they are unimportant or clear from context. For example, the function  $f(x, y) = (x^2, x + 2xy^2, y^3)$  may be assumed to have domain  $\mathbb{R}^2$  and codomain  $\mathbb{R}^3$  whereas the function  $g(x, y) = x/y$  has (largest possible) domain the plane excluding the  $x$ -axis,  $\{(x, y) \in \mathbb{R}^2 \mid y \neq 0\}$ , and codomain  $\mathbb{R}$ . To simplify notation we will often consider functions of the form  $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$ , leaving the reader to make needed modifications (if any) in the case of a function such as  $g$  just above whose domain is a proper subset of  $\mathbb{R}^n$ .

Let  $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$  be a function. For the rest of the chapter, we consider various interpretations of  $f$ , depending on the values of  $n$  and  $m$ :

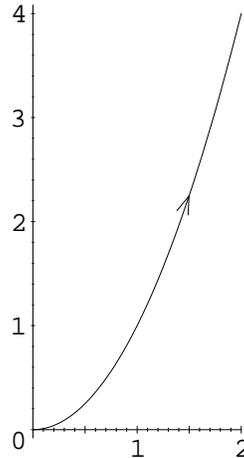
- If  $n \leq m$ , then it is natural to think of  $f$  as parametrizing an  $n$ -dimensional surface in  $m$ -dimensional space.
- If  $m = 1$ , then  $f$  assigns a number to each point in  $\mathbb{R}^n$ ; the number could signify the temperature, density, or height of each point, for example.
- If  $n = m$ , then we may think of  $f$  as a vector field: at each point in space, the function assigns a vector. These types of functions are used to model electric and gravitational fields, flows of fluids, etc.

These points of view are considered separately below.

**2.1. Parametrizations:**  $n \leq m$ . The function  $f$  takes each point in  $\mathbb{R}^n$  and assigns it a place in  $\mathbb{R}^m$ . Thus, we may think of forming the image of  $f$  by imagining that  $\mathbb{R}^n$  is a piece of putty which  $f$  twists and stretches and places in  $\mathbb{R}^m$  (recall the first example of these notes). From this perspective, we say that  $f$  is a *parametrization* of its image. In other words, to say that  $f$  *parametrizes* a blob  $S$  sitting in  $\mathbb{R}^m$  means that  $S = \text{im}(f)$ . If you are asked to give a parametrization of a subset  $S \subset \mathbb{R}^m$ , you are being asked to find a function whose image is  $S$ . We now look at some low-dimensional cases.

2.1.1. *Parametrized curves:  $n = 1$ .* A *parametrized curve in  $\mathbb{R}^m$*  is a function of the form  $f: \mathbb{R} \rightarrow \mathbb{R}^m$ . For example, if  $f(t) = (t, t^2)$ , the image of  $f$  is the set of points  $\{(x, y) \in \mathbb{R}^2 \mid y = x^2\}$ . So  $f$  parametrizes a parabola. The function  $g(t) = (t^3, t^6)$  parametrizes the same parabola. To see this, take any point  $(x, y)$  such that  $y = x^2$ , i.e., a point of the form  $(x, x^2)$ . Define  $a = \sqrt[3]{x}$ ; then  $g(a) = (x, y)$ . Conversely, it is clear that any point of the form  $(t^3, t^6)$  lies on the parabola. Hence, the image of  $g$  is the parabola, as claimed. (Why doesn't this same argument work for the function  $h(t) = (t^2, t^4)$ ? What do you get in that case?)

What is the difference between these two parametrizations of the same parabola? Think of  $t$  as time, and think of  $f$  (respectively,  $g$ ) as describing the motion of a particle in the plane. At time  $t$ , the particle is at the point  $f(t)$  (respectively,  $g(t)$ ). As time goes from  $t = 0$  to  $t = 2$ , the particle described by  $f$  has moved along the parabola from  $(0, 0)$  to  $(2, 4)$ .

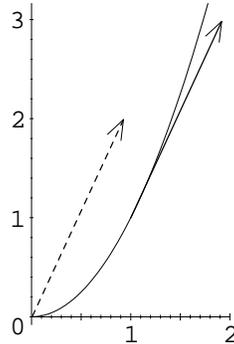


On the other hand, the particle described by  $g$  has moved along the parabola from  $(0, 0)$  to  $(8, 64)$ . So as time goes on, both particles sweep out a parabola, but they are *moving at different speeds*. This points out the essential difference between a parabola and a *parametrized* parabola. A parabola is just a set of points in the plane, whereas a parametrized parabola is a set of points that is explicitly described as being swept out by the motion of a particle.

What does differential calculus do for us in this situation? The Jacobian matrix for  $f$  is

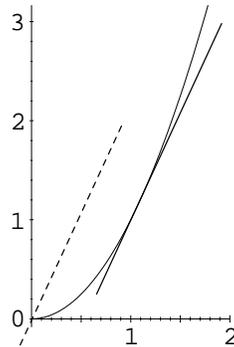
$$Jf = \begin{pmatrix} 1 \\ 2t \end{pmatrix}$$

Define  $f'(t) := (1, 2t)$ ; so  $f'$  is the vector formed by the single column of the Jacobian matrix of  $f$ . One of the tasks which we will take up later is to show that  $f'$  is the *velocity* of the particle described by  $f$  at time  $t$ . For instance, at time  $t = 1$ , the particle is at  $(1, 1)$  moving with velocity  $f'(1) = (1, 2)$ :



Although the vector  $(1, 2)$  is normally thought of as an arrow with tail at the origin,  $(0, 0)$ , and head at the point  $(1, 2)$ , in this case, it makes sense to translate the vector out to the point  $(1, 1)$ , where the action is taking place.

The derivative for  $f$  at  $t = 1$  is the linear function corresponding to  $Jf_1$ , namely,  $Df_1(t) = (t, 2t)$ . This is a (parametrized) line through the origin and passing through the point  $(1, 2)$ , i.e., a line pointing in the same direction as the velocity vector. The best affine approximation to  $f$  at  $t = 1$  parametrizes this same line but translated out so that it passes through  $f(1)$ :



Thus, the best affine approximation is a parametrization of the tangent line to  $f$  at  $t = 1$ .

What we have just seen about the derivative holds for all parametrized curves.

**Definition 2.1.** Let  $f: \mathbb{R} \rightarrow \mathbb{R}^m$  be a parametrized curve. The tangent vector or velocity for  $f$  at time  $t = a$  is the vector

$$\text{velocity}(f)_{t=a} := f'(a) = (f'_1(a), f'_2(a), \dots, f'_m(a)),$$

i.e., the single column of the Jacobian matrix,  $Jf_a$ , considered as a vector (as usual,  $f_i$  is the  $i$ -th component function of  $f$ ). The speed of  $f$  at time  $a$  is the length of the velocity vector:

$$\text{speed}(f)_{t=a} := |\text{velocity}(f)_{t=a}| := \sqrt{f'_1(a)^2 + f'_2(a)^2 + \dots + f'_m(a)^2}.$$

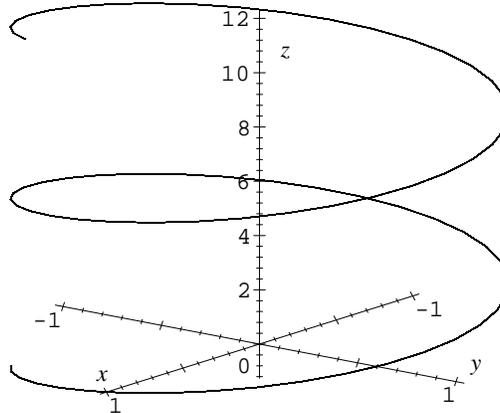
The tangent line to  $f$  at time  $a$  is the line parametrized by the best affine approximation:

$$\begin{aligned} Af_a(t) &:= f(a) + f'(a)t \\ &:= (f_1(a) + f'_1(a)t, f_2(a) + f'_2(a)t, \dots, f_m(a) + f'_m(a)t). \end{aligned}$$

Note the definition of *speed* as the length of the velocity vector, in turn defined to be the square root of the sums of the squares of the components of the velocity vector. For

instance, the speed of  $f(t) = (t, t^2)$  at  $t = 1$  is  $|(1, 2)| = \sqrt{1 + 4} = \sqrt{5}$  while the speed of  $g(t) = (t^3, t^6)$  at  $t = 1$  is  $|(3, 6)| = 3\sqrt{5}$  since  $g'(t) = (3t^2, 6t^5)$ .

We now look at a curve in  $\mathbb{R}^3$ . Let  $h(t) = (\cos(t), \sin(t), t)$ . It describes a particle moving along a helix in  $x, y, z$ -space:



If we *project* onto the first two coordinates, we get the parametrized circle,  $t \mapsto (\cos(t), \sin(t))$ . As time moves forward, the particle described by  $h$  spins around in a circle, increasing its height linearly with  $t$ . The velocity vector at time  $t$  is  $h'(t) = (-\sin(t), \cos(t), 1)$  and the speed is its length,  $\sqrt{\sin^2(t) + \cos^2(t) + 1} = \sqrt{2}$ . The speed is constant; it does not depend on  $t$ . The equation for the tangent line at time  $t = a$  is

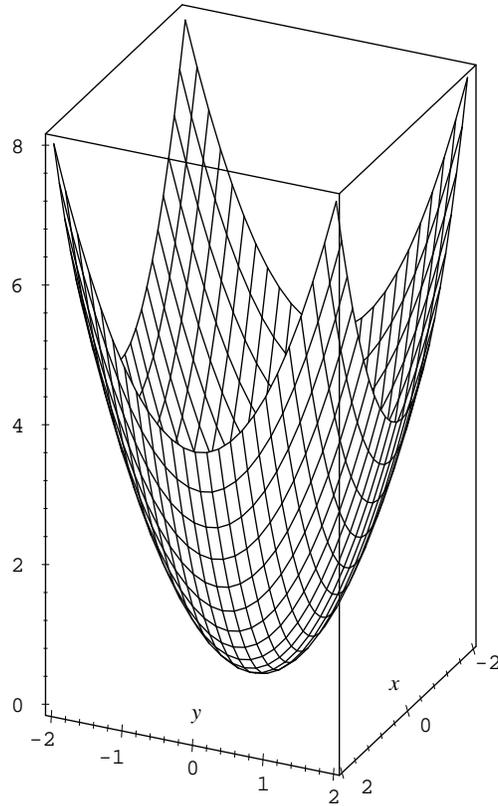
$$Ah_a(t) = (\cos(a) - \sin(a)t, \sin(a) + \cos(a)t, a + t).$$

We are thinking of  $a$  as fixed, and as the parameter  $t$  varies,  $Ah_a(t)$  sweeps out the tangent line. For example, at time  $a = 0$ , the tangent line is parametrized by  $Ah_0(t) = (1, t, t)$ . Try to picture this line attached to the helix. At time  $t = 0$ , it passes through the point  $Ah_0(0) = h(0) = (1, 0, 0)$  with a velocity vector  $(0, 1, 1)$ . The line sits in a plane parallel to the  $y, z$ -plane with an upwards slope of  $45^\circ$ .

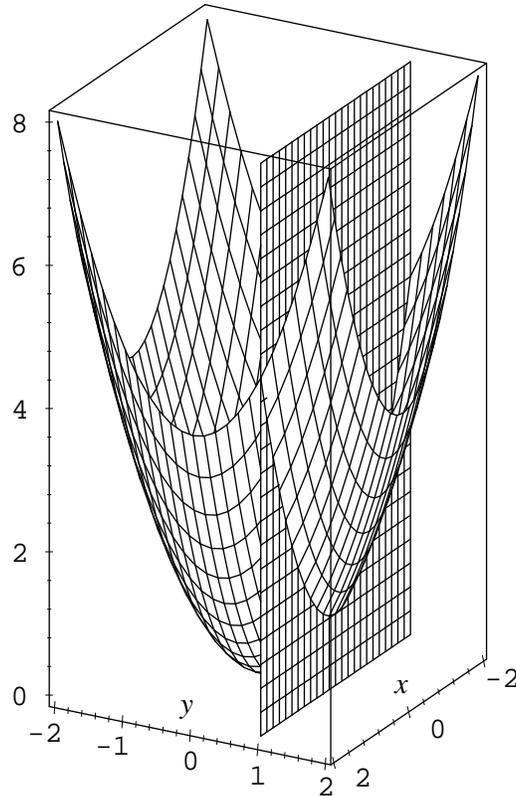
We can write down parametrizations for curves in higher dimensions, for example,  $c(t) = (t, t^2, t^3, t^4)$  is a curve in  $\mathbb{R}^4$ . It gets harder to picture these curves. In four dimensions, you might code the fourth coordinate by using color (try this on a computer). In general, when a mathematician thinks of a curve in dimension higher than three, the vague mental picture is probably of a curve in 3-space.

Note that by our definition, the function  $f(t) = (0, 0)$  qualifies as a parametrized curve even though it just sits at the origin, never moving. In given situations, one might want to require a parametrized curve to be 1–1, at least most of the time.

**2.1.2. Parametrized surfaces:  $n = 2$ .** A *parametrized surface* in  $\mathbb{R}^m$  is a function of the form  $f: \mathbb{R}^2 \rightarrow \mathbb{R}^m$ . For example,  $f(u, v) = (u, v, u^2 + v^2)$  is a parametrized surface in  $\mathbb{R}^3$ . It turns out to be a paraboloid. The image under  $f$  of a square grid centered at the origin in  $\mathbb{R}^2$  is pictured below:



Think of the paraboloid as  $\mathbb{R}^2$  warped and placed into space by  $f$ . To get an idea of how  $f$  is stretching the plane, consider what happens to lines in the plane. For instance, take a line of the form  $v = a$  where  $a$  is a constant. This is a line parallel to the  $u$  axis. Let  $x$ ,  $y$ , and  $z$  denote the coordinates in  $\mathbb{R}^3$ . Plugging our line into  $f$ , we get  $f(u, a) = (u, a, u^2 + a^2)$ , which describes the parabola which is the intersection of the paraboloid with the plane  $y = a$ .



By symmetry, the same kind of thing happens to lines parallel to the  $v$  axis. How does  $f$  transform circles in the plane? Fix a constant  $r$  and consider points in the plane of the form  $(r \cos(t), r \sin(t))$  as  $t$  varies, i.e., the circle of radius  $r$  centered at the origin. Plugging this circle into  $f$  we get  $f(r \cos(t), r \sin(t)) = (r \cos(t), r \sin(t), r^2)$  (using the fact that  $\cos^2(t) + \sin^2(t) = 1$ ). As  $t$  varies, we get the circle which is the intersection of the plane  $z = r^2$  and the paraboloid. So  $f$  is taking concentric circles in the plane, and lifting them to a height equal to the square of their radii. This gives a pretty good picture of how  $f$  turns the plane into a paraboloid in  $\mathbb{R}^3$ .

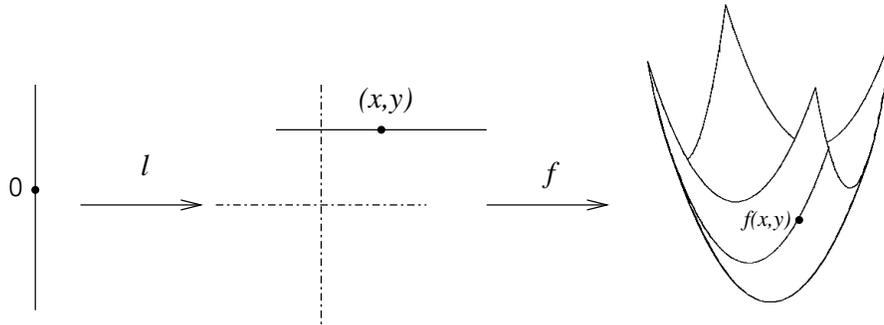
What about derivatives? The Jacobian matrix for  $f$  is

$$Jf = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 2x & 2y \end{pmatrix}$$

It now has two columns: the vectors  $\partial f/\partial x$  and  $\partial f/\partial y$ . To get at the geometry of these vectors, fix the point  $(x, y)$  and consider a line passing through this point, parallel to the  $x$ -axis:  $\ell(t) = (x + t, y)$ . Composing with  $f$  defines a curve:

$$\begin{aligned} c &= f \circ \ell: \mathbb{R} \rightarrow \mathbb{R}^3 \\ t &\mapsto (x + t, y, (x + t)^2 + y^2) \end{aligned}$$

In other words,  $c(t) = f(\ell(t)) = f(x + t, y) = (x + t, y, (x + t)^2 + y^2)$ . The domain of  $c$  is  $\mathbb{R}$ , so it really is a curve (remembering that we have fixed  $(x, y)$ ). We say  $c$  is a *curve on the surface*  $f$ . Explanation: the function  $\ell$  is a curve in  $\mathbb{R}^2$ ; it takes the real number line and places it in the plane. Composing with  $f$  takes this line and puts it on the paraboloid:



To generalize, if  $g: \mathbb{R}^2 \rightarrow \mathbb{R}^m$  is any surface, and  $e: \mathbb{R} \rightarrow \mathbb{R}^2$  is any curve in  $\mathbb{R}^2$ , then the composition,  $g \circ e$  can be thought of as a curve lying on the surface  $g$ . Hence, in this situation, we have a geometric interpretation of the composition of functions.

The tangent to our curve  $c$  at time  $t = 0$  is

$$c'(0) = (1, 0, 2x) = \frac{\partial f}{\partial x}.$$

Thus, to understand the first column of the Jacobian matrix,  $Jf$ , first get in your car and drive parallel to the  $x$  axis in the plane at unit speed. Press the special button labeled  $f$  on your dashboard that magically takes the whole plane, including you, and warps it into  $\mathbb{R}^3$  as a paraboloid. Your path in the plane,  $\ell$ , is transformed into a path on the paraboloid,  $c$ . As you pass through the fixed point  $(x, y)$ , your velocity vector on the paraboloid,  $c'(0)$ , is the first column of  $Jf$ . Similarly, the second column of  $Jf$  comes from taking a path parallel to the  $y$  axis. This whole idea generalizes: the  $k$ -th column of the Jacobian matrix for any function (no matter what the domain and codomain) is the tangent of the curve formed by taking a path parallel to the  $k$ -th coordinate axis and composing with the function.

The two columns of the Jacobian matrix are thus two vectors, tangents to curves on the surface. These two vectors determine a plane—we will usually say that they “span” a plane—which is called the tangent plane to the surface  $f$  at the point  $(x, y)$ . Of course, this plane passes through the origin, and we will want to think of it as translated out to the point in question on the surface,  $f(x, y)$ . It turns out that the derivative function  $Df$  parametrizes the plane spanned by the two tangent vectors which are the columns of the Jacobian matrix. The best affine approximation,  $Af$ , parametrizes this plane, transformed out to the point  $f(x, y)$ . We will take these as the definitions; the geometry will become clearer after the following chapter on linear algebra.

**Definition 2.2.** Let  $f: \mathbb{R}^2 \rightarrow \mathbb{R}^m$  be a surface in  $\mathbb{R}^m$ . The tangent space to  $f$  at a point  $p \in \mathbb{R}^m$  is the plane spanned by the columns of the Jacobian matrix  $Jf(p)$ ; it is parametrized by the derivative  $Df_p$ . The (affine) tangent plane to  $f$  at  $p$  is the plane parametrized by the best affine approximation,  $Af_p$ ; it is the translation of the tangent space out to the point  $f(p)$ .

For example, let’s calculate the tangent plane to our paraboloid  $f$  at the point  $p = (1, 2)$ . The Jacobian at that point is

$$Jf(1, 2) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 2 & 4 \end{pmatrix}$$

Hence the tangent space is spanned by the vectors  $(1, 0, 2)$  and  $(0, 1, 4)$ ; it is parametrized by the derivative:

$$Df_{(1,2)}(x, y) = x(1, 0, 2) + y(0, 1, 4) = (x, y, 2x + 4y).$$

The translation out to  $f(1, 2) = (1, 2, 5)$  is given by the best affine approximation:

$$Af_{(1,2)}(x, y) = (1, 2, 5) + Df_{(1,2)}(x, y) = (1 + x, 2 + y, 5 + 2x + 4y).$$

As in the case of curves, degeneracies can occur. The function  $f(x, y) = (0, 0)$  qualifies as parametrized surface, as does  $f(x, y) = (x, 0)$ , both of which don't look much like what we would want to call surfaces. Again, we may want to add qualifications in the future forcing our functions to be 1–1 most of the time. Even if the function is 1–1, however, there may be places where the two tangent vectors we have discussed point do not span a plane, they can even both be zero vectors. For instance, consider the surface  $f(x, y) = (x^3, y^3, x^3y^3)$ . Even though this function is 1–1, the Jacobian matrix at the origin consists entirely of zeroes (check!). Hence, the tangent space there consists of a single point,  $(0, 0, 0)$ . Such points are especially interesting; they are called *singularities*.

As with curves, there is nothing stopping us from considering surfaces in dimensions higher than three, and in fact they often come up in applications. As far as picturing the geometry, the example of a surface in  $\mathbb{R}^3$  will usually serve as a good guide. For a surface in four dimensions, we may code the last component using color if we want to draw a picture, i.e., the last component can be used to specify the color to paint the point specified by the first three components. Try it on a computer.

2.1.3. *Parametrized solids:  $n = 3$ .* Consider the function

$$\begin{aligned} f: \mathbb{R}^2 \times [0, 1] &\rightarrow \mathbb{R}^3 \\ (x, y, r) &\mapsto (rx, ry, x^2 + y^2) \end{aligned}$$

where  $x, y \in \mathbb{R}$  and  $0 \leq r \leq 1$ . If  $r = 1$ , we get the paraboloid that we just considered. As  $r$  shrinks to zero, the point  $(rx, ry, x^2 + y^2)$  stays at the same height,  $x^2 + y^2$ , but moves radially in towards a central axis. Thus,  $f$  maps a slab, one unit high, to a *solid* paraboloid, i.e.,  $f$  parametrizes the paraboloid and all of its “interior” points. Fixing  $r$ , the function  $f$  maps a slice of this slab, a plane, onto a paraboloid. As  $r$  shrinks, these paraboloids get thinner until  $r = 0$ , at which point we get  $f(x, y, 0) = (0, 0, x^2 + y^2)$ ; then, a whole plane is sent to half of a line (which we may image as an extremely skinny paraboloid).

To understand the derivative in this situation, recall the first example in the previous chapter. There, we had a function mapping a square in  $\mathbb{R}^2$  into  $\mathbb{R}^2$ . If we took a small square about a point in the domain, the image under the function was a warped parallelogram. The derivative mapped that same square to a true parallelogram closely approximating the warped image of the function. A similar thing happens in our present situation. Consider a small three-dimensional box about a point in the domain. The image of this box under  $f$  will be a warped, tilted, box. The image of a box under the derivative of  $f$  will be a tilted box but with straight sides. If the original box is small enough, the image under  $f$  and under its derivative will closely resemble each other; the derivative just straightens things out.

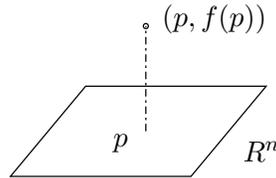
We could consider “solids” in higher dimensions and more general functions of the form  $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$  with  $n \leq m$ , but we have probably gone far enough for now. For these higher-dimensional parametrizations, a secure understanding of a surface in  $\mathbb{R}^3$  may be the best guide.

**2.2. Graphs:**  $m = 1$ . We now come to a second main class of functions. These are functions of the form  $f: \mathbb{R}^n \rightarrow \mathbb{R}$ , having codomain equal to  $\mathbb{R}$ . In the previous section, when  $n \leq m$ , we thought of a function as being characterized mainly by its image. That no longer makes much sense since now our function is squeezing all of  $\mathbb{R}^n$ —think of  $n$  as being at least 2—down into one dimension. Just looking at the image would throw out a lot of information.

The function  $f$  associates a single number with each point in  $\mathbb{R}^n$ . In practice,  $f$  may be specifying the temperature for each point in space or the density at each point of a solid, etc. We will call such functions, *real-valued functions*; they are sometimes called *scalar fields*. To picture  $f$ , we look at its graph, i.e., the set of points

$$\Gamma_f := \{(p, f(p)) \in \mathbb{R}^{n+1} \mid p \in \mathbb{R}^n\}.$$

To imagine the graph, think of  $\mathbb{R}^n$  lying horizontally (imagine  $n = 2$ ), and think of the last coordinate in the graph as specifying height:



Consider the function  $f(x, y) = x^2 + y^2$  with its graph  $\Gamma_f = \{(x, y, x^2 + y^2) \mid (x, y) \in \mathbb{R}^2\}$ . Hence, the graph of  $f$  is the paraboloid we spoke about above. However, note the difference. Before, we had a parametrized paraboloid: it was the image of a function. Many different functions can have the same image. Now it is the graph of a function. The graph contains more information; in fact, we can completely reconstruct a function from its graph (cf. Definition 1.1). For example, since  $(2, -1, 5)$  is in the graph of  $f$ , we know that  $f(2, -1) = 5$ .

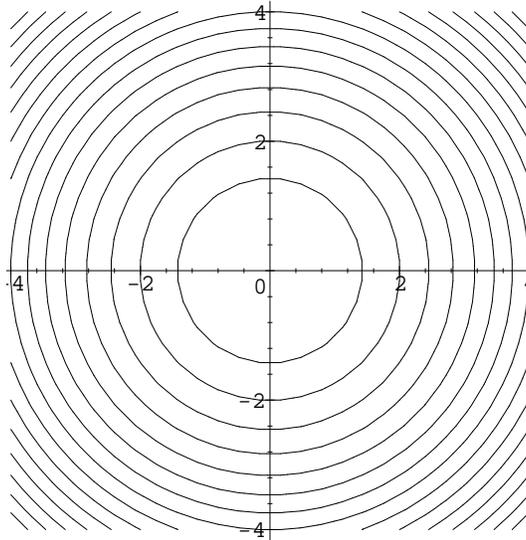
Anyone who has looked at a topographical map knows another way of picturing a real-valued function. A topographical map depicts a three-dimensional landscape using just two dimensions, by drawing *contours*. A contour is the set of points on the map that represent points that are all at the same height on the landscape. By drawing a separate contour for, say, each 10 feet change in elevation, we get a good idea of the corresponding landscape. If consecutive contours are close at a certain region on the map, that means the landscape is steep there. Where the contours are far apart, the landscape is fairly flat. We now generalize this idea to take care of real-valued functions in general.

**Definition 2.3.** Let  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  be a real-valued function. A contour or level set of  $f$  at height  $a \in \mathbb{R}$  is the set of points

$$f^{-1}(a) := \{x \in \mathbb{R}^n \mid f(x) = a\}.$$

A collection of level sets of  $f$  is called a contour diagram or topographical map of  $f$ .

Thus, a contour of  $f$  is the inverse image of a point,  $f^{-1}(a)$ , or equivalently, the set of solutions  $x$  to an equation of the form  $f(x) = a$ . Continuing our example from above, the solutions  $(x, y)$  to  $f(x, y) = x^2 + y^2 = a$  form a circle of radius  $\sqrt{a}$  for each  $a$ . (As special cases, we get the origin when  $a = 0$  and we get the empty set when  $a < 0$ .) Here is a contour diagram for  $f$ ; you can tell there is a deep hole at the origin:



The Jacobian matrix for  $f$  consists of a single row (contrasting with the case of a parametrized curve whose Jacobian matrix consists of a single column):

$$Jf = ( 2x \ 2y )$$

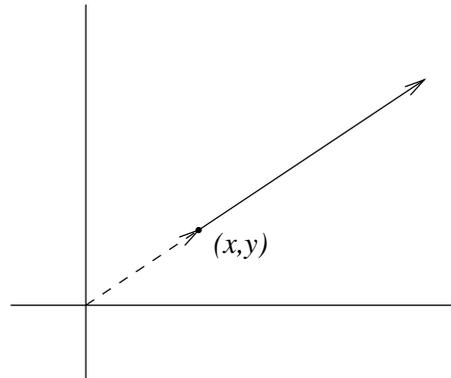
The single row of the Jacobian is often thought of as a vector; it is called the *gradient* of  $f$  and denoted  $\nabla f$ . Thus,  $\nabla f = (2x, 2y)$ . In fact  $\nabla f$  is a function: as  $x$  and  $y$  change, so does the value of  $\nabla f$ . So we could write  $\nabla f(x, y) = (2x, 2y)$ , or making the domain and codomain explicit:

$$\begin{aligned} \nabla f: \mathbb{R}^2 &\rightarrow \mathbb{R}^2 \\ (x, y) &\mapsto (2x, 2y) \end{aligned}$$

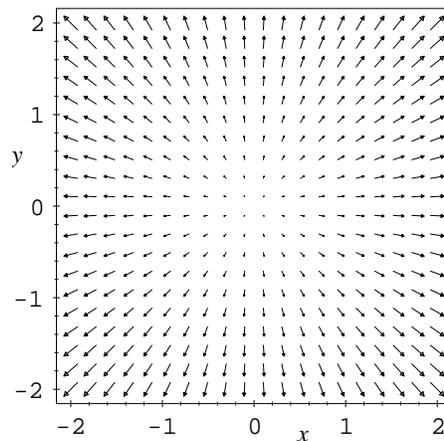
Thus, for example,  $\nabla f(1, 3) = (2, 6)$ .

To see the geometrical significance of the gradient, think of the function  $f$  as reporting the amount of profit obtained by adjusting certain production levels,  $x$  and  $y$ , of two commodities. For example, producing 2 units of the first commodity and 5 of the second yields a profit of  $f(2, 5) = \$29$ . You are a greedy two-dimensional being walking around in the plane. Your position  $(x, y)$  will determine production levels, yielding a profit of  $f(x, y) = x^2 + y^2$ . Of course, you want to maximize profits, so you want to walk in the direction that determines production levels which increase profits most dramatically. Since you are only a two-dimensional being, you cannot see the big picture, the paraboloid, which would make it obvious in which direction you must go, namely, straight away from the origin so that profits are pushed up the uphill. However, you are equipped with a magical watch. The watch has only one hand, and it grows and shrinks in length and changes direction as you walk. This one hand is the gradient vector of  $f$ , having its tail at the center of the watch, which you might think of as the origin (you could also think of this vector as having its tail at your current position). The secret behind your watch is that it always points in the direction you must go in order to increase profits most quickly. The length of the hand on your watch indicates how quickly profits will increase if you head in that direction.

If you are standing at a specific point  $(x, y)$  and look at your watch, you will see the vector  $\nabla f = (2x, 2y)$ . Think of the vector that starts at the origin of the plane and goes out to where you are standing. On your watch, you have a vector that continues in this direction and has length twice that of your distance to the origin:



If we go around to lots of points in the plane and draw the vector represented by the hand on our watch, we get a picture like this:



Thus, it is clear what you'll do to increase profits most quickly. Wherever you stand, you must walk out radially from the origin. The rate of change will be two times your current distance from the origin, the length of the gradient vector.

We are ready for the formal definition.

**Definition 2.4.** Let  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  be a real-valued function. The gradient of  $f$  is the single row of the Jacobian matrix for  $f$ , considered as a vector in  $\mathbb{R}^n$ :

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

Note that  $\nabla f$  is a vector that sits in  $\mathbb{R}^n$ , the domain of  $f$ . One of the main tasks of these notes is to explain the following facts about the gradient: (i) the gradient points in the direction of quickest increase of the function; (ii) the length of the gradient gives the rate of increase of the function in the direction of its quickest increase; and (iii) the gradient is perpendicular to its corresponding level set.

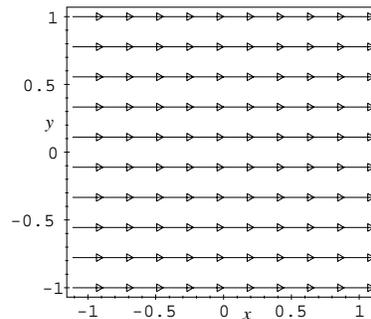
**2.3. Vector fields:**  $n = m$ . Imagine a flow, say water running down a stream or air circulating in a room. At each point there is a particle moving with a certain velocity. We will now consider functions that can be used to glue velocity vectors to points in space in order to model this behavior.

There is technically no difference between a point in  $\mathbb{R}^n$  and a vector in  $\mathbb{R}^n$ . The  $n$ -tuple  $(x_1, \dots, x_n) \in \mathbb{R}^n$  can be thought of in two ways: it can specify a point that is a distance  $x_1$  along the first axis,  $x_2$  along the second axis, and so on, or it can be thought of as a vector, an arrow with its tail at the origin and head at the point in question. The trick behind modeling a flow with a function is to use both of these interpretations at once. Here is the definition:

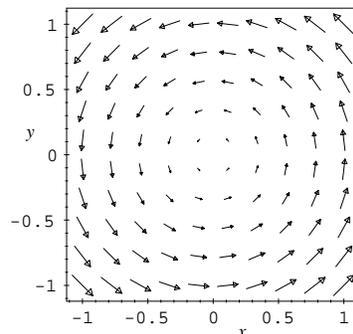
**Definition 2.5.** A vector field in  $\mathbb{R}^n$  is a function having both domain and codomain equal to  $\mathbb{R}^n$ , i.e., a function of the form  $F: \mathbb{R}^n \rightarrow \mathbb{R}^n$ .

For each point  $p \in \mathbb{R}^n$ , the function associates a vector  $v := F(p) \in \mathbb{R}^n$ . To picture the function at a point  $p$ , take the corresponding vector  $F(p)$ , which officially has its tail at the origin in  $\mathbb{R}^n$ , and translate it out so that its tail is sitting at  $p$ . In this way, we think of  $F$  as gluing a vector to each point in space, and we can imagine the corresponding flow of particles. When  $n = 2$  or  $n = 3$ , one typically draws these vectors for lots of different choices of  $p$ , and a pattern develops that lets you visualize the flow.

**Example 2.6.** The vector field  $F(x, y) = (1, 0)$  is a constant vector field. It models a uniform flow in the plane.



**Example 2.7.** The function  $F(x, y) = (-y, x)$  describes the following vector field:



The vector  $(-y, x)$  is perpendicular to the line segment going out from the origin to the point  $(x, y)$ , and both the vector and the line segment have the same length (cf. Chapter 3).

**Example 2.8.** If  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  is a real-valued function, its gradient,  $\nabla f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ , is a vector field on  $\mathbb{R}^n$ . In this situation, the function  $f$  is called a *potential function* for the vector field  $\nabla f$ . Continuing a previous example, if  $f(x, y) = x^2 + y^2$ , then  $\nabla f(x, y) = (2x, 2y)$ . We drew a picture of this vector field on page 29.

When thinking about vector fields, questions easily arise which will take you outside of the realm of differential calculus. For one, does every vector field have a potential function, i.e., is every vector field actually a gradient vector field? This is an important question which you'll take up in a course on integral calculus. In integral calculus, you'll also learn how to calculate the amount of stuff flowing through a given surface. Another naturally occurring question is: can I determine the path a particle would take when flowing according to a given vector field? That turns out to be a whole subject in itself: differential equations.

### 3. Conclusion

In this chapter, we have formally defined functions and introduced some of the basic terminology for describing them. We have also just begun to consider the various ways in which multivariate functions can be used, e.g., to parametrize surfaces, to describe temperature distributions, or to model fluid flow. We will take a unified and coherent approach to the subject. For instance, things as disparate as tangent vectors and gradients will both be considered as instances of the same thing: the derivative of a multivariate function.

Your interpretation of a function will depend on the context. A function may be open to multiple interpretations. For example, if  $f: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ , you might think of  $f$  as a parametrized surface or as a vector field. In the very first example in these notes, we considered the plane as a piece of putty that gets folded and stretched according to instructions encoded by  $f$ . On the other hand, in the present chapter, we looked at a function of the same form as a gradient vector field, showing us which direction to go in order to maximize another function most quickly. Ironically, the type of function studied in one variable calculus, having the form  $f: \mathbb{R} \rightarrow \mathbb{R}$ , is even more ambiguous; it can be thought of under each of the three basic categories of functions introduced in this chapter. For instance,  $f(x) = x^2$  can be thought of as parametrizing a curve in  $\mathbb{R}$ ; it describes the motion of a particle which at time  $x$  is at position  $x^2$  along the real number line. Secondly, since  $f$  is a real-valued function, we can think of it as telling us the temperature of the number line at each point; the graph of position vs. temperature is a parabola. Finally, you could interpret  $f$  as a vector field describing the flow of a particle which moves with velocity  $x^2$  along the number line when it is at position  $x$ .

**3.1. To do.** In this chapter, we have added to our list of items which you are asked to take on faith for now but which need to be explained in the rest of the notes.

- (1) In the case of a parametrized curve, why is the single column of the Jacobian matrix a good definition for velocity? More generally, why can the columns of the Jacobian matrix of a parametrized surface be thought of as tangent vectors?
- (2) Explain the basic facts about the gradient of a real-valued function:
  - The gradient points in the direction of quickest increase of the function.
  - The length of the gradient gives the rate of increase of the function in the direction of its quickest increase.
  - The gradient is perpendicular to its corresponding level set.

Again, the idea behind introducing these concepts early is to advertise up front what differential calculus is about so that you can see the motivation for the rest of the notes. By the time we get to the explanations, you'll be ready to hear them.

## EXERCISES

- (1) Let  $X = \{1, 2, 3, 4\}$  and  $Y = \{a, b\}$ . Write out all elements of the Cartesian product,  $X \times Y$ .
- (2) Describe the image of each of the following functions:
  - (a)  $f(x) = \cos(x)$ .
  - (b)  $g(u, v) = (u^2, v, 0)$ .
  - (c)  $h(x, y) = x^2 - y^2$  where both  $x$  and  $y$  are restricted to lie in the interval  $[0, 1]$ , i.e.,  $0 \leq x \leq 1$  and  $0 \leq y \leq 1$ .
- (3) For each of the following functions, state whether the function is 1–1 and whether it is onto. If it is not 1–1, provide two explicit distinct points that get mapped to the same point; if it is not onto, exhibit a point in the codomain that is not in the image of the function.
  - (a)  $f: \mathbb{R} \rightarrow \mathbb{R}$  defined by  $f(x) = \cos(x)$ .
  - (b)  $f: \mathbb{R} \rightarrow [-1, 1]$  defined by  $f(x) = \cos(x)$  where  $[-1, 1] := \{x \in \mathbb{R} \mid -1 \leq x \leq 1\}$ .
  - (c)  $f: [0, \pi) \rightarrow [-1, 1]$  defined by  $f(x) = \cos(x)$  where  $[0, \pi) := \{x \in \mathbb{R} \mid 0 \leq x < \pi\}$ .
  - (d)  $g: \mathbb{R}^2 \rightarrow \mathbb{R}^3$  defined by  $g(u, v) = (u, v, uv)$ .
  - (e)  $g: \mathbb{R}^2 \rightarrow \mathbb{R}^3$  defined by  $g(u, v) = (u^2, v, uv)$ .
- (4) Let  $f(x, y) = x - y$ . What is  $f^{-1}(0)$ , the inverse image of 0?
- (5) Let  $f(x, y) = (x + y, 2y)$  and  $g(u, v) = (u^2, u + v)$ . What is the composition,  $g \circ f$ ? (Be careful about the order.)
- (6) We have discussed two different parametrizations of the parabola:  $f(t) = (t, t^2)$  and  $g(t) = (t^3, t^6)$ .
  - (a) When is the speed of  $f$  greater than the speed of  $g$ ? When is it less than the speed of  $g$ ? When are the speeds equal?
  - (b) Give a parametrization of this same parabola which describes the motion of a particle moving backwards, i.e., in the opposite direction to that given by  $f$ .
- (7) Let  $c(t) = (t, t^3)$ . Suppose that  $c$  denotes the position of a particle in the plane at time  $t$ .
  - (a) Draw a picture of the image of  $c$ .
  - (b) What is the velocity of the particle when  $t = 2$ ?
  - (c) What is the speed of the particle when  $t = 2$ ?
  - (d) Give an equation parametrizing the tangent line (the best affine approximation) at time  $t = 2$ , and draw this tangent line in your picture of  $c$ .
- (8) Sketch the curve  $c(t) = (t, t^n)$  for  $n = 1, 2, 3, 4, 5, 6$ . Describe the basic behavior for general  $n$ .
- (9) Let  $f(t) = (t^2, t^3)$ .
  - (a) Draw a picture of the image of  $f$ .
  - (b) Find the speed of  $f$  at an arbitrary time  $t = a$ . What happens when  $a = 0$ ? Use this information to give a rough description of the motion of a particle whose position at time  $t$  is  $f(t)$ .
- (10) Consider the curve  $c(t) = (t^2 - 1, t(t^2 - 1))$ .
  - (a) Show that every point  $(x, y)$  in the image of  $c$  satisfies the equation  $y^2 = x^3 + x^2$ . (For extra credit, prove the converse: every point  $(x, y)$  satisfying  $y^2 = x^3 + x^2$  is in the image of  $c$ .)

- (b) If  $c$  passes through the point  $(x, y)$  when  $t = a$ , at what time does  $c$  pass through the point  $(x, -y)$ ? This shows that the image of  $c$  is symmetric about the  $x$ -axis.
  - (c) Name the three times  $c$  passes through the  $x$ -axis, i.e., when it passes through points of the form  $(x, 0)$ .
  - (d) Sketch the image of  $c$ .
  - (e) Find parametric equations for the tangent lines (the best affine approximation) at the two different times  $c$  passes through the origin. Draw these two tangent lines in your picture of  $c$ .
- (11) Give a parametric equation for the tangent line to the curve  $c(t) = (t, t^2, t^3, t^4)$  when  $t = 1$ .
- (12) Let  $f(x, t) = (t, tx^2)$ . We could think of  $f$  as a parametrized surface or as a vector field. Another way to think about  $f$  is as a curve, parametrized by  $x$  which evolves over time,  $t$ . Draw the parametrized curves represented by  $f(x, -2)$ ,  $f(x, -1)$ ,  $f(x, 0)$ ,  $f(x, 1)$ , and  $f(x, 2)$ .
- (13) Consider the parametrized surface  $f(u, v) = (u, v, u^2 + v^2)$  in  $\mathbb{R}^3$ . Let  $x$ ,  $y$ , and  $z$  denote the coordinates in  $\mathbb{R}^3$ . We want to show that the tangent plane to  $f$  at the origin can be thought of as the collection of all tangent vectors to curves on  $f$  passing through the origin.
- (a) Calculate the parametrization of the tangent plane (the best affine approximation) to  $f$  at  $(0, 0)$ . You should get a parametrization of the  $x, y$ -plane.
  - (b) Let  $c(t)$  be any curve in  $\mathbb{R}^2$  passing through the origin at time  $t = 0$ . So  $c(t) = (c_1(t), c_2(t))$  and  $c(0) = (0, 0)$ ; hence,  $c_1(0) = c_2(0) = 0$ . Let  $e := f \circ c$ . Thus  $e$  is a parametrized curve on  $f$ , passing through the point  $f(0, 0) = (0, 0, 0)$  at time  $t = 0$ . Calculate the velocity vector of  $e$  when  $t = 0$ , and show that it lies in the tangent plane to  $f$  at the origin.
  - (c) Conversely, given any vector  $(a, b, 0) \in \mathbb{R}^3$ , i.e., in the tangent plane to  $f$  at the origin, explicitly find an example of a curve  $c$  in the plane, such that  $c(0) = (0, 0)$  and such that if  $e := f \circ c$ , then  $e'(0) = (a, b, 0)$ .
- (14) Let  $f(u, v) = (u^2, u + v, v^2, uv^3, u^2 - 3v)$  be a surface in  $\mathbb{R}^5$ . Calculate the tangent plane to  $f$  at the point  $(2, -1)$ .
- (15) Let  $f(x, y) = xy$ .
- (a) Make a contour map for  $f$  by drawing the contours,  $f = a$ , for  $a = -3, -2, -1, 0, 1, 2, 3$ .
  - (b) Calculate the gradient of  $f$  at the point  $(1, 1)$ , and draw it on your contour diagram.
  - (c) Make a sketch of the graph of  $f$ . Your contour diagram should help.
- (16) Let  $f(x, y) = x^2y$ .
- (a) Make a contour map for  $f$  by drawing the contours,  $f = a$ , for  $a = -3, -2, -1, 0, 1, 2, 3$ .
  - (b) Calculate the gradient of  $f$  at the point  $(1, 1)$ , and draw it on your contour diagram.
  - (c) Make a sketch of the graph of  $f$ . Your contour diagram should help.
- (17) Why can't a point be on two different contours for the same real-valued function?
- (18) Let  $F(x, y) = (1, 2y)$  be a vector field.
- (a) Draw a picture of the vector field  $F$ .
  - (b) Find two different potential functions for  $F$ .

- (19) Draw the vector field  $F(x, y) = (y, x)$ . Your drawing should make the general behavior of  $F$  reasonably clear.
- (20) The vector fields we have described in these notes are static, they do not evolve over time. To model a vector field that evolves over time, we could use a function of the form  $F: \mathbb{R}^{n+1} \rightarrow \mathbb{R}^n$  and consider the last coordinate to be time. For example, let  $F(x, y, t) = (tx, ty)$ . Draw the vector fields represented by  $F(x, y, -2)$ ,  $F(x, y, -1)$ ,  $F(x, y, 0)$ ,  $F(x, y, 1)$ , and  $F(x, y, 2)$ . Imagine these vector fields as they continuously evolve.

# Linear algebra

We have already defined Euclidean  $n$ -space to be the Cartesian product of  $\mathbb{R}$  with itself  $n$ -times:

$$\begin{aligned}\mathbb{R}^n &:= \{(x_1, \dots, x_n) \mid x_i \in \mathbb{R} \text{ for } i = 1, \dots, n\} \\ &= \underbrace{\mathbb{R} \times \dots \times \mathbb{R}}_{n \text{ times}}.\end{aligned}$$

The elements of  $\mathbb{R}^n$  are called *points* or *vectors*. The real number  $x_i$  is called the  $i$ -th *component* or  $i$ -th *coordinate* of  $(x_1, \dots, x_n)$ . In this chapter, we will put what are called a *linear structure* and a *metric* on  $\mathbb{R}^n$ , and study the functions between Euclidean spaces that “preserve” linear structures, namely, linear functions.

## 1. Linear structure

The linear structure on  $\mathbb{R}^n$  consists of a rule for adding points and for scaling by a real number. We did this on page 12, but will repeat it here for convenience. Let  $x = (x_1, \dots, x_n)$  and  $y = (y_1, \dots, y_n)$  be two points in  $\mathbb{R}^n$ . Their *sum* is defined component-wise:

$$x + y = (x_1, \dots, x_n) + (y_1, \dots, y_n) := (x_1 + y_1, \dots, x_n + y_n).$$

If  $t \in \mathbb{R}$ , define the *scalar multiplication* of  $x = (x_1, \dots, x_n)$  by  $t$  component-wise:

$$tx = t(x_1, \dots, x_n) := (tx_1, \dots, tx_n).$$

Thus, for instance,

$$(1, -2, 0, 4) + 4(2, 7, 3, 1) = (1, -2, 0, 4) + (8, 28, 12, 4) = (9, 26, 12, 8).$$

We will never add points that lie in spaces of different dimensions, e.g., we will never try to add  $(1, 2)$  and  $(3, 4, 5, 6)$ .

Addition is a function  $+: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ . This merely says that addition associates an element of  $\mathbb{R}^n$  with each pair of elements in  $\mathbb{R}^n$ . Similarly, scalar multiplication is a function  $\cdot: \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ ; it takes a scalar (a real number) and an element of  $\mathbb{R}^n$  and returns an element of  $\mathbb{R}^n$ .

Define the *origin* in  $\mathbb{R}^n$  to be the vector  $\vec{0} = (0, \dots, 0)$ . If  $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ , define  $-x := (-x_1, \dots, -x_n)$ . Thus, for example,  $-(1, 2, 3) = (-1, -2, -3)$ . The following proposition summarizes some basic properties of addition and scalar multiplication:

**Proposition 1.1.** *Let  $x, y, z \in \mathbb{R}^n$ , and let  $s, t \in \mathbb{R}$ ; then*

1.  $x + y = y + x$  (commutativity of vector addition);
2.  $(x + y) + z = x + (y + z)$  (associativity of vector addition);
3.  $\vec{0} + x = x + \vec{0} = x$  ( $\vec{0}$  is the additive identity);
4.  $x + (-x) = (-x) + x = \vec{0}$  ( $-x$  is the additive inverse of  $x$ );
5.  $1x = x$  and  $(-1)x = -x$ ;
6.  $(st)x = s(tx)$  (associativity of scalar multiplication);
7.  $(s + t)x = sx + tx$  (distributivity);
8.  $s(x + y) = sx + sy$  (distributivity).

PROOF: We will prove a couple of items and leave the rest as exercises. They all consist of reducing to the corresponding result for real numbers, which we take as known (cf. Appendix B). For the first item,

$$\begin{aligned} x + y &= (x_1, \dots, x_n) + (y_1, \dots, y_n) \\ &= (x_1 + y_1, \dots, x_n + y_n) \\ &= (y_1 + x_1, \dots, y_n + x_n) \\ &= y + x. \end{aligned}$$

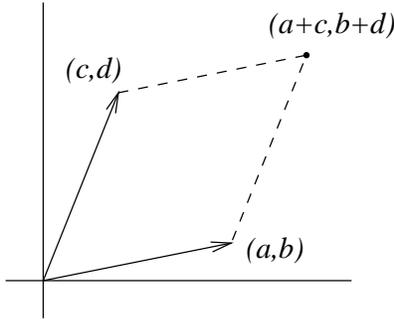
The only step which was not purely definitional was the third; there, we used commutativity of addition for real numbers,  $x_i + y_i = y_i + x_i$ .

For the last item,

$$\begin{aligned} s(x + y) &= s((x_1, \dots, x_n) + (y_1, \dots, y_n)) \\ &= s(x_1 + y_1, \dots, x_n + y_n) \\ &= (s(x_1 + y_1), \dots, s(x_n + y_n)) \\ &= (sx_1 + sy_1, \dots, sx_n + sy_n) \\ &= (sx_1, \dots, sx_n) + (sy_1, \dots, sy_n) \\ &= s(x_1, \dots, x_n) + s(y_1, \dots, y_n) \\ &= sx + sy. \end{aligned}$$

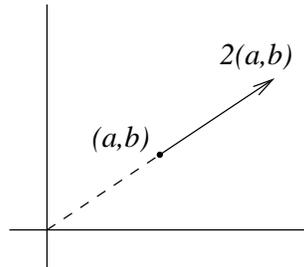
The main step is on the fourth line. It uses the fact that for real numbers, multiplication distributes over addition.  $\square$

Vector addition and scalar multiplication have geometric interpretations. It is essential to understand these interpretations in order to understand the geometry discussed throughout these notes. You are probably familiar with the *parallelogram rule* for addition of points in the plane. To visualize the addition of points  $(a, b)$  and  $(c, d)$  in  $\mathbb{R}^2$ , draw edges from the origin out to these two points. These give to edges for a unique parallelogram whose final vertex is the sum  $(a, b) + (c, d) = (a + c, b + d)$ . Another way to picture this is to imagine that the vector  $(a, b)$  has been translated out by the vector  $(c, d)$ , or vice-versa.



We will think of addition in higher dimensions the same way. In fact, any two vectors that we might want to add will actually sit in a plane, and we can think of using the ordinary parallelogram rule in that plane (imagine this in  $\mathbb{R}^3$ ). If  $x, y \in \mathbb{R}^n$ , we define the *translation of  $x$  by  $y$*  to be the sum  $x + y$ . Of course, by commutativity of addition, this is the same as the translation of  $y$  by  $x$ .

Multiplying a vector by a scalar  $t \in \mathbb{R}$  turns out to yield a vector pointing in the same direction but whose length has been changed by a factor of  $t$ . This is easy to see in examples in  $\mathbb{R}^2$ :



The proof of this fact in all dimensions appears below in the next section.

## 2. Metric structure

Measurements in  $\mathbb{R}^n$  are all based on the following simple algebraic device.

**Definition 2.1.** *The dot product of  $x, y \in \mathbb{R}^n$  (also called the scalar product or inner product) is*

$$\langle x, y \rangle := x \cdot y := \sum_{i=1}^n x_i y_i.$$

The dot product is a function from  $\mathbb{R}^n \times \mathbb{R}^n$  to  $\mathbb{R}$ . For example,

$$(3, -2, 5, 1) \cdot (-1, 0, 3, 2) = 3(-1) + (-2)0 + 5 \cdot 3 + 1 \cdot 2 = 14.$$

We interchangeably use the notation  $\langle (3, -2, 5, 1), (-1, 0, 3, 2) \rangle = 14$ . The algebra defining the dot product is simple, but it will take a while to understand the underlying geometry. The following basic properties are easily verified.

**Proposition 2.2.** *Let  $x, y, z \in \mathbb{R}^n$  and  $s \in \mathbb{R}$ ; then*

1.  $x \cdot y = y \cdot x$ ;
2.  $x \cdot (y + z) = x \cdot y + x \cdot z$ ;
3.  $(sx) \cdot y = x \cdot (sy) = s(x \cdot y)$ ;

4.  $x \cdot x \geq 0$ ;  
 5.  $x \cdot x = 0$  if and only if  $x = \vec{0}$ .

PROOF: The first item is proved using commutativity of multiplication for real numbers:

$$x \cdot y = \sum_{i=1}^n x_i y_i = \sum_{i=1}^n y_i x_i = y \cdot x.$$

Proofs for the rest of the items are left as exercises.  $\square$

We sum up this proposition by saying that the dot product is a symmetric, bilinear, positive, definite form. Symmetry is item 1; linearity is items 2 and 3; positivity is item 4; and definiteness is item 5.

We use the inner product to define length:

**Definition 2.3.** The length of  $x \in \mathbb{R}^n$  (also called the norm or absolute value) is

$$|x| := \sqrt{x \cdot x} = \sqrt{\sum_{i=1}^n x_i^2}.$$

This definition generalizes the usual notion of length in  $\mathbb{R}^2$  and  $\mathbb{R}^3$ : it is the square root of the sum of the squares of the coordinates. Thus,

$$|(5, 3, -2, 1, 2)| = \sqrt{5^2 + 3^2 + (-2)^2 + 1^2 + 2^2} = \sqrt{43}.$$

In dimension one, for  $x \in \mathbb{R}$ , we have  $|x| = \sqrt{x^2}$ , which agrees with the ordinary notion of absolute value.

**2.1. Perpendicularity, components, and projections.** If you would like to look ahead, our next goal is Proposition 2.8, especially a proof of the triangle inequality. Before taking that task on, it is best to consider when two vectors are perpendicular and about how to find the component of one vector along another.

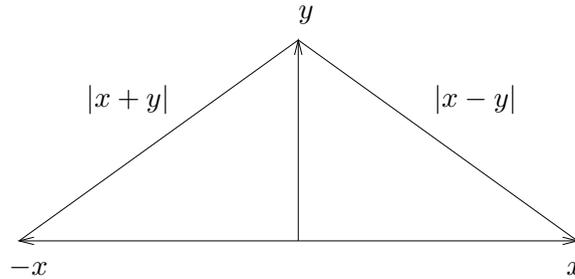
The notion of perpendicularity has a simple formulation in terms of the dot product:

**Definition 2.4.** The vectors  $x, y \in \mathbb{R}^n$  are perpendicular or orthogonal if  $x \cdot y = 0$ .

For instance, since  $(2, -1) \cdot (1, 2) = 2(1) + (-1)2 = 0$ , the vectors  $(2, -1)$  and  $(1, 2)$  are perpendicular. The vectors  $(1, 2, 3)$  and  $(2, 1, 0)$  are not perpendicular:  $(1, 2, 3) \cdot (2, 1, 0) = 4 \neq 0$ . The zero vector,  $\vec{0}$ , is perpendicular to every vector.

Note that no proof of this characterization of perpendicularity is required. We have simply *defined* two vectors to be perpendicular if their dot product is zero. You may have some prior notion of perpendicularity in the case of  $\mathbb{R}^2$  or  $\mathbb{R}^3$ , and in that case, you could try to show that the two notions, the new and the old, are equivalent. We can, however, argue that our definition is reasonable, at least if you feel that our definition of length is reasonable. (In any case, our argument should resolve any doubts you may have in  $\mathbb{R}^2$  and  $\mathbb{R}^3$ .)

Let  $x, y \in \mathbb{R}^n$ . By the parallelogram rule for vector addition, the following picture shows that it is reasonable to say that  $x$  and  $y$  are perpendicular if and only if  $|x + y| = |x - y|$ :



Now,

$$\begin{aligned}
 |x + y| = |x - y| &\iff |x + y|^2 = |x - y|^2 \\
 &\iff (x + y) \cdot (x + y) = (x - y) \cdot (x - y) \\
 &\iff x^2 + 2x \cdot y + y^2 = x^2 - 2x \cdot y + y^2 \\
 &\iff 2x \cdot y = -2x \cdot y \\
 &\iff 4x \cdot y = 0 \\
 &\iff x \cdot y = 0.
 \end{aligned}$$

Let's pause for a few remarks about notation. First, the symbol " $\iff$ " means "if and only if." It does *not* mean "equal." If you use this symbol in your own writing, make sure that you can directly substitute the words "if and only if" wherever  $\iff$  appears. Second,  $x^2$  in the above context is shorthand for  $x \cdot x$ . Now back to the proof: the first line follows from the fact that two non-negative real numbers are equal if and only if their squares are equal. The second line follows directly from the definition of the norm: for any vector  $z$ , we have  $|z|^2 = z \cdot z$ . The rest follows from the basic properties of the dot product (cf. Proposition 1.1).

We can now prove a generalization of the Pythagorean theorem:

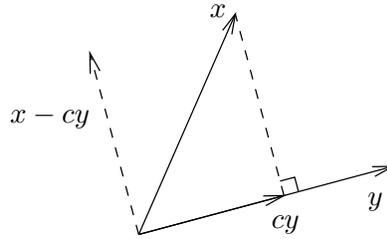
**Theorem 2.5. (Pythagorean Theorem)** *Let  $x, y \in \mathbb{R}^n$  be perpendicular; then*

$$|x|^2 + |y|^2 = |x + y|^2.$$

PROOF: From the basic properties of the dot product, using the fact that  $x \cdot y = 0$ , it follows that  $|x + y|^2 = (x + y) \cdot (x + y) = x^2 + 2x \cdot y + y^2 = x^2 + y^2 = |x|^2 + |y|^2$ .  $\square$

This really is the Pythagorean theorem. Draw a triangle having one edge equal to  $x$  and hypotenuse equal to the vector  $x + y$ . The remaining edge of the triangle is the vector  $y$  translated out to the tip of the vector  $x$ . These edges fit together to form a triangle by the parallelogram rule for vector addition.

Suppose we are given two vectors  $x, y \in \mathbb{R}^n$ . A useful geometric operation is to break  $x$  into two parts, one of which lies along the vector  $y$ . Given any real number  $c$ , the vector  $cy$  lies along  $y$  and we can evidently write  $x$  as the sum of two vectors:  $x = (x - cy) + cy$ . In addition, though, we would like to require, by adjusting  $c$ , that the vector  $x - cy$  is perpendicular to  $y$ :



Thus, we will have decomposed  $x$  into a vector that is perpendicular to  $y$  and a vector that runs along  $y$ . Let's calculate the required  $c$ :

$$(x - cy) \cdot y = 0 \iff x \cdot y - cy \cdot y = 0 \iff c = \frac{x \cdot y}{y \cdot y} = \frac{x \cdot y}{|y|^2}.$$

**Definition 2.6.** Let  $x, y \in \mathbb{R}^n$  with  $y \neq \vec{0}$ . The component of  $x$  along  $y$  is the real number

$$c := \frac{x \cdot y}{y \cdot y} = \frac{x \cdot y}{|y|^2}.$$

The projection of  $x$  along  $y$  is the vector  $cy$ .

It is probably difficult to remember, but we take the word “component” to mean the real number  $c$  and the word “projection” to refer to the corresponding vector  $cy$ .

**Example 2.7.** Let  $x = (2, 4)$  and  $y = (5, 1)$ . The component of  $x$  along  $y$  is

$$c = \frac{(2, 4) \cdot (5, 1)}{(5, 1) \cdot (5, 1)} = \frac{14}{26} = \frac{7}{13},$$

so the projection of  $x$  along  $y$  is

$$cy = \frac{7}{13}(5, 1) = \left( \frac{35}{13}, \frac{7}{13} \right).$$

You may want to draw a picture to convince yourself that the calculation has worked. Note that the component of  $x$  along  $y$  and the projection will be different from the component of  $y$  along  $x$  and the corresponding projection (try it for this example!).

For a higher-dimensional example, let  $u = (1, 0, 3, -2)$  and  $v = (2, 1, 1, 2)$ . The component of  $u$  along  $v$  is

$$c = \frac{(1, 0, 3, -2) \cdot (2, 1, 1, 2)}{(2, 1, 1, 2) \cdot (2, 1, 1, 2)} = \frac{1}{10},$$

and the projection of  $u$  along  $v$  is

$$cv = \frac{1}{10}(2, 1, 1, 2) = (0.2, 0.1, 0.1, 0.2).$$

**2.2. Cauchy-Schwarz, distance, and angles.** The basic properties of the norm are summarized below. They show that the norm is a reasonable measure of length. The tricky result is the Cauchy-Schwarz inequality, from which the triangle inequality easily follows.

**Proposition 2.8.** Let  $x, y \in \mathbb{R}^n$  and  $s \in \mathbb{R}$ ; then

1.  $|x| \geq 0$  (positive);
2.  $|x| = 0$  if and only if  $x = 0$  (definite);
3.  $|sx| = |s||x|$ ;
4.  $|x \cdot y| \leq |x||y|$  (Cauchy-Schwarz inequality);
5.  $|x + y| \leq |x| + |y|$  (triangle inequality).

PROOF: The first three items follow directly from the definition of the norm and are left as exercises. To prove the Cauchy-Schwarz inequality, first note that it is clearly true if  $y = \vec{0}$ . So suppose  $y \neq \vec{0}$ , and let  $c$  be the component of  $x$  along  $y$ . By construction,  $x - cy$  is perpendicular to  $y$ , hence to  $cy$ , so we can apply the Pythagorean theorem:

$$|x - cy|^2 + |cy|^2 = |(x - cy) + cy|^2 = |x|^2.$$

Since  $|x - cy|^2 \geq 0$  by item 1, we can throw it out of the above equation to get the inequality

$$|cy|^2 \leq |x|^2$$

Hence, writing this inequality in reverse and taking square roots,

$$|x| \geq |cy| = |c||y| = \left| \frac{x \cdot y}{y \cdot y} \right| |y| = \frac{|x \cdot y|}{|y|^2} |y| = \frac{|x \cdot y|}{|y|}.$$

Multiplying through by  $|y|$  yields Cauchy-Schwarz.

The triangle inequality is an easy consequence of the Cauchy-Schwarz inequality:

$$\begin{aligned} |x + y|^2 &= (x + y) \cdot (x + y) \\ &= x \cdot x + 2x \cdot y + y \cdot y = |x|^2 + 2x \cdot y + |y|^2 \\ &\leq |x|^2 + 2|x \cdot y| + |y|^2 \quad (\text{ordinary absolute value for } \mathbb{R}) \\ &\leq |x|^2 + 2|x||y| + |y|^2 \quad (\text{Cauchy-Schwarz}) \\ &= (|x| + |y|)^2. \end{aligned}$$

The result follows by taking square roots (since we are dealing with non-negative quantities).  $\square$

2.2.1. *Distance.* The distance between two points in  $\mathbb{R}^n$  is defined to be the length of their difference.

**Definition 2.9.** Let  $x, y \in \mathbb{R}^n$ . The distance between  $x$  and  $y$  is

$$d(x, y) := |x - y| = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}.$$

For example,

$$d((3, 1), (5, -1)) = |(3, 1) - (5, -1)| = |(-2, 2)| = \sqrt{8} = 2\sqrt{2},$$

and

$$d((1, 0, 3, -2, 4), (3, 1, -1, 0, 1)) = |(-2, -1, 4, -2, 3)| = \sqrt{34}.$$

The norm itself can be thought of as the distance between a point and the origin:  $d(x, \vec{0}) = |x - \vec{0}| = |x|$ .

The main properties of the distance function are given by the following proposition.

**Proposition 2.10.** Let  $x, y, z \in \mathbb{R}^n$ ; then

1.  $d(x, y) = d(y, x)$  (symmetric);
2.  $d(x, y) \geq 0$  (positive);
3.  $d(x, y) = 0 \iff x = y$  (definite);
4.  $d(x, y) \leq d(x, z) + d(z, y)$  (triangle inequality).

PROOF: The proof, which follows immediately from Proposition 2.8, is left as an exercise.  $\square$

The function  $d: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$  is called a *distance function* or a *metric* on  $\mathbb{R}^n$ .

### 2.2.2. Angles.

**Definition 2.11.** Let  $x, y \in \mathbb{R}^n \setminus \{\vec{0}\}$ , i.e.,  $x$  and  $y$  are nonzero vectors. The angle between  $x$  and  $y$ , in radians, is the real number  $\theta \in [0, \pi]$  such that

$$\cos \theta = \frac{x \cdot y}{|x||y|}, \text{ i.e., } \theta := \cos^{-1} \left( \frac{x \cdot y}{|x||y|} \right).$$

**Example 2.12.** The angle between  $(1, 0)$  and  $(1, 1)$  is

$$\cos^{-1} \left( \frac{(1, 0) \cdot (1, 1)}{|(1, 0)|| (1, 1)|} \right) = \cos^{-1} \frac{1}{\sqrt{2}} = \frac{\pi}{4},$$

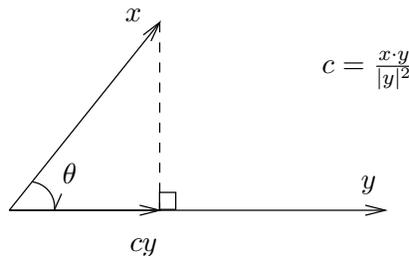
and the angle between  $(1, 0, 2, 1)$  and  $(2, -1, 3, 2)$  is

$$\cos^{-1} \left( \frac{(1, 0, 2, 1) \cdot (2, -1, 3, 2)}{|(1, 0, 2, 1)|| (2, -1, 3, 2)|} \right) = \cos^{-1} \frac{10}{\sqrt{6}\sqrt{18}} \approx 15.8^\circ.$$

Why is our definition of angle reasonable? First off, Cauchy-Schwarz shows that the angle  $\theta$  actually exists since

$$-1 \leq \frac{x \cdot y}{|x||y|} \leq 1.$$

Otherwise, we would not be able to take the inverse cosine. More importantly, our definition of angle makes sense geometrically. Suppose that  $0 \leq \theta \leq \pi/2$ , and consider the picture:



The vector  $cy$  is the projection of  $x$  along  $y$ ; so the triangle drawn is a right triangle, and the cosine should be the ratio of the lengths of the adjacent side and the hypotenuse:

$$\cos \theta = \frac{|cy|}{|x|} = |c| \frac{|y|}{|x|} = \frac{x \cdot y}{|y|^2} \frac{|y|}{|x|} = \frac{x \cdot y}{|x||y|}.$$

The case of  $\pi/2 \leq \theta \leq \pi$  is left as an exercise.

By writing the definition of the angle between vectors in a slightly different form, we finally arrive at a **geometric explanation of the dot product**. It is the product of the lengths of the vectors, scaled by the cosine of angle between them:

$$x \cdot y = |x||y| \cos \theta.$$

**2.2.3. Unit vectors.** According to our definition, given a nonzero vector  $x$  and a nonzero real number  $s$ , then scaling  $x$  by  $s$  gives a vector which points either in the same or opposite direction to that of  $x$ . To see this, let  $\theta$  be the angle between  $x$  and  $sx$ ; then

$$\cos \theta = \frac{x \cdot (sx)}{|x||sx|} = \frac{s}{|s|} \frac{x \cdot x}{|x|^2} = \frac{s}{|s|} = \begin{cases} 1 & \text{if } s > 0 \\ -1 & \text{if } s < 0. \end{cases}$$

Thus,  $\theta$  is 0 or  $\pi$  depending on whether  $s$  is positive or negative, respectively.

A *unit vector* is a vector of length equal to one. For each  $i = 1, \dots, n$ , define the  *$i$ -th standard basis vector in  $\mathbb{R}^n$* , denoted  $e_i$ , to be the vector whose components are all zero except the  $i$ -th component, which is 1: so  $e_i = (0, \dots, 0, 1, 0, \dots, 0)$  with a 1 in the  $i$ -th spot. It is a unit vector pointing in the direction of the  $i$ -th axis. Given any nonzero vector  $x$ , the vector  $x/|x|$  is a unit vector pointing in the same direction (proof: its length is  $|x/|x|| = |1/|x|||x| = 1$  by item 3 of Proposition 2.8).

If  $y$  is a unit vector and  $x$  is any vector, then the component of  $x$  along  $y$  is  $c = x \cdot y/|y|^2 = x \cdot y$ . Thus, in this special case,  $x \cdot y$  is the amount by which you must scale the unit vector  $y$  in order to get the projection of  $x$  along  $y$ , another geometric interpretation of the dot product. For example, note that  $x \cdot e_i = x_i$ , the  $i$ -th component of  $x$ .

You may also want to use unit vectors to think about the cosine of the angle  $\theta$  between any two vectors nonzero  $x$  and  $y$  (where  $y$  is no longer necessarily a unit vector). Since

$$\cos \theta = \frac{x \cdot y}{|x||y|} = \frac{x}{|x|} \cdot \frac{y}{|y|},$$

the cosine is visibly the dot product of the two unit vectors pointing in the same direction as  $x$  and  $y$ .

### 3. Linear subspaces

A linear subspace is a generalization of a line or plane, passing through the origin.

**Definition 3.1.** *A nonempty subset  $W \subseteq \mathbb{R}^n$  is a linear subspace if it is closed under vector addition and scalar multiplication, i.e.,*

1. if  $x, y \in W$ , then  $x + y \in W$ , and
2. if  $x \in W$  and  $s \in \mathbb{R}$ , then  $sx \in W$ .

Think about this definition geometrically. We have already seen that if you scale a vector, you get a vector pointing in the same or opposite direction. By taking all possible scalings of a nonzero vector, you get the line containing that vector and passing through the origin. Thus, the second part of the definition of a linear subspace says that if a point is in a linear subspace, then so is the line through the origin containing that point. Suppose you took two distinct, nonzero points,  $x, y \in \mathbb{R}^n$ , not scalar multiples of each other. By scaling, each determines a line passing through the origin. The union of those two lines is a big X sitting in  $\mathbb{R}^n$  which satisfies the second part of the definition; however, it is not a linear subspace. For instance, take nonzero real numbers  $s$  and  $t$ ; then the points  $sx$  and  $ty$  are from distinct lines in the X. These points and the origin determine a parallelogram whose fourth vertex is the sum  $sx + ty$ . In this way, we can produce any point in the plane determined by the vectors  $x$  and  $y$ . Thus, the first part of the definition implies that any linear subspace containing the X would also need to contain the plane containing the X.

There are only two linear subspaces of  $\mathbb{R}$ ; they are  $\{0\}$  and  $\mathbb{R}$ , itself. The linear subspaces of  $\mathbb{R}^2$  are  $\{(0, 0)\}$ , any line passing through the origin, and  $\mathbb{R}^2$ , itself. The linear subspaces of  $\mathbb{R}^3$  are  $\{(0, 0, 0)\}$ , any line through the origin, any plane through the origin, and  $\mathbb{R}^3$ , itself. Linear subspaces always contain the origin; to see this, let  $s = 0$  in part 2 of the definition. The zero vector, by itself, is always a linear subspace (check!). Similarly  $\mathbb{R}^n$  is always a linear subspace of itself.

**Example 3.2.** There are three main ways you are likely to see linear subspaces presented. We'll look at examples of each.

- (1) Let  $U := \{s(1, 2, 5) + t(0, 1, 8) \mid s, t \in \mathbb{R}\}$ . The linear subspace  $U$  is presented as the *span* of a set consisting of two vectors,  $(1, 2, 5)$  and  $(0, 1, 8)$ .
- (2) Let

$$\begin{aligned} L: \mathbb{R}^2 &\rightarrow \mathbb{R}^3 \\ (u, v) &\mapsto (u, 2u + v, 5u + 8v) \end{aligned}$$

and define  $V := \text{im}(L)$ . The linear subspace  $V$  is presented as the image of a linear function.

- (3) Let  $W := \{(x, y, z) \in \mathbb{R}^3 \mid 11x - 8y + z = 0\}$ . The linear subspace  $W$  is the set of solutions to a homogeneous linear equation.

We'll now prove that each of  $U$ ,  $V$ , and  $W$  is a linear subspace.

- (1) To show that  $U$  is a linear subspace, we need to show that it is nonempty and closed under vector addition and scalar multiplication. The zero vector is clearly an element of  $U$ —let  $s = t = 0$  in the definition of  $U$ —so  $U \neq \emptyset$ . Two arbitrary elements of  $U$  have the form  $x = s(1, 2, 5) + t(0, 1, 8)$  and  $y = s'(1, 2, 5) + t'(0, 1, 8)$ . Their sum is

$$\begin{aligned} x + y &= s(1, 2, 5) + t(0, 1, 8) + s'(1, 2, 5) + t'(0, 1, 8) \\ &= (s + s')(1, 2, 5) + (t + t')(0, 1, 8) \\ &= s''(1, 2, 5) + t''(0, 1, 8), \end{aligned}$$

where  $s'' := s + s'$  and  $t'' := t + t'$ . Thus,  $x + y$  is visibly an element of  $U$ , showing that  $U$  is closed under vector addition. As for scalar multiplication, take any  $s' \in \mathbb{R}$ , then

$$s'x = s'(s(1, 2, 5) + t(0, 1, 8)) = (s's)(1, 2, 5) + (s't)(0, 1, 8) = s''(1, 2, 5) + t''(0, 1, 8),$$

where this time  $s'' := s's$  and  $t'' := s't$ . Hence  $s'x \in U$ , showing that  $U$  is closed under scalar multiplication, as well. Thus, we have shown that  $U$  is a linear subspace of  $\mathbb{R}^3$ . See Definition 3.4 for a generalization.

- (2) To see that  $V = \text{im}(L)$  is a linear subspace, note that the image of  $L$  consists exactly of points of the form

$$(u, 2u + v, 5u + 8v) = u(1, 2, 5) + v(0, 1, 8).$$

Thus,  $V = U$ , and we have already shown that  $U$  is a linear subspace. Proposition 4.11 gives a generalization.

- (3) It turns out that  $W = U$ , too, but we will show that it is a linear subspace directly from the definition. First note that the zero vector is in  $W$ , so  $W$  is nonempty. To show  $W$  is closed under addition, let  $(x, y, z)$  and  $(x', y', z')$  be arbitrary points of  $W$ . It follows that their sum is in  $W$  since

$$11(x + x') - 8(y + y') + (z + z') = (11x - 8y + z) + (11x' - 8y' + z') = 0 + 0 = 0.$$

Note how we have used the fact that  $(x, y, z)$  and  $(x', y', z')$  are in  $W$  in the above calculation. To see that  $W$  is closed under scalar multiplication, take  $t \in \mathbb{R}$  and  $(x, y, z) \in W$ . Then

$$11(tx) - 8(ty) + (tz) = t(11x - 8y + z) = t \cdot 0 = 0,$$

hence,  $t(x, y, z) \in W$ , as required. Therefore, we have shown that  $W$  is a linear subspace. In this example,  $W$  is defined by a single linear equation. There could have been more: the set of simultaneous solutions to a collection of equations like these will always be a linear subspace.

**3.1. Linear combinations, spanning sets, and dimension.** By iterating the two parts of the definition of a linear subspace, we produce what are called *linear combinations* of vectors.

**Definition 3.3.** A linear combination of vectors  $v_1, \dots, v_k \in \mathbb{R}^n$  is any vector of the form

$$v = \sum_{i=1}^k a_i v_i,$$

where  $a_1, \dots, a_k$  are real numbers.

For instance,

$$3(1, -2, 4) + 2(2, 1, 3) - 5(0, 4, -5) = (7, -24, 43)$$

is a linear combination of the three vectors  $(1, -2, 4)$ ,  $(2, 1, 3)$ , and  $(0, 4, -5)$ .

It follows easily from Definition 3.1 that  $W \subseteq \mathbb{R}^n$  is a linear subspace of  $\mathbb{R}^n$  if and only if it is closed under taking linear combinations of its elements, i.e., if  $v$  is a linear combination of elements of  $W$ , then  $v \in W$ . In fact, one often sees a linear subspace defined as the set of all linear combinations of a given set of vectors, as in part 1 of Example 3.2.

**Definition 3.4.** The span of a set of vectors  $v_1, \dots, v_k \in \mathbb{R}^n$  is the linear subspace of  $\mathbb{R}^n$  formed from all linear combinations of the  $v_i$ :

$$\text{Span}(v_1, \dots, v_k) = \left\{ \sum_{i=1}^k a_i v_i \mid a_i \in \mathbb{R} \text{ for } i = 1, \dots, k \right\}.$$

If  $W = \text{Span}(v_1, \dots, v_k)$ , we say that  $W$  is spanned by  $v_1, \dots, v_k$ . If  $S$  is any subset of  $\mathbb{R}^n$ , even an infinite one, we define  $\text{Span}(S)$  to be the set of all linear combinations of all finite subsets of  $S$ .

A trivial way in which to see that a linear subspace  $W \subseteq \mathbb{R}^n$  is the span of some set is to note that  $W = \text{Span}(W)$ . However, using all of  $W$  to span itself is overkill. In fact a theorem from linear algebra, which we'll take on faith in these notes, states that there will always be a *finite* set of spanning vectors. We'll say a set of vectors  $S$  is a *minimal spanning set* of the linear subspace  $W$  if  $W = \text{Span}(S)$  and no proper subset of  $S$  spans  $W$ . A minimal spanning set is not unique, for example  $\{(1, 0), (0, 1)\}$  and  $\{(1, 0), (1, 1)\}$  both span the linear subspace  $\mathbb{R}^2$ . However, in a course on linear algebra, one shows that the number of elements of a minimal spanning set is independent of the minimal spanning set.

**Definition 3.5.** Let  $W$  be a linear subspace of  $\mathbb{R}^n$ , and let  $S$  be a minimal spanning set for  $W$ . The number of elements of  $S$  is called the *dimension* of  $W$ , denoted  $\dim W$ . The set  $S$  is called a *basis* for  $W$ .

**Example 3.6.**

- (1) The dimension of  $\mathbb{R}^n$  is  $n$ . The set  $\{e_1, \dots, e_n\}$  of standard basis vectors is a basis (cf. page 43).
- (2) Let  $S = \{(1, 0, 0), (0, 1, 0), (1, 2, 0)\}$  and define

$$W = \text{Span}(S) = \{(x, y, 0) \in \mathbb{R}^3 \mid x, y \in \mathbb{R}\}.$$

The subset  $S$  is not a minimal spanning set for  $W$ , however any subset of  $S$  consisting of two elements of  $S$  is a minimal spanning set. Thus,  $\dim W = 2$ .

As the second part of this example shows, a linear subspace  $W$  may be presented as the span of a set  $S$  which is not a basis. However, it turns out that one can always find a subset of  $S$  which is a basis. The way to do this is to find an element  $v$  of  $S$  which can be written as a linear combination of the remaining elements. If the set is not a spanning set, there will always be such an element. We say it is *linearly dependent* on the remaining elements. Throw  $v$  out of  $S$  to get a smaller set  $S' \subset S$ . The set  $S'$  still spans  $W$ : any element in  $W$  that you might have written as a linear combination involving  $v$  can be written as a linear combination of elements of the smaller set  $S'$  by writing  $v$  as a linear combination of elements of  $S'$  and substituting. Hence,  $W = \text{Span}(S')$ . If  $S'$  is not a basis, repeat. Keep throwing out linearly dependent elements until you arrive at a basis.

Continuing the previous example, the vector  $(1, 2, 0)$  is linearly dependent on the other two elements:

$$(1, 2, 0) = (1, 0, 0) + 2(0, 1, 0).$$

Throwing it out yields the set  $S' := \{(1, 0, 0), (0, 1, 0)\}$  which forms a basis. The element  $(2, 3, 0) = (1, 0, 0) + (0, 1, 0) + (1, 2, 0)$ , which apparently depends on all three elements of  $S$ , can be written instead as

$$(2, 3, 0) = (1, 0, 0) + (0, 1, 0) + [(1, 0, 0) + 2(0, 1, 0)] = 2(1, 0, 0) + 3(0, 1, 0),$$

so it is in the span of the smaller set  $S'$ .

**Definition 3.7.** Let  $v$  and  $v_1, \dots, v_k$  be elements of  $\mathbb{R}^n$ . The vector  $v$  is linearly dependent on  $v_1, \dots, v_k$  if there are scalars  $a_1, \dots, a_k$  such that

$$v = \sum_{i=1}^k a_i v_i.$$

The vectors  $v_1, \dots, v_k$  are linearly independent if no  $v_i$  is linearly dependent on the remaining, equivalently, if there are no scalars  $a_1, \dots, a_k$ , not all zero, such that

$$\sum_{i=1}^k a_i v_i = \vec{0}.$$

A basis for a linear subspace is a set of vectors that span the space and are linearly independent.

**3.2. Affine subspaces.** To generalize arbitrary lines and planes, not just those that pass through the origin, we need to translate linear subspaces.

**Definition 3.8.** A subset  $A \subseteq \mathbb{R}^n$  is an affine subspace if there is a point  $p \in \mathbb{R}^n$  and a linear subspace  $W \subseteq \mathbb{R}^n$  such that

$$A = p + W := \{p + w \mid w \in W\}.$$

If  $\dim W = k$ , then  $A$  is called a  $k$ -plane, and in this case, we say that the dimension of  $A$  is  $k$ , and we write  $\dim A = k$ . In particular, a 1-plane is a line, a 2-plane is a plane, and an  $(n - 1)$ -plane in  $\mathbb{R}^n$  is a hyperplane.

We have introduced the notation  $A = p + W$  to denote the set obtained by adding  $p$  to every element of  $W$ , i.e., the translation of  $W$  by  $p$ . Note that since  $W$  is a linear subspace,  $\vec{0} \in W$ , hence,  $p = p + \vec{0} \in A$ .

**Example 3.9.**

- (1) Every linear subspace of  $W$  is also an affine subspace: let  $p = \vec{0}$  in the definition.
- (2) The representation of an affine subspace  $A$  as  $p + W$  is not unique. We have  $p + W = q + W$  if and only if  $p - q \in W$ .
- (3) The affine space  $A = (3, 1) + \text{Span}((1, 1))$  is a line passing through the point  $(3, 1)$  and parallel to the line passing through the origin and the point  $(1, 1)$ .
- (4) The affine space  $A = (0, 0, 3) + \text{Span}((1, 0, 0), (0, 1, 0))$  is a plane in  $\mathbb{R}^3$  passing through  $(0, 0, 3)$  and parallel to a coordinate plane.

Hyperplanes are affine subspaces of dimension one less than the ambient space. We say they have *codimension* one. For example, let

$$W = \{(x_1, x_2, x_3, x_4) \in \mathbb{R}^4 \mid 3x_1 - 2x_2 + x_3 + 4x_4 = 0\}.$$

The set  $W$  is an affine subspace; it is easy to check that it is even a linear subspace. We will show that  $W$  is a hyperplane by exhibiting a basis for  $W$  consisting of three elements, thus showing that  $\dim W = 3$ . The basis is formed by the vectors

$$v_1 = (4, 0, 0, -3), \quad v_2 = (0, 4, 0, 2), \quad v_3 = (0, 0, 4, -1).$$

First, note that each  $v_i$  is actually in  $W$ . For instance,  $v_1 \in W$  since it satisfies the single linear equation defining  $W$ :  $3 \cdot 4 - 2 \cdot 0 + 1 \cdot 0 + 4 \cdot (-3) = 0$ . Furthermore, since each  $v_i$  is in  $W$ , we have  $\text{Span}(v_1, v_2, v_3) \subseteq W$ . Second, one may check that *every* element of  $W$  is in  $\text{Span}(v_1, v_2, v_3)$ . For instance, try to show that  $(1, 3, -5, 2)$ , which is an element of  $W$ , is in span of the  $v_i$ . Thus,  $\text{Span}(v_1, v_2, v_3) = W$ . Finally, check that none of  $v_1, v_2$ , or  $v_3$  is in the span of the other two. For instance, an arbitrary linear combination of  $v_2$  and  $v_3$  has the form

$$s(0, 4, 0, 2) + t(0, 0, 4, -1) = (0, 4s, 4t, 2s - t).$$

Note that the first coordinate is zero independent of the choice of  $s$  and  $t$ ; so we can never get  $v_1$ , which has a nonzero first coordinate. We have shown that  $W = \text{Span}(v_1, v_2, v_3)$  and that no subset of  $\{v_1, v_2, v_3\}$  spans  $W$ , hence the vectors form a basis for  $W$ . More generally, we can prove that every hyperplane has a similar form.

**Theorem 3.10.** *Let  $a = (a_1, \dots, a_n) \neq \vec{0}$ , and let  $p = (p_1, \dots, p_n)$ . Define*

$$H = \{x \in \mathbb{R}^n \mid (x - p) \cdot a = 0\}.$$

*Then  $H$  is a hyperplane in  $\mathbb{R}^n$ . Conversely, for every hyperplane  $H$  in  $\mathbb{R}^n$  there are  $a$  and  $p$  in  $\mathbb{R}^n$  so that  $H$  has the above form.*

PROOF: First suppose that

$$H = \{x \in \mathbb{R}^n \mid (x - p) \cdot a = 0\}.$$

Define  $W = \{x \in \mathbb{R}^n \mid x \cdot a = 0\}$ . We have  $x \in H$  if and only if  $(x - p) \cdot a = 0$ , i.e., if and only if  $x - p \in W$ . This shows that  $H = p + W$ . We need to show that  $W$  is a linear subspace of dimension  $n - 1$ .

To see that  $W$  is a linear subspace, let  $x, y \in W$ , and let  $s \in \mathbb{R}$ . Hence,  $x \cdot a = y \cdot a = 0$ . Using the standard properties of the dot product, it follows that  $(x + y) \cdot a = x \cdot a + y \cdot a = 0 + 0 = 0$ , whence  $x + y \in W$ . Similarly,  $(sx) \cdot a = s(x \cdot a) = s(0) = 0$ , whence  $sx \in W$ . We have shown that  $W$  is closed under vector addition and scalar multiplication. The zero vector is clearly in  $W$ , so  $W$  is nonempty.

We must now exhibit a basis for  $W$  consisting of  $n - 1$  elements. Since  $a \neq \vec{0}$ , we may assume without loss of generality that  $a_n \neq 0$ . For  $i = 1, \dots, n - 1$ , define  $v_i := a_n e_i - a_i e_n$ , i.e.,

$$\begin{aligned} v_1 &= a_n(1, 0, \dots, 0) - a_1(0, \dots, 0, 1) = (a_n, 0, \dots, 0, -a_1) \\ v_2 &= (0, a_n, 0, \dots, 0, -a_2) \\ &\vdots \\ v_{n-1} &= (0, \dots, 0, a_n, -a_{n-1}). \end{aligned}$$

Each  $v_i \in W$  since  $v_i \cdot a = a_n a_i - a_i a_n = 0$ . Hence,  $\text{Span}(v_1, \dots, v_{n-1}) \subseteq W$ . To see that  $\text{Span}(v_1, \dots, v_{n-1}) = W$ , take  $w = (w_1, \dots, w_n) \in W$ . We write  $w$  as a linear combination of the  $v_i$  as follows:

$$\begin{aligned} \sum_{i=1}^{n-1} \frac{w_i}{a_n} v_i &= \frac{w_1}{a_n} (a_n, 0, \dots, 0, -a_1) + \frac{w_2}{a_n} (0, a_n, 0, \dots, 0, -a_2) + \\ &\quad \dots + \frac{w_{n-1}}{a_n} (0, \dots, 0, a_n, -a_{n-1}) \\ &= (w_1, w_2, \dots, w_{n-1}, (-a_1 w_1 - a_2 w_2 - \dots - a_{n-1} w_{n-1})/a_n). \end{aligned}$$

Since  $w \in W$ , it follows that  $w \cdot a = a_1 w_1 + \dots + a_n w_n = 0$ . Solving for  $w_n$  gives  $w_n = (-a_1 w_1 - a_2 w_2 - \dots - a_{n-1} w_{n-1})/a_n$ . So  $\sum_{i=1}^{n-1} \frac{w_i}{a_n} v_i = w$ , as required.

It remains to be shown that none of  $v_1, \dots, v_{n-1}$  is in the span of the remaining  $v_i$ . This follows because for each  $j$ , the sum  $\sum_{i:i \neq j} s_i v_i$  has  $j$ -th coordinate equal to zero for all choices of scalars  $s_i$  while the  $j$ -th coordinate of  $v_j$  is  $a_n \neq 0$ .

We have shown that every subset of a certain form is a hyperplane. The converse is a standard result from linear algebra which we will not cover in these notes.  $\square$

Note that if  $H$  is defined as in the theorem, it is easily visualized as the set of all points which when translated by  $-p$  are perpendicular to the vector  $a$ . In other words, to get  $H$ , take all points perpendicular to the vector  $a$  (obtaining what we called  $W$ ) then translate out by  $p$ .

**Example 3.11.**

- (1) A hyperplane in  $\mathbb{R}^2$  is just a line. For example,

$$\ell := \{(x, y) \in \mathbb{R}^2 \mid ((x, y) - (3, -2)) \cdot (4, 1) = 0\}$$

defines a line passing through the point  $(3, -2)$  and perpendicular to the vector  $(4, 1)$ . It is defined via the equation  $4(x - 3) + (y + 2) = 0$ , i.e.,  $y = -4x + 10$ .

- (2) The following defines a hyperplane in  $\mathbb{R}^4$  which passes through the point  $(1, 2, 3, 4)$  and is perpendicular to the vector  $(2, 1, 3, 2)$ :

$$\begin{aligned} H &:= \{(w, x, y, z) \in \mathbb{R}^4 \mid ((w, x, y, z) - (1, 2, 3, 4)) \cdot (2, 1, 3, 2) = 0\} \\ &= \{(w, x, y, z) \in \mathbb{R}^4 \mid (w - 1, x - 2, y - 3, z - 4) \cdot (2, 1, 3, 2) = 0\} \\ &= \{(w, x, y, z) \in \mathbb{R}^4 \mid 2(w - 1) + (x - 2) + 3(y - 3) + 2(z - 4) = 0\} \\ &= \{(w, x, y, z) \in \mathbb{R}^4 \mid 2w + x + 3y + 2z = 21\} \end{aligned}$$

- (3) Let

$$H := \{(x, y, z) \in \mathbb{R}^3 \mid 3x - 2y + 3z = 6\}.$$

By applying a trick, we can write  $H$  in the form given in the theorem and thus show it is a hyperplane. The trick is to write the constant 6 as  $(3, -2, 3) \cdot (u, v, w)$  for some  $(u, v, w) \in \mathbb{R}^3$ . The vector  $(3, -2, 3)$  comes from the coefficients of the linear equation defining  $H$ . There are many choices for  $(u, v, w)$ ; for instance,  $6 = (3, -2, 3) \cdot (2, 0, 0)$ . Then

$$H := \{(x, y, z) \in \mathbb{R}^3 \mid ((x, y, z) - (2, 0, 0)) \cdot (3, -2, 3) = 0\}.$$

A similar trick shows that hyperplanes can be exactly characterized as the sets of solutions  $(x_1, \dots, x_n)$  to nonzero equations of the form  $a_1x_1 + \dots + a_nx_n = d$ .

#### 4. Linear functions

You should be familiar with the ad hoc introduction to linear functions and matrices from the beginning of these notes (cf. page 10). We now study the topic more carefully, starting with the actual definition of a linear function.

**Definition 4.1.** A function  $L: \mathbb{R}^n \rightarrow \mathbb{R}^m$  is a linear function, also called a linear mapping or linear transformation, if

1.  $L(x + y) = L(x) + L(y)$  for all  $x, y \in \mathbb{R}^n$ ;
2.  $L(sx) = sL(x)$  for all  $s \in \mathbb{R}$  and  $x \in \mathbb{R}^n$ .

The first part of the definition says that  $L$  preserves vector addition, and the second part says that it preserves scalar multiplication. A linear function is a function which preserves linear structure. It does not matter if we add or scale vectors before or after applying the function; the result is the same.

**Example 4.2.** The following functions are linear:

1.  $L(x, y) = (3x - 4y, 2x + 5y)$ ;
2.  $L(x, y) = (2x + 3y, x - 5y, 0, 6x + 2y, 2x)$ ;
3.  $W(x, y, z) = 5x - 2y + 3z$ ;
4.  $f(t) = (5t, 3t, -6t)$ ;
5.  $\ell(u) = 3u$ ;
6.  $L(u, v) = u(4, 3, 2) + v(7, -1, 3)$ ;
7.  $E(s, t, u) = (0, 0)$ .

The following functions are *not* linear:

1.  $f(x) = x^2$ ;
2.  $L(x, y) = (x^2 + 2y, x - 5y^4)$ ;
3.  $E(u, v) = 7 \cos(uv)$ ;
4.  $r(a, b) = (3 + a + b, a + 4b)$ ;
5.  $L(x) = 4$ .

To prove that  $L(x, y) = (3x - 4y, 2x + 5y)$  is linear, first show that it preserves addition. Let  $(x, y)$  and  $(x', y')$  be any two points in  $\mathbb{R}^2$ ; then

$$\begin{aligned} L((x, y) + (x', y')) &= L(x + x', y + y') \\ &= (3(x + x') - 4(y + y'), 2(x + x') + 5(y + y')) \\ &= ((3x - 4y) + (3x' - 4y'), (2x + 5y) + (2x' + 5y')) \\ &= (3x - 4y, 2x + 5y) + (3x' - 4y', 2x' + 5y') \\ &= L(x, y) + L(x', y'), \end{aligned}$$

as required. Now show it preserves scalar multiplication. Take any point  $(x, y) \in \mathbb{R}^2$  and scalar  $s \in \mathbb{R}$ ; then

$$L(s(x, y)) = L(sx, sy) = (3sx - 4sy, 2sx + 5sy) = s(3x - 4y, 2x + 5y) = s(L(x, y)).$$

To prove that a function is not linear, it is best to find a simple explicit example which shows that it does not preserve linear structure. For instance  $f(x) = x^2$  is not linear since  $f(1 + 1) = f(2) = 4 \neq f(1) + f(1) = 2$ . It is easy to show that a linear function sends the zero vector in its domain to the zero vector in its codomain (do it!); so this is often an easy way to see that a function is not linear. It doesn't work for  $f(x) = x^2$ , but it does for  $r(a, b) = (3 + a + b, a + 4b)$  or  $L(x) = 4$ .

**4.1. Picturing linear functions.** A key to understanding linear functions is to see that they are completely determined by their action on the standard basis vectors. For example, let  $L: \mathbb{R}^2 \rightarrow \mathbb{R}^3$  be a linear function such that  $L(1, 0) = (2, 3, -1)$  and  $L(0, 1) = (4, 1, 2)$ . What is  $L(2, 3)$ ? Since  $L$  is linear, we know that

$$L(2, 3) = L(2(1, 0) + 3(0, 1)) = 2L(1, 0) + 3L(0, 1) = 2(2, 3, -1) + 3(4, 1, 2) = (16, 9, 4).$$

Similarly,  $L(-1, 3) = -1(2, 3, -1) + 3(4, 1, 2) = (10, 0, 7)$ . In general, if  $L: \mathbb{R}^n \rightarrow \mathbb{R}^m$ , and  $L(e_i) = v_i \in \mathbb{R}^m$ , then

$$L(x_1, \dots, x_n) = L(x_1 e_1 + \dots + x_n e_n) = \sum_{i=1}^n x_i L(e_i) = \sum_{i=1}^n x_i v_i.$$

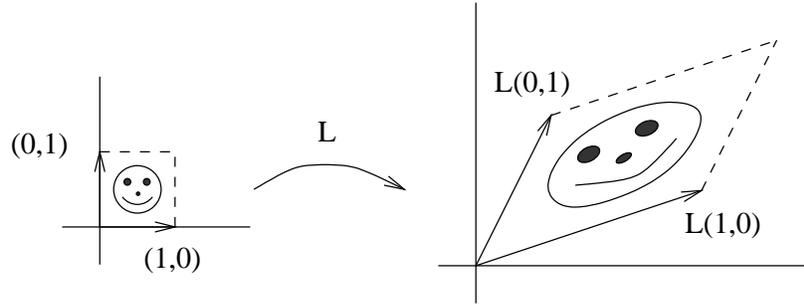
Knowing the image of each standard basis vector  $e_i$  allows us to calculate the image of any vector  $(x_1, \dots, x_n)$ .

To picture a linear function, imagine the images of the standard basis vectors. Using the geometric interpretation of scaling and of vector addition, you can then imagine the image of any point.

**Example 4.3.** Consider the function

$$\begin{aligned} L: \mathbb{R}^2 &\rightarrow \mathbb{R}^2 \\ (x, y) &\mapsto (3x + y, x + 2y) \end{aligned}$$

So  $L(1, 0) = (3, 1)$  and  $L(0, 1) = (1, 2)$ . To picture the image of  $(2, 3) = 2(1, 0) + 3(0, 1)$ , imagine doubling the length of  $(3, 1)$  and tripling the length of  $(1, 2)$ , then using the parallelogram law for vector addition. Next, try imagining the image of each point of a unit square:



## 4.2. Matrices.

**Definition 4.4.** An  $m \times n$  matrix is a rectangular box with  $m$  rows and  $n$  columns. The entry in the  $i$ -th row and  $j$ -th column is denoted  $A_{i,j}$ .\*

Let  $A$  and  $B$  be  $m \times n$  matrices. Define their sum,  $A + B$ , to be the  $m \times n$  matrix with  $i, j$ -th entry  $(A + B)_{i,j} := A_{i,j} + B_{i,j}$ . If  $s \in \mathbb{R}$ , define  $sA$  to be the  $m \times n$  matrix with  $i, j$ -th entry  $(sA)_{i,j} := s(A_{i,j})$ . In this way we have a linear structure on the set of  $m \times n$  matrices. For example,

$$\begin{pmatrix} 1 & 2 \\ 5 & 7 \end{pmatrix} + \begin{pmatrix} 2 & 4 \\ 0 & 7 \end{pmatrix} = \begin{pmatrix} 3 & 6 \\ 5 & 14 \end{pmatrix} \quad \text{and} \quad 5 \begin{pmatrix} 2 & 0 & 3 \\ -1 & 6 & 4 \end{pmatrix} = \begin{pmatrix} 10 & 0 & 15 \\ -5 & 30 & 20 \end{pmatrix}$$

You can only add matrices with equal dimensions. For example, you cannot add a  $2 \times 3$  matrix to a  $2 \times 2$  matrix.

The following proposition states the main properties of matrix addition and scalar multiplication. The symbol “0” appearing in the proposition denotes the  $m \times n$  zero matrix. It is a matrix whose entries are all zeroes. Its dimensions can usually be inferred from the context.

**Proposition 4.5.** Let  $A, B$ , and  $C$  be  $m \times n$  matrices, and let  $s, t \in \mathbb{R}$ ; then

1.  $A + B = B + A$  (commutativity);
2.  $A + (B + C) = (A + B) + C$  (associativity);
3.  $A + 0 = 0 + A = A$  (additive identity exists);
4.  $A + (-1)A = (-1)A + A = 0$  (additive inverses exist);
5.  $s(A + B) = sA + sB$  (distributivity);
6.  $(s + t)A = sA + tA$  (distributivity);
7.  $(st)A = s(tA)$  (associativity);
8.  $1 \cdot A = A$ .

**PROOF:** All parts of the proposition follow directly from the definitions and the fact that two matrices are equal exactly when all of their corresponding entries are equal. For instance,  $A + B = B + A$  because

$$(A + B)_{i,j} = A_{i,j} + B_{i,j} = B_{i,j} + A_{i,j} = (B + A)_{i,j}$$

for all row indices  $i$  and column indices  $j$ . The first and last steps come from the definition of matrix addition; the main step is in the middle, where we appeal to the commutative

\*If this definition is not rigorous enough for you, define an  $m \times n$  matrix with entries in a set  $S$  to be a function  $A: \{1, 2, \dots, m\} \times \{1, 2, \dots, n\} \rightarrow S$ . The value of the function  $A(i, j)$  is the  $i, j$ -th entry  $A_{i,j}$ , and we represent  $A$  using a rectangular box.

property of addition of real numbers.  $\square$

According to part 4 of the proposition, we can denote  $(-1)A$  by just  $-A$ . It is the additive inverse of  $A$ . Define subtraction by  $A - B = A + (-B)$ , as usual.

Matrices can also be multiplied. If  $A$  is an  $m \times r$  matrix and  $B$  is an  $r \times n$  matrix, then their product,  $AB$  is the  $m \times n$  matrix whose  $i, j$ -th entry is

$$(AB)_{i,j} := \sum_{k=1}^r A_{i,k} B_{k,j}.$$

It is probably easier to think of this as saying the  $i, j$ -th entry of  $AB$  is the dot product of the  $i$ -th row of  $A$  with the  $j$ -th column of  $B$ . For example,

$$\begin{pmatrix} 1 & 0 & 2 \\ 3 & -4 & 6 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 & 1 \\ 3 & 1 & 1 & 1 \\ 2 & 6 & 4 & 3 \end{pmatrix} = \begin{pmatrix} 5 & 13 & 8 & 7 \\ 3 & 35 & 20 & 17 \end{pmatrix}$$

The entry in the second row and third column of the product comes from taking the dot product of the second row of the first matrix and the third column of the second matrix:

$$(3, -4, 6) \cdot (0, 1, 4) = 20.$$

Note that to multiply two matrices, the length of a row of the first matrix (i.e., the number of columns) must equal the length of a column of the second matrix (i.e., the number of rows).

A few more examples:

$$\begin{aligned} \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 4 & 5 \end{pmatrix} &= \begin{pmatrix} 8 & 5 \\ 14 & 15 \end{pmatrix} \\ (1 & 2 & 3 & 5) \begin{pmatrix} 1 \\ -1 \\ 2 \\ 3 \end{pmatrix} &= (20) \\ \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix} (4 & 2 & 1) &= \begin{pmatrix} 4 & 2 & 1 \\ 12 & 6 & 3 \\ 8 & 4 & 2 \end{pmatrix} \\ \begin{pmatrix} 1 & -3 & 1 \\ 2 & 0 & 2 \end{pmatrix} \begin{pmatrix} 2 & 3 \\ 1 & 3 \\ 1 & 0 \end{pmatrix} + 4 \begin{pmatrix} -1 & 2 \\ 1 & 3 \end{pmatrix} &= \begin{pmatrix} -4 & 2 \\ 10 & 18 \end{pmatrix} \end{aligned}$$

Some of the main properties of matrix multiplication are summarized below.

**Proposition 4.6.** *Suppose  $A$  is an  $m \times n$  matrix and  $B$  is an  $n \times r$  matrix. Let  $t \in \mathbb{R}$ .*

1.  $t(AB) = (tA)B = A(tB)$ ;
2.  $A(BC) = (AB)C$  for  $C$  an  $r \times s$  matrix (associativity);
3.  $A(B + C) = AB + AC$  for  $C$  an  $n \times r$  matrix (distributivity).

PROOF: The proofs all follow from the definition of matrix multiplication, although the definition is a bit cumbersome. The most difficult part is the middle one. Again we prove

the equality by proving it on the level of individual entries:

$$\begin{aligned}
 [A(BC)]_{i,j} &= \sum_{k=1}^n A_{i,k}(BC)_{k,j} \\
 &= \sum_{k=1}^n A_{i,k} \left( \sum_{\ell=1}^r B_{k,\ell} C_{\ell,j} \right) \\
 &= \sum_{k=1}^n \sum_{\ell=1}^r A_{i,k} B_{k,\ell} C_{\ell,j} \\
 &= \sum_{\ell=1}^r \sum_{k=1}^n A_{i,k} B_{k,\ell} C_{\ell,j} \\
 &= \sum_{\ell=1}^r \left( \sum_{k=1}^n A_{i,k} B_{k,\ell} \right) C_{\ell,j} \\
 &= \sum_{\ell=1}^r (AB)_{i,\ell} C_{\ell,j} \\
 &= [(AB)C]_{i,j}.
 \end{aligned}$$

□

Be careful, matrix multiplication is not commutative. For example

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} = \begin{pmatrix} 19 & 22 \\ 43 & 50 \end{pmatrix} \quad \text{but} \quad \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 23 & 34 \\ 31 & 46 \end{pmatrix}$$

In fact, if the dimensions are not right, multiplication might make sense in one order but not the other:

$$\text{does not make sense} \rightsquigarrow \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix} \quad \text{makes sense} \rightsquigarrow \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

The cancellation law does not necessarily hold, i.e., it is possible for matrices  $A$ ,  $B$ , and  $C$  to satisfy  $AB = AC$  with  $A \neq 0$  even though  $B \neq C$ :

$$\begin{pmatrix} 0 & 1 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} 3 & 5 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} 6 & 2 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 3 & 3 \end{pmatrix} \quad \text{but} \quad \begin{pmatrix} 3 & 5 \\ 1 & 1 \end{pmatrix} \neq \begin{pmatrix} 6 & 2 \\ 1 & 1 \end{pmatrix}$$

It is also possible for matrices  $A$  and  $B$  to satisfy  $AB = 0$  even though  $A \neq 0$  and  $B \neq 0$ :

$$\begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 4 & 2 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

Another special matrix is  $n \times n$  *identity matrix*, denoted  $I_n$  or just  $I$ , is the square matrix whose diagonal entries are 1s and the other entries are 0, so

$$(I_n)_{i,j} = \delta_{i,j} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

The identity matrix works like the number 1 but for square matrices (“square” means the number of rows equals the number of columns). If  $A$  is any  $n \times n$  matrix, then  $AI_n = I_n A = A$ .

4.2.1. *Matrices and linear functions revisited.* In section 2.5 you saw how matrices correspond to linear functions. The  $i$ -th row of a matrix corresponds to the coefficients of the  $i$ -th component function of the corresponding linear function. For example,

$$A = \begin{pmatrix} 3 & 2 \\ 5 & -1 \\ 7 & 6 \end{pmatrix} \rightsquigarrow L: \mathbb{R}^2 \rightarrow \mathbb{R}^3 \\ (x, y) \mapsto (3x + 2y, 5x - y, 7x + 6y)$$

Each column of  $A$ , thought of as an element in  $\mathbb{R}^3$ , is the image of the corresponding standard basis vector under  $L$ :

$$L(e_1) = L(1, 0) = (3, 5, 7) \quad \text{and} \quad L(e_2) = L(0, 1) = (2, -1, 6).$$

We think of each column of  $A$  as a  $3 \times 1$  matrix, and identify it with a point in  $\mathbb{R}^3$  in the natural way:

$$\begin{pmatrix} 3 \\ 5 \\ 7 \end{pmatrix} \rightsquigarrow (3, 5, 7) \quad \text{and} \quad \begin{pmatrix} 2 \\ -1 \\ 6 \end{pmatrix} \rightsquigarrow (2, -1, 6)$$

The image of a general point  $(x, y) = \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2$  can be found through matrix multiplication:

$$\begin{pmatrix} 3 & 2 \\ 5 & -1 \\ 7 & 6 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 3x + 2y \\ 5x - y \\ 7x + 6y \end{pmatrix} = (3x + 2y, 5x - y, 7x + 6y) = L(x, y),$$

again, identifying column vectors with points in Euclidean space. The whole image of  $L$  is the span of the columns of  $A$ , as can be seen explicitly as follows:

$$L(x, y) = (3x + 2y, 5x - y, 7x + 6y) = x(3, 5, 7) + y(2, -1, 6).$$

All of this generalizes to arbitrary matrices and linear functions. We summarize our observations below. They are essential for understanding the relation between matrices and linear functions. From now on, we will identify points in  $\mathbb{R}^k$  with *column matrices*, i.e., matrices with a single column (also called *column vectors*):

$$x = (x_1, \dots, x_k) \rightsquigarrow x = \begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix}$$

Let  $A$  be an  $m \times n$  matrix with corresponding linear function  $L: \mathbb{R}^n \rightarrow \mathbb{R}^m$ .

- The  $j$ -th column of  $A$  is  $L(e_j)$ , the image of the  $j$ -th standard basis vector under  $L$ .
- We can calculate  $L(x)$  through matrix multiplication:  $L(x) = Ax$ .
- The image of  $L$  is the span of the columns of  $A$ .

We finish this section by putting a linear structure on the set of all linear functions with fixed domain and codomain and then showing how basic operations on linear functions correspond to operations on their corresponding matrices.

**Definition 4.7.** Let  $L: \mathbb{R}^n \rightarrow \mathbb{R}^m$  and  $M: \mathbb{R}^n \rightarrow \mathbb{R}^m$  be linear functions, and let  $s \in \mathbb{R}$ . The sum,  $L + M$ , is the linear function

$$L + M: \mathbb{R}^n \rightarrow \mathbb{R}^m \\ x \mapsto L(x) + M(x)$$

and  $sL$  is the linear function:

$$\begin{aligned} sL: \mathbb{R}^n &\rightarrow \mathbb{R}^m \\ x &\mapsto s(L(x)) \end{aligned}$$

Hence,  $(L + M)(x) := L(x) + M(x)$  and  $(sL)(x) := s(L(x))$ .

For example, let  $L(x, y) = (2x + 5y, 3x - y, 2x + 7y)$  and  $M(x, y) = (3x - 2y, 5y, 6x + 2y)$ , both functions from  $\mathbb{R}^2$  to  $\mathbb{R}^3$ . Then

$$\begin{aligned} (L + M)(x, y) &:= L(x, y) + M(x, y) \\ &= (2x + 5y, 3x - y, 2x + 7y) + (3x - 2y, 5y, 6x + 2y) \\ &= (5x + 3y, 3x + 4y, 8x + 9y). \end{aligned}$$

and, for instance,

$$\begin{aligned} (3L)(x, y) &:= 3(L(x, y)) \\ &= 3(2x + 5y, 3x - y, 2x + 7y) \\ &= (6x + 15y, 9x - 3y, 6x + 21y). \end{aligned}$$

We have identified matrices with linear functions, and we have put linear structures on both sets. It is clear from the definitions that the linear structures correspond under the identification. We state this as a formal proposition.

**Proposition 4.8.** *Let  $L: \mathbb{R}^n \rightarrow \mathbb{R}^m$  and  $M: \mathbb{R}^n \rightarrow \mathbb{R}^m$  be linear functions with corresponding matrices  $A$  and  $B$ , respectively. Then the matrix corresponding to the linear function  $L + M$  is  $A + B$ . If  $s \in \mathbb{R}$ , then the matrix corresponding to  $sL$  is  $sA$ .*

Continuing with the example from above, the matrices for  $L$  and  $M$  are

$$L \rightsquigarrow A = \begin{pmatrix} 2 & 5 \\ 3 & -1 \\ 2 & 7 \end{pmatrix} \quad M \rightsquigarrow B = \begin{pmatrix} 3 & -2 \\ 0 & 5 \\ 6 & 2 \end{pmatrix}$$

You can see that the matrices corresponding to  $L + M$  and  $3L$  are

$$\begin{aligned} A + B &= \begin{pmatrix} 2 & 5 \\ 3 & -1 \\ 2 & 7 \end{pmatrix} + \begin{pmatrix} 3 & -2 \\ 0 & 5 \\ 6 & 2 \end{pmatrix} = \begin{pmatrix} 5 & 3 \\ 3 & 4 \\ 8 & 9 \end{pmatrix} \\ 3A &= 3 \begin{pmatrix} 2 & 5 \\ 3 & -1 \\ 2 & 7 \end{pmatrix} = \begin{pmatrix} 6 & 15 \\ 9 & -3 \\ 6 & 21 \end{pmatrix} \end{aligned}$$

You may have been wondering why we defined matrix multiplication as we did. It turns out that this seemingly bizarre operation on matrices corresponds with the natural operation of composition for the associated linear functions.

**Proposition 4.9.** *Let  $L: \mathbb{R}^n \rightarrow \mathbb{R}^t$  and let  $M: \mathbb{R}^t \rightarrow \mathbb{R}^m$  be linear functions with corresponding matrices  $A$  and  $B$ , respectively. Then the matrix corresponding to the composition  $M \circ L: \mathbb{R}^n \rightarrow \mathbb{R}^m$  is the product,  $BA$ .*

PROOF: We have already observed that applying a linear function to  $x \in \mathbb{R}^n$  is the same as multiplying the corresponding matrix with  $x$  provided we think of  $x$  as a column matrix. Thus,  $L(x) = Ax$ , and

$$(M \circ L)(x) = M(L(x)) = M(Ax) = (BA)x,$$

by associativity of matrix multiplication.  $\square$

As an example, define  $L(x, y) = (3x + 2y, x - y, 2x + 2y)$ , and define  $M(x, y, z) = (x + 2y - z, 4x + 2z)$ . Then

$$\begin{aligned} (M \circ L)(x, y) &= M(L(x, y)) \\ &= M(3x + 2y, x - y, 2x + 2y) \\ &= ((3x + 2y) + 2(x - y) - (2x + 2y), 4(3x + 2y) + 2(2x + 2y)) \\ &= (3x - 2y, 16x + 12y). \end{aligned}$$

The matrices corresponding to  $L$  and  $M$  are

$$L \rightsquigarrow \begin{pmatrix} 3 & 2 \\ 1 & -1 \\ 2 & 2 \end{pmatrix} \quad M \rightsquigarrow \begin{pmatrix} 1 & 2 & -1 \\ 4 & 0 & 2 \end{pmatrix}$$

The point is that the matrix corresponding to  $M \circ L$  is the product  $BA$ :

$$BA = \begin{pmatrix} 1 & 2 & -1 \\ 4 & 0 & 2 \end{pmatrix} \begin{pmatrix} 3 & 2 \\ 1 & -1 \\ 2 & 2 \end{pmatrix} = \begin{pmatrix} 3 & -2 \\ 16 & 12 \end{pmatrix} \rightsquigarrow M \circ L.$$

**4.3. Affine functions, parametrizations.** Just as affine subspaces are translations of linear subspaces, affine functions are “translations” of linear functions.

**Definition 4.10.** A function  $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$  is an affine function (or transformation or mapping) if there is a linear function  $L: \mathbb{R}^n \rightarrow \mathbb{R}^m$  and a point  $p \in \mathbb{R}^m$  such that  $f(x) = p + L(x)$  for all  $x \in \mathbb{R}^n$ .

Here are a few examples of affine functions:

1.  $f(x, y) = (3 + 5x - 2y, 1 + 2x + 6y)$ ;
2.  $g(u, v, w) = 3u - 2v + w + 6$ ;
3.  $\ell(t) = 3 + 6t$ ;
4.  $r(s) = 5$ .
5.  $L(x, y) = (4x + 5y, 3x + 3y, 6x - y)$ .

For instance, the first function can be written  $f(x, y) = (3, 1) + (5x - 2y, 2x + 6y)$ , the translation of the linear function  $(x, y) \mapsto (5x - 2y, 2x + 6y)$  by the point  $(3, 1)$ . Note that the last part of the example illustrates the point that every linear function is affine (let  $p = \vec{0}$  in the definition of an affine function).

We say that an affine function *parametrizes* its image, especially in the case when the function is 1–1. The image of a linear function is always a linear subspace, and the image of an affine function is always an affine subspace. In fact, a bit more is true:

**Proposition 4.11.** *Let  $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$  be an affine function, and let  $W \subseteq \mathbb{R}^n$  be an affine subspace. Then the image of  $W$  under  $f$ ,*

$$f(W) := \{f(x) \in \mathbb{R}^m \mid x \in W\},$$

*is an affine subspace of  $\mathbb{R}^m$ . If  $f$  is a linear function, and  $W$  is a linear subspace, then  $f(W)$  is a linear subspace.*

PROOF: Suppose that  $f$  is affine; so  $f(x) = p + L(x)$  for some  $p \in \mathbb{R}^m$  and some linear function  $L: \mathbb{R}^n \rightarrow \mathbb{R}^m$ . Let  $W$  be an affine subspace of  $\mathbb{R}^n$ , which means that there is a point  $q \in \mathbb{R}^n$  and a linear subspace  $V \subseteq \mathbb{R}^n$  such that  $W = q + V$ . An arbitrary point in  $W$  has the form  $q + v$  where  $v \in V$ , and its image under  $f$  is  $f(q + v) = p + L(q + v) = p + L(q) + L(v)$ . This shows that  $f(W) = (p + L(q)) + L(V)$ , where  $p + L(q)$  is just a fixed point in  $\mathbb{R}^m$ . Thus, to show that  $f$  is affine, it suffices to show  $L(V)$  is a linear subspace of  $\mathbb{R}^m$ .

To show that  $L(V)$  is a linear subspace, we need to show that it is closed under vector addition and scalar multiplication. This follows exactly because  $L$  preserves linear structure. Take two arbitrary points in  $L(V)$ ; they will have the form  $L(u)$  and  $L(v)$  with  $u, v \in V$ . We first want to show that  $L(u) + L(v) \in L(V)$ . Since  $L$  is a linear function, we have  $L(u) + L(v) = L(u + v)$ , and  $u + v \in V$  since  $V$  is a linear subspace. Thus, we have exhibited  $L(u) + L(v)$  as the image of a point in  $V$ . Similarly,  $L(V)$  is closed under scalar multiplication. Take  $s \in \mathbb{R}$  and  $v \in V$ ; then  $sL(v) = L(sv)$ , and  $sv \in V$  since  $V$  is a linear subspace. Thus, a scalar multiple of an arbitrary point in  $L(V)$  is still in  $L(V)$ .

Note that we have just shown that the image of a linear subspace under a linear function is a linear subspace of the codomain, which proves the second part of the proposition, as well.  $\square$

Thus, for instance, the image of a line or plane under an affine function is again a line or plane provided the function is 1–1 (otherwise, it could be a smaller-dimensional linear subspace). The image of a derivative of a function is a linear subspace, the tangent space, and the image of a best affine approximation of a function is an affine subspace, the translation of the tangent space out to the point in question.

## 5. Conclusion

We have put a linear structure on  $\mathbb{R}^n$ , which means that we can add and scale vectors. These algebraic operations correspond to the geometric operations of translation and scaling, respectively. In order to make measurements in  $\mathbb{R}^n$ , we defined the dot product. This simple device, alone, allowed us to define length, distance, projections, and angles.

Linear functions are functions that preserve linear structure. A linear function plus a translation gives an affine function. The main point of differential calculus is to reduce the study of complicated functions to linear ones. We have seen how there is a 1–1 correspondence between matrices and linear functions—the same correspondence that occurs between the Jacobian matrix and the derivative—and that under this correspondence, addition and scalar multiplication of matrices coincides with addition and scalar multiplication of linear functions. The product of matrices corresponds to composition of the associated linear functions. One of the most important theorems in calculus is the chain rule, which says, roughly, that the derivative of a composition of functions is the composition of their derivatives. Equivalently, the Jacobian matrix of a composition of functions is the product of their individual Jacobian matrices. We will get to this theorem in the next chapter.

Linear subspaces are subsets closed under vector addition and scalar multiplication. Their translations are affine subspaces. Since linear functions preserve linear structure, the image of a linear subspace under a linear function is again a linear subspace (and similarly for affine subspaces and functions). Thus, these subspaces appear all the time in differential calculus as images of derivatives (or the best affine approximation function).

## EXERCISES

- (1) Let  $v = (1, 5)$ ,  $w = (3, 2)$ ,  $u = (-1, 4)$ , and  $s = -2$ . By explicit calculation, show that
  - (a)  $v \cdot (w + u) = v \cdot w + v \cdot u$ ;
  - (b)  $(sv) \cdot w = v \cdot (sw) = s(v \cdot w)$ .
- (2) Prove the remaining items in Proposition 1.1.
- (3) Prove the remaining items in Proposition 2.2.
- (4) Prove the first three parts of Proposition 2.8.
- (5) Compute:
  - (a)  $(1, 4, 0, 3) \cdot (2, -2, 1, 1)$ .
  - (b)  $|(2, -3, 1, 5, 1)|$ .
  - (c)  $d((1, 5, 3, -1), (3, -2, 5, 0))$ .
- (6) Find three vectors perpendicular to  $(4, 1, 4, 1)$ .
- (7) What are the angles between the following pairs of vectors? Write your answer as the inverse cosine of something, then also use a calculator to express the approximate solution in degrees.
  - (a)  $(1, 0)$  and  $(3, 4)$ .
  - (b)  $(0, 2, -1)$  and  $(5, -1, 3)$ .
  - (c)  $(1, 0, 1, 0)$  and  $(0, 1, 0, 1)$ .
- (8) For each of the following pairs of vectors,  $u, v$ , find (i) the component of  $u$  along  $v$ , and (ii) the projection of  $u$  along  $v$ .
  - (a)  $u = (2, 0)$ ,  $v = (3, 5)$ .
  - (b)  $u = (3, 2, 1)$ ,  $v = (1, 0, -3)$ .
  - (c)  $u = (1, 1, 1, 1, 1)$ ,  $v = (2, 5, -2, 5, 1)$ .
- (9) Find a vector pointing in the same direction as  $(4, 5, 1, 2, -3)$  but whose length is 1.
- (10) Let  $u, v \in \mathbb{R}^n$ .
  - (a) Prove that  $4u \cdot v = |u + v|^2 - |u - v|^2$ .
  - (b) Prove that  $|u + v|^2 + |u - v|^2 = 2|u|^2 + 2|v|^2$ . Show with a picture how this formula states that the sum of the squares of the lengths of the diagonals of a parallelogram equals the sum of the squares of the sides.
- (11) Use the triangle inequality to prove the reverse triangle inequality:  $|x| - |y| \leq |x + y|$  for all  $x, y \in \mathbb{R}^n$ .
- (12) Prove that if two medians of a triangle have the same length, then the triangle is isosceles. (Hint: You may assume that the vertices are  $(0, 0)$ ,  $a$ ,  $b$  and that the medians are the line segments connecting  $(0, 0)$  to  $(a + b)/2$  and  $b$  to  $a/2$ . Draw a picture. As in our proof of the Pythagorean theorem, resist the temptation to use coordinates.)
- (13) In the section on angles, we argued that our definition of angle is reasonable by drawing a picture of a triangle and appealing to the idea that the cosine should be the ratio of the lengths of the adjacent side and the hypotenuse. We assumed  $0 \leq \theta \leq \pi/2$ . Give a similar argument for the case  $\pi/2 \leq \theta \leq \pi$ .
- (14) Prove the law of cosines:  $|a - b|^2 = |a|^2 + |b|^2 - 2|a||b| \cos(\theta)$  for all  $a, b \in \mathbb{R}^n$  where  $\theta$  is the angle between  $a$  and  $b$ . (Work straight from the definition of length and angle.)

- 14.5 Show that two nonzero vectors  $x, y \in \mathbb{R}^n$  point in the same direction if and only if there exists a positive real number  $s$  such that  $x = sy$ . By “point in the same direction”, we mean that the angle between the vectors is 0. [Hint: one direction of this proof is complicated: you will need to examine the proof of the Cauchy-Schwarz inequality to determine when the inequality is an equality.]
- (15) What is the angle between  $c(t) = (t, t^2)$  and  $d(t) = (t, 2 - t^2)$  at the point  $(1, 1)$ ? Express your solution in degrees. Part of this problem is to figure out what is meant by the angle. Hint: it has something to do with tangent vectors. Drawing a picture might help.
- (16) Let  $f: \mathbb{R} \rightarrow \mathbb{R}^n$  and  $g: \mathbb{R} \rightarrow \mathbb{R}^n$  be two differentiable curves in  $\mathbb{R}^n$ .
- Show that  $(f(t) \cdot g(t))' = f'(t) \cdot g(t) + f(t) \cdot g'(t)$ , a sort of product rule for derivatives of dot products. You may assume the product rule from one-variable calculus.
  - Suppose that  $|f(t)| = r$  for some constant  $r \in \mathbb{R}$ . Prove that  $f(t) \cdot f'(t) = 0$ , and describe what this relation means for the motion of a particle whose position at time  $t$  is  $f(t)$ .
- (17) Sketch the set  $\{(x, y) \in \mathbb{R}^2 \mid (x, y) \cdot (1, 2) = n\}$  for  $n = 0, 1, 2, 3$ . (You should see lines perpendicular to  $(1, 2)$  with  $n$  regulating how far the lines have been shifted in the direction of  $(1, 2)$ .)
- (18) Which of the following are linear subspaces? Provide a proof (directly from the definition of a linear subspace) or a counter-example.
- $\{(x, y, z) \mid x + 2y + 3z = 0\}$ .
  - $\{(x, y) \mid x^2 + y^2 = 1\}$ .
  - $\{(x, y, z) \mid x \neq 0\}$ .
  - $\{s(1, 2, 3, 6) + t(1, 4, 3, 2) \mid s, t \in \mathbb{R}\}$ .
  - $\{s^3(1, 1) \mid s \in \mathbb{R}\}$ .
- (19) Let  $U$  and  $V$  be linear subspaces of  $\mathbb{R}^n$ . Define

$$U + V := \{u + v \mid u \in U, v \in V\}.$$

Prove that  $U + V$  is a linear subspace of  $\mathbb{R}^n$ . (Do not forget the subtle point of showing that  $U + V$  is nonempty. For instance, you can argue that  $\vec{0} \in U + V$ .)

- (20) (a) Is  $(-10, 6, 1, -7)$  in the span of  $(1, 3, 5, -2)$  and  $(4, 0, 3, 1)$ ? Prove or disprove.  
 (b) What is the dimension of the space spanned by  $(-10, 6, 1, -7)$ ,  $(1, 3, 5, -2)$ , and  $(4, 0, 3, 1)$ ? No proof is required, but you should explain your thinking.
- (21) Let  $W$  be a linear subspace of  $\mathbb{R}^n$ . Let  $p$  and  $q$  be points of  $\mathbb{R}^n$ . Show that  $p + W = q + W$  if and only if  $p - q \in W$ .
- (22) Write the equation for the hyperplane through the point  $(1, 2, 0, 4)$  and perpendicular to  $(1, 2, 7, -3)$ .
- (23) Give an equation for a hyperplane perpendicular to  $(3, 5, 2, 1)$  and passing through the point  $(6, 3, 4, 9)$ .
- (24) Give an equation for a (hyper)plane perpendicular to  $(4, 7, -2)$  and passing through the point  $(5, 7, 1)$ .
- (25) Theorem 3.10 shows that hyperplanes in  $\mathbb{R}^n$  are exactly sets of the form

$$H = \{x \in \mathbb{R}^n \mid (x - p) \cdot a = 0\}.$$

for some fixed  $a \neq \vec{0}$  and  $p$  in  $\mathbb{R}^n$ . Use the trick in Example 3.11 to prove that hyperplanes in  $\mathbb{R}^n$  are exactly subsets of the form

$$H = \{(x_1, \dots, x_n) \in \mathbb{R}^n \mid a_1x_1 + \dots + a_nx_n = d\}.$$

for some fixed  $(a_1, \dots, a_n) \neq \vec{0}$  and  $d \in \mathbb{R}$ . (There are two parts to this exercise: if  $H$  has this form, use the trick and the theorem to show that  $H$  is a hyperplane. Conversely, if  $H$  has the form given in the theorem, find  $(a_1, \dots, a_n)$  and  $d$ .)

- (26) Which of the following functions are linear? No proof is necessary.
- $L(x, y) = (x + y, 2x - 3y)$ .
  - $L(t) = 2t$ .
  - $L(x, y, z) = xyz$ .
  - $L(x, y, z) = (x + 2y, z, 3x - y + 2z)$ .
  - $L(x, y) = (x^2 + 2y, \cos(y))$ .
- (27) Prove that  $L(x, y) = (3x - 5y, x + 4y)$  is a linear function.
- (28) Prove that  $L(x, y) = (2x + y + 4, 3x - 6y + 2)$  is not linear. (In other words, give a simple counter-example.)
- (29) Prove that  $L(x, y, z) = 3x + 2y - 5z$  is a linear function.
- (30) Prove that  $L(x, y) = (x^2, x + 3y)$  is not a linear function. (In other words, give a simple counterexample.)
- (31) Let  $L: \mathbb{R}^n \rightarrow \mathbb{R}^m$  be a linear function. Let  $\vec{0}_n$  and  $\vec{0}_m$  denote the zero vectors in  $\mathbb{R}^n$  and  $\mathbb{R}^m$ , respectively. Show that  $L(\vec{0}_n) = \vec{0}_m$  in two different ways: the first using the fact that  $L$  preserves addition, and the second using the fact that  $L$  preserves scalar multiplication.
- (32) (a) Prove that  $L: \mathbb{R} \rightarrow \mathbb{R}^m$  is a linear function if and only if there is a vector  $v \in \mathbb{R}^m$  such that  $L(x) = xv$ . In other words, linear functions with domain  $\mathbb{R}$  are exactly given by scaling a fixed vector.
- (b) Prove that  $L: \mathbb{R}^n \rightarrow \mathbb{R}$  is a linear function if and only if there is a vector  $v \in \mathbb{R}^n$  such that  $L(x) = v \cdot x$ . In other words, a linear functions with codomain  $\mathbb{R}$  are exactly given by taking the dot product with a fixed vector.

(33) Calculate:

(a)

$$\begin{pmatrix} 1 & 2 & 0 \\ -1 & 3 & 2 \\ 4 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 3 & 2 \\ -2 & 5 \end{pmatrix}$$

(b)

$$3 \begin{pmatrix} 2 & 3 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 3 \\ -1 & 2 \end{pmatrix} + 4 \begin{pmatrix} 1 & 5 & 0 \\ 3 & 2 & 3 \end{pmatrix} \begin{pmatrix} 1 & 5 \\ 3 & 0 \\ 4 & 2 \end{pmatrix}$$

(c)

$$\begin{pmatrix} 1 \\ 3 \\ 7 \end{pmatrix} (6)$$

(d)

$$(1 \ 2 \ 3) \begin{pmatrix} 4 & 2 \\ 6 & -3 \\ 0 & 8 \end{pmatrix} + 4(3 \ 2)$$

- (34) Let  $L(x, y) = (5x + 2y, x)$  and  $M(x, y) = (x - 2y, 5x + 4y)$ .
- Find the matrices  $A_L$  for  $L$  and  $A_M$  for  $M$ .
  - $(L + M)(x, y) = ?$
  - Check that the matrix for  $L + M$  is  $A_L + A_M$ .
  - $4L(x, y) = ?$ .
  - Check that the matrix for  $4L$  is  $4A_L$ .
- (35) Which linear maps correspond to the given matrices?
- $\begin{pmatrix} 2 & 3 \\ -1 & 4 \end{pmatrix}$
  - $\begin{pmatrix} 2 & 1 \\ 3 & 2 \\ 5 & 0 \end{pmatrix}$
  - $\begin{pmatrix} 1 & 2 & 3 \end{pmatrix}$
  - $\begin{pmatrix} 1 \\ -1 \\ 0 \\ 2 \end{pmatrix}$
- (36) Let  $L(x, y) = (2x + 3y, x + y, 3y)$  and  $M(x, y, z) = (x + 2y - z, x + z)$ .
- Find the matrices  $A_L$  for  $L$  and  $A_M$  for  $M$ .
  - $M \circ L(x, y) = ?$
  - Check that the matrix for  $M \circ L$  is the product  $A_M A_L$ .
- (37) Suppose that  $L: \mathbb{R}^2 \rightarrow \mathbb{R}^4$  is a linear function and that  $L(1, 0) = (3, 2, 0, 5)$  and  $L(0, 1) = (-6, 3, 2, 9)$ . What is  $L(4, -3)$ ?
- (38) Give a parametric equation for the plane spanned by  $(6, 2, -4, 1)$  and  $(3, 5, 1, 7)$ .
- (39) Give a parametric equation for the line passing through  $(1, 4, 2)$  and  $(7, 3, 0)$ .
- (40) Give a parametric equation for the line through the points  $(1, 4, 5)$  and  $(3, -1, 2)$ .
- (41) Let  $L(x, y, z) = (2x - 4y + z, 3x + 7z)$  and  $M(x, y, z) = (x + y + z, 3x + 5y + z)$ .
- Find the matrices,  $A_L$  for  $L$  and  $A_M$  for  $M$ .
  - $(L + M)(x, y, z) = ?$
  - Check that the matrix for  $L + M$  is  $A_L + A_M$ .
- (42) Let  $L(x, y) = (ax + by, cx + dy)$  and  $M(x, y) = (a'x + b'y, c'x + d'y)$  be arbitrary linear functions, i.e., the coefficients,  $a, a', \dots, d, d'$  are arbitrary real numbers.
- Find the matrices,  $A_L$  for  $L$  and  $A_M$  for  $M$ .
  - $M \circ L(x, y) = ?$
  - Check that the matrix for  $M \circ L$  is the product  $A_M A_L$ .
- (43) Let  $f(x, y, z) = (x + 2y + z + x^2 - y^2, 2x - y + 3z + 2x^2 + xy + z^3)$ .
- Calculate  $Jf(0, 0, 0)$ , the Jacobian of  $f$  at the point  $(0, 0, 0)$ .
  - Calculate  $Df_{(0,0,0)}$ , the linear map associated with  $Jf(0, 0, 0)$ .
  - Do you see in what sense  $Df_{(0,0,0)}$  gives just the “linear terms”  $f$ ?
- (44) Let  $f: \mathbb{R}^2 \rightarrow \mathbb{R}^3$  be defined by  $f(x, y) = (x + y, x^2 + 3xy, -4y + xy)$ , and let  $g: \mathbb{R}^3 \rightarrow \mathbb{R}$  be defined by  $g(x, y, z) = x^2 + 3y - yz$ .
- Calculate the Jacobian matrix of  $f$  at  $(1, 2)$ , i.e.,  $Jf(1, 2)$ .
  - Calculate the derivative of  $f$  at  $(1, 2)$ ,  $Df_{(1,2)}$ , as the linear function associated with the Jacobian matrix.
  - Calculate the Jacobian matrix of  $g$  at  $f(1, 2) = (3, 7, -6)$ , i.e.,  $Jg(3, 7, -6)$ .

- 
- (d) Calculate the derivative of  $g$  at  $(3, 7, -6)$ ,  $Dg_{(3,7,-6)}$ , as the linear function associated with the Jacobian matrix.
  - (e)  $g \circ f(x, y) = ?$
  - (f) Find the Jacobian matrix of  $g \circ f$  at the point  $(1, 2)$ , i.e.,  $J(g \circ f)(1, 2)$ .
  - (g) Calculate the derivative of  $g \circ f$  at  $(1, 2)$ ,  $Dg \circ f_{(1,2)}$ , as the linear function associated with the Jacobian matrix.
  - (h) Verify that  $J(g \circ f)(1, 2) = Jg(f(1, 2))Jf(1, 2)$ .
  - (i) Verify that  $D(g \circ f)_{(1,2)} = Dg_{f(1,2)} \circ Df_{(1,2)}$  by composing the derivatives from parts (b) and (d).



# The derivative

## 1. Introduction

The purpose of this chapter is to finally give the definition of the derivative and prove that the derivative is the linear function associated with the Jacobian matrix. The definition will make it immediately clear in what sense the derivative is a good linear approximation of the function. It is more difficult to see the connection with the Jacobian matrix, which we do by using the chain rule. The chain rule appears to be a merely technical result relating the derivative of a composition of functions to the derivatives of each function separately, but it turns out to be one of the most meaningful and useful results in multivariate calculus. To understand the definition of the derivative requires understanding limits of multivariate functions and a bit of topology of  $\mathbb{R}^n$ , which we cover first.

## 2. Topology

In this section, we introduce the metric topology on  $\mathbb{R}^n$ , starting with a generalization of the usual solid ball in  $\mathbb{R}^3$ .

**Definition 2.1.** *Let  $r > 0$  and  $p \in \mathbb{R}^n$ , The open ball of radius  $r$  centered at  $p$  is the set*

$$B_r(p) := \{x \in \mathbb{R}^n \mid d(x, p) < r\} = \{x \in \mathbb{R}^n \mid |x - p| < r\}.$$

*The closed ball of radius  $r$  centered at  $p$  is the set*

$$\bar{B}_r(p) := \{x \in \mathbb{R}^n \mid d(x, p) \leq r\}.$$

*The outer shell of the closed ball is the sphere of radius  $r$  centered at  $p$ :*

$$S_r(p) := \{x \in \mathbb{R}^n \mid d(x, p) = r\}.$$

Open balls in  $\mathbb{R}$  are open intervals. For example  $B_1(5) \subset \mathbb{R}$  is the interval  $(4, 6)$ . A closed ball in  $\mathbb{R}$  is a closed interval, e.g.,  $\bar{B}_1(5) = [4, 6]$ . In the plane, open balls are solid discs, without their circular boundaries. Closed balls in  $\mathbb{R}^2$  are discs, including their boundaries:



In  $\mathbb{R}^3$ , balls are solid spheres. In  $\mathbb{R}^4$ , we meet something new: a ball in four dimensions. The general open ball centered at the origin has the form

$$B_r = \{(t, x, y, z) \in \mathbb{R}^4 \mid t^2 + x^2 + y^2 + z^2 < r^2\}.$$

To picture this set, you could think of  $t$  as representing time. At a given time  $t$ , our glimpse of the ball is the set of points  $(x, y, z)$  such that  $x^2 + y^2 + z^2 < r^2 - t^2$ . The ball itself, then, is a movie. At time  $t = 0$ , we see a ball of radius  $r$  in  $\mathbb{R}^3$ , and as time moves forward, the ball shrinks (having radius  $\sqrt{r^2 - t^2}$  at time  $t$ ), finally disappearing when  $t = r$ .

Open balls are “fuzzy” on the outside. Near any point in an open ball, you can travel in any direction, maybe only a small way, and still remain in the ball. A general set of this type is called “open.”

**Definition 2.2.** *A set  $U \subseteq \mathbb{R}^n$  is open if it contains an open ball about each of its points; that is, for each  $p \in U$ , there is a real number  $r > 0$ , which may depend on  $p$ , such that  $B_r(p) \subseteq U$ .*

**Example 2.3.**

1. Every open ball  $B_r(p)$  in  $\mathbb{R}^n$  is open. To see this, take any point  $q$  in the ball. We need to find a ball  $B_s(q)$  about  $q$ , completely contained inside  $B_r(p)$ . Take  $s = r - d(p, q)$ , the distance from  $q$  to the boundary of the ball. We then have

$$\begin{aligned} q' \in B_s(q) &\Rightarrow d(q, q') < s = r - d(p, q) \\ &\Rightarrow d(p, q) + d(q, q') < r \\ &\Rightarrow d(p, q') < r \\ &\Rightarrow q' \in B_r(p). \end{aligned}$$

The key step is the triangle inequality for the implication on the third line. As a special case, this says that open intervals in  $\mathbb{R}$  are open sets.

2. The set  $\mathbb{R}^n$ , itself, is open. For instance, if  $p \in \mathbb{R}^n$ , then  $B_1(p) \subset \mathbb{R}^n$ . It would do no harm to think of  $\mathbb{R}^n$  as an open ball itself, with radius  $r = \infty$ .
3. The empty set  $\emptyset \subset \mathbb{R}^n$  is open. If it weren't, there would need to be a point in  $\emptyset$  such that no open ball about that point is completely contained inside  $\emptyset$ . That cannot happen since there are no points in  $\emptyset$  to begin with.
4. Removing a single point from  $\mathbb{R}^n$  leaves an open set. We'll leave this as an exercise. In fact, removing any finite number of points leaves an open set.
5. The half-open interval,  $[0, 1)$ , is not open. The problem is that every open ball about the point 0 will contain points that are negative and hence outside of the interval.
6. The set consisting of a single point  $p \in \mathbb{R}^n$  is not open. Every open ball about  $p$  will contain points not equal to  $p$ .

**Proposition 2.4.**

1. Let  $I$  be an arbitrary index set, and let  $U_\alpha$  be an open subset of  $\mathbb{R}^n$  for each  $\alpha \in I$ . Then their union,  $\cup_{\alpha \in I} U_\alpha$ , is open.
2. Let  $k$  be a positive integer. If  $W_1, \dots, W_k$  are open subsets of  $\mathbb{R}^n$ , then so is their intersection,  $\cap_{i=1}^k W_i$ .

PROOF: Let  $p \in \cup_{\alpha \in I} U_\alpha$ . Since  $p$  is in the union, it is in  $U_\beta$  for some  $\beta \in I$ , and since  $U_\beta$  is open, there is an  $r > 0$  such that  $B_r(p) \subseteq U_\beta$ . It follows that

$$B_r(p) \subseteq U_\beta \subseteq \cup_{\alpha \in I} U_\alpha,$$

as required.

For the second part of the proposition, take  $p \in \cap_{i=1}^k W_i$ . Since each  $W_i$  is open, there is an  $r_i > 0$  such that  $B_{r_i}(p) \subseteq W_i$  for each  $i = 1, \dots, k$ . Define  $r = \min\{r_1, \dots, r_k\}$ , then  $B_r(p) \subseteq B_{r_i}(p) \subseteq W_i$  for each  $i$ ; hence,  $B_r(p) \subseteq \cap_{i=1}^k W_i$ , as required.  $\square$

The collection of open sets in  $\mathbb{R}^n$  which we have just defined forms what is called a *topology* on  $\mathbb{R}^n$ .

**Definition 2.5.** Let  $S$  be a set. A topology on  $S$  is a collection of subsets  $\tau$  of  $S$  satisfying the following four axioms:

1.  $\emptyset \in \tau$ ;
2.  $S \in \tau$ ;
3. If  $U_\alpha \in \tau$  for all  $\alpha$  in some index set  $I$ , then  $\cup_{\alpha \in I} U_\alpha \in \tau$ ;
4. Let  $k$  be a positive integer. If  $W_1, \dots, W_k$  are in  $\tau$ , then so is  $\cap_{i=1}^k W_i$ .

The elements of  $\tau$  are called the open sets of the topology. A set with a topology is called a topological space.

We paraphrase the last two parts of the definition by saying that  $\tau$  is closed under arbitrary unions and finite intersections. Looking back in this section, you will see that the collection of open sets we have defined for  $\mathbb{R}^n$  forms a topology. It is called a metric topology, having been induced from a notion of distance.

**2.1. Limit points, closed sets.** The limit points of a set are any points which can be approximated arbitrarily closely by points within the set.

**Definition 2.6.** Let  $S \subseteq \mathbb{R}^n$ . A point  $p \in \mathbb{R}^n$  is a limit point of  $S$  if every open set containing  $p$  contains a point of  $S$  besides  $p$ .

Equivalently,  $p \in \mathbb{R}^n$  is a limit point of  $S$  if every open ball about  $p$  contains a point of  $S$  besides  $p$ .

**Example 2.7.**

1. Every point of an open set is a limit point of that open set.
2. The limit points of the open interval  $S = (0, 1) \subset \mathbb{R}$  are all the points in the closed interval  $[0, 1]$ .
3. In  $\mathbb{R}^2$ , let  $S = B_1(0, 0) \cup \{(10, 0)\}$ , a unit ball plus one isolated point. The limit points of  $S$  are then the points of the closed unit ball  $\bar{B}_1(0, 0)$ .
4. Let  $S = \{\frac{1}{n} \mid n \text{ a positive integer}\}$ . Then  $S$  has only one limit point: 0.
5. Let  $S = \mathbb{Q} \subset \mathbb{R}$  be the set of rational numbers. Then every real number is a limit point of  $\mathbb{Q}$ .

**Definition 2.8.** A subset  $C \subset \mathbb{R}^n$  is closed if its complement,  $C^c = \mathbb{R}^n \setminus C$  is open.

**Example 2.9.**

1. The closed ball  $\bar{B}_r(p) \subset \mathbb{R}^n$  is closed. Given any point  $q \in \mathbb{R}^n$  which is not in  $\bar{B}_r(p)$ , define  $s = d(q, p) - r$ . Then  $s > 0$  and  $B_s(q)$  is completely contained in the complement of  $\bar{B}_r(p)$ . This shows that the complement of  $\bar{B}_r(p)$  is open, i.e.,  $\bar{B}_r(p)$  is closed. In particular, we have shown that a closed interval on the real number line is a closed set.
2. A similar argument shows that a set consisting of a single point  $p \in \mathbb{R}^n$  is closed.
3. Arbitrary intersections of closed sets and finite unions of closed sets are closed (exercise). For instance, any finite set of points in  $\mathbb{R}^n$  is closed.
4. “Closed” does not mean “not open.” There are sets which are neither open nor closed, for instance,  $(0, 1] \subset \mathbb{R}$ . There are exactly two sets of  $\mathbb{R}^n$  that are simultaneously open and closed. Can you find them?

**Proposition 2.10.** *A subset  $C \subset \mathbb{R}^n$  is closed if and only if it contains all of its limit points.*

PROOF: First suppose that  $C$  is closed and that  $p$  is a point outside of  $C$ . Since  $p$  is in the complement of  $C$ , an open set, there is an open ball about  $p$  that does not intersect  $C$ . So  $p$  is not a limit point of  $C$ . This shows that  $C$  contains all of its limit points.

Conversely, suppose that  $C$  contains all of its limit points, and let  $p$  be a point outside of  $C$ . Since  $p$  is not a limit point of  $C$ , there is an open ball about  $p$  which is completely contained in the complement of  $C$ . This shows that the complement of  $C$  is open, i.e.,  $C$  is closed.  $\square$

### 3. Limits, continuity

**Definition 3.1.** *Let  $S \subseteq \mathbb{R}^n$ , and let  $f: S \rightarrow \mathbb{R}^m$  be a function. Let  $s$  be a limit point of  $S$ . The limit of  $f$  as  $x$  approaches  $s$  is  $p \in \mathbb{R}^m$ , denoted  $\lim_{x \rightarrow s} f(x) = p$ , if for all  $\varepsilon > 0$  there is a  $\delta > 0$  such that*

$$d(f(x), p) < \varepsilon$$

whenever

$$x \in S \quad \text{and} \quad 0 < d(x, s) < \delta.$$

Just as in one-variable calculus,  $\lim_{x \rightarrow s} f = p$  roughly says that at points close to  $s$ , the function  $f$  is close to  $p$ . If you want the function to be within a distance of  $\varepsilon$  of  $p$ , you just need to stay within a suitable distance  $\delta$  of  $s$ . However, note that the definition does not care what happens when  $d(x, s) = 0$ , i.e., when  $x = s$ . In other words, the value of  $f$  at  $s$  is irrelevant;  $f$  need not even be defined at  $s$  (this case arises in the definition of the derivative, for instance). The number  $\delta$  may depend on  $\varepsilon$  and on  $s$ . The requirement that  $x \in S$  in the last line of the definition is just so that  $f(x)$  is defined.

**Theorem 3.2.** *Let  $f: S \rightarrow \mathbb{R}^m$  and  $g: S \rightarrow \mathbb{R}^m$  where  $S$  is a subset of  $\mathbb{R}^n$ .*

1. *The limit of a function is unique: if  $\lim_{x \rightarrow s} f = p$  and  $\lim_{x \rightarrow s} f = q$ , then  $p = q$ .*
2. *The limit  $\lim_{x \rightarrow s} f(x)$  exists if and only if the corresponding limits for each of the component functions,  $\lim_{x \rightarrow s} f_i(x)$ , exists. In that case,*

$$\lim_{x \rightarrow s} f(x) = \left( \lim_{x \rightarrow s} f_1(x), \dots, \lim_{x \rightarrow s} f_m(x) \right).$$

3. Define  $f + g: S \rightarrow \mathbb{R}^m$  by  $(f + g)(x) := f(x) + g(x)$ . If  $\lim_{x \rightarrow s} f(x) = a$  and  $\lim_{x \rightarrow s} g(x) = b$ , then  $\lim_{x \rightarrow s} (f + g)(x) = a + b$ . Similarly, if  $t \in \mathbb{R}$ , define  $tf: S \rightarrow \mathbb{R}^m$  by  $(tf)(x) := t(f(x))$ . If  $\lim_{x \rightarrow s} f(x) = a$ , then  $\lim_{x \rightarrow s} (tf)(x) = ta$ .
4. If  $m = 1$ , define  $(fg)(x) := f(x)g(x)$  and  $(f/g)(x) := f(x)/g(x)$  (provided  $g(x) \neq 0$ ). If  $\lim_{x \rightarrow s} f(x) = a$  and  $\lim_{x \rightarrow s} g(x) = b$ , then  $\lim_{x \rightarrow s} (fg)(x) = ab$  and, if  $b \neq 0$ , then  $\lim_{x \rightarrow s} (f/g)(x) = a/b$ .
5. If  $m = 1$  and  $g(x) \leq f(x)$  for all  $x$ , then  $\lim_{x \rightarrow s} g(x) \leq \lim_{x \rightarrow s} f(x)$  provided these limits exist.
6. Let  $h: \mathbb{R}^n \rightarrow \mathbb{R}$  and  $m = 1$ . Suppose that  $g(x) \leq h(x) \leq f(x)$  for all  $x \in S$ . If the limits of  $g(x)$  and  $f(x)$  as  $x$  approaches  $s$  exist and are equal, then the limit of  $h(x)$  as  $x$  approaches  $s$  exists and is equal to that of  $f$  and  $g$ . This is called the squeeze principle.

PROOF: We will prove part 2, and leave the rest as exercises. First assume that  $\lim_{x \rightarrow s} f = p$ . Given  $\varepsilon > 0$ , take  $\delta$  so that  $d(f(x), p) < \varepsilon$  whenever  $x \in S$  and  $0 < d(x, s) < \delta$ . Then, for each  $i = 1, \dots, n$ , we have

$$\varepsilon > d(f(x), p) = |f(x) - p| = \left| \sqrt{\sum_{i=1}^m (f_i(x) - p_i)^2} \right| \geq |f_i - p_i|$$

provided  $x \in S$  and  $0 < d(x, s) < \delta$ . This shows that  $\lim_{x \rightarrow s} f_i = p_i$  for each  $i$ . Conversely, assume  $\lim_{x \rightarrow s} f_i = p_i$  for each  $i$ , and take  $\varepsilon > 0$ . Take  $\delta_i > 0$  so that  $d(f_i(x), p_i) < \varepsilon/\sqrt{m}$  whenever  $x \in S$  and  $0 < d(x, s) < \delta_i$ . Define  $\delta = \min\{\delta_1, \dots, \delta_n\}$ . Then, if  $x \in S$  and  $0 < d(x, s) < \delta$ , it follows that

$$d(f(x), p) = \left| \sqrt{\sum_{i=1}^m (f_i(x) - p_i)^2} \right| \leq \left| \sqrt{\sum_{i=1}^m (\varepsilon/\sqrt{m})^2} \right| = \varepsilon.$$

Hence,  $\lim_{x \rightarrow s} f = p$ . □

**Definition 3.3.** Let  $f: S \rightarrow \mathbb{R}^m$  with  $S \subseteq \mathbb{R}^n$ , and let  $s \in S$ . Then  $f$  is continuous at  $s$  if for all  $\varepsilon > 0$ , there exists a  $\delta > 0$  such that  $d(f(x), f(s)) < \varepsilon$  whenever  $x \in S$  and  $d(x, s) < \delta$ . The function  $f$  is continuous on  $S$  if  $f$  is continuous at all  $s \in S$ .

Unlike with the definition of a limit, the value of  $f$  at the point in question is crucial for continuity. You should check that if  $s$  is a limit point of  $S$ , then  $f$  is continuous at  $s$  if and only if  $\lim_{x \rightarrow s} f(x) = f(s)$ . On the other hand, if  $s$  is not a limit point, then  $f$  is automatically continuous at  $s$ . The former case is the only one that arises when  $S$  is an open subset of  $\mathbb{R}^n$ .

**Theorem 3.4.** Let  $f: S \rightarrow \mathbb{R}^m$  and  $g: S \rightarrow \mathbb{R}^m$  where  $S$  is a subset of  $\mathbb{R}^n$ .

1. The function  $f$  is continuous if and only if the inverse image of every open subset of  $\mathbb{R}^m$  under  $f$  is the intersection of an open subset of  $\mathbb{R}^n$  with  $S$ . In other words, if and only if for each open set  $U \subseteq \mathbb{R}^m$ , the set  $f^{-1}(U) := \{s \in S \mid f(s) \in U\}$  has the form  $W \cap S$  for some open set  $W \subset \mathbb{R}^n$ .
2. The function  $f$  is continuous at  $s$  if and only if each of its component functions is continuous at  $s$ .
3. The composition of continuous functions is continuous.
4. The functions  $f + g$  and  $tf$  for  $t \in \mathbb{R}$  defined as in Theorem 3.2 are continuous at  $s \in S$  provided  $f$  and  $g$  are continuous at  $s$ .

5. If  $m = 1$  and  $f$  and  $g$  are continuous at  $s \in S$ , then  $fg$  and  $f/g$  defined as in Theorem 3.2 are continuous at  $s$  (provided  $g(s) \neq 0$  in the latter case).
6. A function whose component functions are polynomials is continuous.

PROOF: The proof is left as an exercise. We can make  $S$  into a topological space by declaring a subset of  $S$  to be open if and only if it is the intersection of an open subset of  $\mathbb{R}^n$  with  $S$ . In that case, part 1 reads that  $f$  is continuous if and only if the inverse image of an open set is open. In fact, this elegant formulation is the usual *definition* of continuity in the case of a function between arbitrary topological spaces (not just subsets of Euclidean spaces). No  $\varepsilon$ 's or  $\delta$ 's!

Part 2 is similar to the analogous result for limits, proved above. Part 3 is most easily proved using part 1 and the fact that for any sets  $X$  and  $Y$  of  $\mathbb{R}^m$ ,  $f^{-1}(X \cap Y) = f^{-1}(X) \cap f^{-1}(Y)$ . Part 4 is straightforward, and hints to part 5 are given in the exercise at the end of this chapter pertaining to the analogous result for limits.

Part 6 is fairly easy to prove using the previous parts. To take a specific case, suppose you want to show that the function  $f(x, y) = (x^2y + 3xy^3, x - 4y^5, x^4 + 2xy - y^5)$  is continuous. By part 2, it suffices to show each of the three component functions is continuous. For instance, we would need to show that the function  $f_1(x, y) = x^2y + 3xy^3$  is continuous. By part 4, it suffices to show that the functions  $(x, y) \mapsto x^2y$  and  $(x, y) \mapsto 3xy^3$  are continuous. By part 5, we only need to show that each of the functions  $(x, y) \mapsto 3$ ,  $(x, y) \mapsto x$ , and  $(x, y) \mapsto y$  is continuous. First, a constant mapping is easily seen to be continuous by using part 1. Then you must show that a “projection” mapping such as  $(x, y) \mapsto x$  is continuous.  $\square$

#### 4. The definition of the derivative

By now, you know how to calculate the derivative as the linear function associated with the Jacobian matrix, and you have taken on faith that the derivative is a good linear approximation of a function at a point. In this section, we finally state the definition of the derivative. From the definition, it will be clear that the derivative gives a good linear approximation, but we will have to work to see that our calculation via the Jacobian matrix actually produces it.

**Definition 4.1.** Let  $f: U \rightarrow \mathbb{R}^m$  with  $U$  an open subset of  $\mathbb{R}^n$ . The function  $f$  is differentiable at  $p \in U$  if there is a linear function  $Df_p: \mathbb{R}^n \rightarrow \mathbb{R}^m$  such that

$$\lim_{h \rightarrow \vec{0}} \frac{|f(p+h) - f(p) - Df_p(h)|}{|h|} = 0.$$

The linear function  $Df_p$  is called the derivative of  $f$  at  $p$ . The function  $f$  is differentiable on  $U$  if it is differentiable at each point of  $U$ .

Observe that in the definition, we divide by the *length* of  $h$ , a real number. It would not make sense in general to divide by the vector  $h$ , itself.

**Example 4.2.** Let  $f(x, y) = (x - 2xy, y^2)$ . At the point  $(2, 3)$  you can readily calculate that the linear function associated with the Jacobian matrix  $Jf_{(2,3)}$  is given by  $L(x, y) = (-5x - 4y, 6y)$ . Let's check from the definition that  $L = Df_{(2,3)}$ , i.e., that  $L$  is the derivative

of  $f$  at  $(2, 3)$ . Plugging into the definition,

$$\begin{aligned}
& \lim_{(a,b) \rightarrow \vec{0}} \frac{|f((a, b) + (2, 3)) - f(2, 3) - L(a, b)|}{|(a, b)|} \\
&= \lim_{(a,b) \rightarrow \vec{0}} \frac{|((a+2) - 2(a+2)(b+3), (b+3)^2) - (-10, 9) - (-5a - 4b, 6b)|}{|(a, b)|} \\
&= \lim_{(a,b) \rightarrow \vec{0}} \frac{|(-2ab, b^2)|}{|(a, b)|} \\
&= \lim_{(a,b) \rightarrow \vec{0}} \frac{|b|\sqrt{4a^2 + b^2}}{\sqrt{a^2 + b^2}} \\
&\leq \lim_{(a,b) \rightarrow \vec{0}} \frac{|b|\sqrt{4a^2 + 4b^2}}{\sqrt{a^2 + b^2}} \\
&= \lim_{(a,b) \rightarrow \vec{0}} 2|b| = 0.
\end{aligned}$$

The result follows from the squeeze principle of Theorem 3.2. To verify the last step of the calculation, let  $\varepsilon > 0$  be given. If  $d((a, b), \vec{0}) < \varepsilon/2$ , we have

$$2|b| \leq 2\sqrt{a^2 + b^2} = 2|(a, b)| < \varepsilon.$$

Thus,  $\lim_{(a,b) \rightarrow \vec{0}} 2|b| = 0$ . This completes the verification that  $L = Df_{(2,3)}$ .

**4.1. One variable calculus.** If  $n = m = 1$ , then  $f$  is a function of one variable. Let us see how the new definition of the derivative of  $f$  compares with the usual definition from one variable calculus. We will take the usual definition to be that  $f$  is differentiable at  $p$  with derivative  $f'(p) \in \mathbb{R}$  if

$$\lim_{h \rightarrow 0} \frac{f(p+h) - f(p)}{h} = f'(p).$$

If  $f$  is differentiable in the new sense, then its derivative is a linear function of the form  $Df_p: \mathbb{R} \rightarrow \mathbb{R}$ . Any linear function with domain and codomain  $\mathbb{R}$  is simply multiplication by a constant (cf. exercises for Chapter 3). In other words, there is a real number  $\alpha$  such that  $Df_p(x) = \alpha x$ . In this case, it turns out that  $f$  is differentiable in the old sense of one variable calculus and  $f'(p) = \alpha$ . Conversely, if  $f$  is differentiable in the old sense, then it is differentiable in the new sense and the derivative is the linear function formed by multiplication by the real number  $f'(p)$ , i.e.,  $Df_p(x) = f'(p)x$ . This is verified by the following computation.

$$\begin{aligned}
0 &= \lim_{h \rightarrow 0} \frac{|f(p+h) - f(p) - Df_p(h)|}{|h|} \\
&= \lim_{h \rightarrow 0} \frac{|f(p+h) - f(p) - \alpha h|}{|h|} \\
&= \lim_{h \rightarrow 0} \left| \frac{f(p+h) - f(p)}{h} - \alpha \right| \\
&\Leftrightarrow \lim_{h \rightarrow 0} \frac{f(p+h) - f(p)}{h} = \alpha.
\end{aligned}$$

To repeat, our two definitions of the derivative in the case where  $n = m = 1$  agree provided one identifies the real number  $f'(p)$  with the linear function  $Df_p(x) = f'(p)x$ .

## 5. The best affine approximation revisited

Let  $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$  be differentiable at a point  $p \in \mathbb{R}^n$ . In Chapter 1, we defined the best affine approximation of  $f$  at  $p$  to be the affine function

$$Af_p(x) = f(p) + Df_p(x),$$

and claimed that for values of  $x$  near the origin,  $Af_p(x)$  is a good approximation for  $f$  near  $p$ . We are now ready to make that statement precise.

The definition of the derivative says that

$$\lim_{h \rightarrow \vec{0}} \frac{|f(p+h) - f(p) - Df_p(h)|}{|h|} = 0.$$

Thus, as  $h \rightarrow \vec{0}$ , the numerator,  $|f(p+h) - f(p) - Df_p(h)|$ , goes to zero faster than the length,  $|h|$ . This means that for all points sufficiently close to  $p$ , that is, all points of the form  $p+h$  with  $h$  sufficiently close to  $\vec{0}$ , we have

$$f(p+h) \approx f(p) + Df_p(h) \tag{1}$$

On the right hand side of the equality we have  $Af_p(h)$ , and we have just shown that if  $h$  is close to the origin, then  $Af_p(h)$  is approximately  $f(p+h)$ , the value of  $f$  at a certain point near  $p$ .\*

The definition of the derivative shows that as  $h \rightarrow \vec{0}$ , the function  $Af_p(h)$  approaches  $f(p+h)$  faster than the length of  $h$  approaches 0. We say that  $Af_p$  is a good *first order approximation* of  $f$  near  $p$ .<sup>†</sup> Why is  $Af_p$  *the best* affine approximation? Suppose there were another affine function approximating  $f$ . It would have the form  $B(x) = q + L(x)$  for some  $q \in \mathbb{R}^m$  and some linear function  $L: \mathbb{R}^n \rightarrow \mathbb{R}^m$ . If  $B(x)$  were a good first order approximation, taking values close to  $f(p)$  for values of  $x$  close to  $\vec{0}$ , we would have

$$\lim_{h \rightarrow \vec{0}} \frac{|f(p+h) - B(h)|}{|h|} = 0,$$

just as we have for  $Af_p$ . It follows that

$$\begin{aligned} \lim_{h \rightarrow \vec{0}} \frac{|Af_p(h) - B(h)|}{|h|} &= \lim_{h \rightarrow \vec{0}} \frac{|(f(p+h) - B(h)) - (f(p+h) - Af_p(h))|}{|h|} \\ &\leq \lim_{h \rightarrow \vec{0}} \frac{|(f(p+h) - B(h))|}{|h|} + \lim_{h \rightarrow \vec{0}} \frac{|(f(p+h) - Af_p(h))|}{|h|} \\ &= 0 + 0 = 0. \end{aligned}$$

\*In Chapter 1, we also defined a shifted version of the best affine approximation,  $Tf_p(x) = f(p) + Df_p(x-p)$ . Our argument has shown in what sense  $Tf_p$  is an approximation for  $f$  for values near  $p$  (substituting  $x$  for  $p+h$  in the displayed equation 1 shows that  $f(x) \approx f(p) + Df_p(x-p)$ ).

<sup>†</sup>Later in the notes, we will consider multivariate Taylor polynomials. They give *higher order* (more accurate) approximations.

Thus,  $\lim_{h \rightarrow \vec{0}} |Af_p(h) - B(h)|/|h| = 0$ . We will show that  $Af_p(x) = B(x)$  for all points  $x \in \mathbb{R}^n$ . Take  $x \in \mathbb{R}^n$  with  $x \neq \vec{0}$ . From what we have already shown it follows that

$$\begin{aligned} 0 &= \lim_{t \rightarrow 0} \frac{|Af_p(tx) - B(tx)|}{|tx|} \\ &= \lim_{t \rightarrow 0} \frac{|(f(p) + Df_p(tx)) - (q + L(tx))|}{|tx|} \\ &= \lim_{t \rightarrow 0} \left| \frac{f(p) - q}{t|x|} + \frac{Df_p(tx) - L(tx)}{t|x|} \right| \\ &= \lim_{t \rightarrow 0} \left| \frac{f(p) - q}{t|x|} + \frac{Df_p(x) - L(x)}{|x|} \right|. \end{aligned}$$

The first step, that the limit is 0, is left as an exercise. The last step follows since  $Df_p$  and  $L$  are linear; so the scalar  $t$  factors out. Since  $f(p) - q$  and  $(Df_p(x) - L(x))/|x|$  do not depend on  $t$ , the only way the limit can be zero is if

$$f(p) = q \quad \text{and} \quad \frac{Df_p(x) - L(x)}{|x|} = 0.$$

Multiplying through by  $|x|$  in the latter equality yields  $Df_p(x) = L(x)$ , which holds when  $x = \vec{0}$ , as well, since  $Df_p$  and  $L$  are linear. Hence,  $B(x) = Af_p(x)$  for all  $x \in \mathbb{R}^n$ , and in this sense there is only one good first order affine approximation to  $f$  near  $p$ .

A slight rephrasing of the argument we just gave shows that derivatives are unique.

**Theorem 5.1.** *Let  $f: U \rightarrow \mathbb{R}^m$  with  $U$  an open subset of  $\mathbb{R}^n$ . Let  $p \in U$ . There is at most one linear function  $L: \mathbb{R}^n \rightarrow \mathbb{R}^m$  such that*

$$\lim_{h \rightarrow \vec{0}} \frac{|f(p+h) - f(p) - L(h)|}{|h|} = 0.$$

## 6. The chain rule

The chain rule is one of the most important results in calculus. It says, roughly, that the derivative of a composition of functions is the composition of the corresponding derivatives. It is the result that allows us to quickly compute the following derivatives in one variable calculus:

$$\begin{aligned} ((x^2 + 4x + 25)^{23})' &= 23(2x + 4)(x^2 + 4x + 25)^{22} \\ (\cos(x^4))' &= -4x^3 \sin(x^4). \end{aligned}$$

Here is the statement of the chain rule in several variables. Before proving it, we will look at a couple examples and prove a preliminary lemma.

**Theorem 6.1.** *(Chain rule) Let  $f: U \rightarrow \mathbb{R}^k$  and let  $g: V \rightarrow \mathbb{R}^m$  with  $U \subseteq \mathbb{R}^n$  and  $V \subseteq \mathbb{R}^k$  open subsets. Suppose that  $f(U) \subseteq V$  so that we can form the composition  $g \circ f: U \rightarrow \mathbb{R}^m$ . If  $f$  is differentiable at  $p \in U$  and  $g$  is differentiable at  $f(p) \in V$ , then  $g \circ f$  is differentiable at  $p$  and*

$$D(g \circ f)_p = Dg_{f(p)} \circ Df_p.$$

*In terms of Jacobian matrices,  $J(g \circ f)(p) = Jg(f(p))Jf(p)$ .*

**Example 6.2.** If  $f$  and  $g$  are functions from one variable calculus ( $k = m = n = 1$ ) then all the relevant Jacobian matrices are  $1 \times 1$  matrices containing the ordinary derivatives. For instance,  $Jf(x) = (f'(x))$ . Thus, in this case the chain rule says that  $(g \circ f)'(x) = g'(f(x))f'(x)$ . As a concrete example, take  $f(x) = x^2 + 4x + 25$  and  $g(x) = x^{23}$ . The chain rule says

$$((x^2 + 4x + 25)^{23})' = (g \circ f)'(x) = g'(f(x))f'(x) = 23(x^2 + 4x + 25)^{22}(2x + 4),$$

the example with which we began this section.

**Example 6.3.** For an example of the chain rule in several variables, take  $f(x, y) = x^2 + y + 2$  and  $g(t) = (t, t^2)$ . Then

$$(g \circ f)(x, y) = g(f(x, y)) = g(x^2 + y + 2) = (x^2 + y + 2, (x^2 + y + 2)^2).$$

Thus,  $(g \circ f)(x, y) = (x^2 + y + 2, x^4 + 2x^2y + 4x^2 + y^2 + 4y + 4)$ . The relevant Jacobian matrices are

$$Jf(x, y) = \begin{pmatrix} 2x & 1 \end{pmatrix}, \quad Jg(t) = \begin{pmatrix} 1 \\ 2t \end{pmatrix},$$

$$J(g \circ f)(x, y) = \begin{pmatrix} 2x & 1 \\ 4x^3 + 4xy + 8x & 2x^2 + 2y + 4 \end{pmatrix}.$$

Let  $p = (1, 2)$ ; so  $f(p) = f(1, 2) = 5$ . Evaluating the Jacobian matrices at the relevant points:

$$Jf(1, 2) = \begin{pmatrix} 2 & 1 \end{pmatrix}, \quad Jg(5) = \begin{pmatrix} 1 \\ 10 \end{pmatrix},$$

$$J(g \circ f)(1, 2) = \begin{pmatrix} 2 & 1 \\ 20 & 10 \end{pmatrix}.$$

The point of the chain rule is that we can calculate the Jacobian matrix of  $J(g \circ f)$  by multiplying the Jacobian matrices for  $g$  and  $f$ :

$$J(g \circ f)(1, 2) = \begin{pmatrix} 1 \\ 10 \end{pmatrix} \begin{pmatrix} 2 & 1 \end{pmatrix} = J(g \circ f)(1, 2) = \begin{pmatrix} 2 & 1 \\ 20 & 10 \end{pmatrix}.$$

We will need the following result to prove the chain rule. It says that the factor by which a linear function stretches the length of a vector is bounded, independent of the vector.

**Lemma 6.4.** *Let  $L: \mathbb{R}^n \rightarrow \mathbb{R}^m$  be a linear function. There exists a constant  $c_L \geq 0$  such that*

$$|L(x)| \leq c_L|x|$$

for all  $x \in \mathbb{R}^n$ .

PROOF: Let  $r_i$  be the  $i$ -th row of the  $m \times n$  matrix corresponding to  $L$ . Then the  $i$ -th component of  $L(x)$  is the dot product  $r_i \cdot x$ , i.e.,  $L(x) = (r_1 \cdot x, \dots, r_m \cdot x)$ . For instance, if  $L(x_1, x_2) = (2x_1 + 3x_2, 4x_1 - 5x_2)$ , then  $r_1 = (2, 3)$  and  $r_2 = (4, -5)$ . We would have, for

example,  $L_1(x_1, x_2) = r_1 \cdot (x_1, x_2) = 2x_1 + 3x_2$ . For general  $L$ , define  $c_L := \sum_{i=1}^m |r_i|$ . Then

$$\begin{aligned} |L(x)| &= |(r_1 \cdot x, \dots, r_m \cdot x)| = \left| \sum_{i=1}^m (r_i \cdot x) e_i \right| \\ &\leq \sum_{i=1}^m |(r_i \cdot x) e_i| = \sum_{i=1}^m |(r_i \cdot x)| |e_i| = \sum_{i=1}^m |(r_i \cdot x)| \\ &\leq \sum_{i=1}^m |r_i| |x| = c_L |x|. \end{aligned}$$

The first inequality is the triangle inequality, and the second is the Cauchy-Schwarz inequality. Also, recall that  $|e_i| = 1$ .  $\square$

We are now ready to prove the chain rule.

PROOF: (adapted from Spivak's *Calculus on Manifolds*) Define the functions

$$\begin{aligned} \phi(h) &= f(p+h) - f(p) - Df_p(h) \\ \psi(k) &= g(f(p)+k) - g(f(p)) - Dg_{f(p)}(k) \\ \rho(h) &= (g \circ f)(p+h) - (g \circ f)(p) - Dg_{f(p)} \circ Df_p(h). \end{aligned}$$

Since  $f$  is differentiable at  $p$  and  $g$  is differentiable at  $f(p)$ , it follows that

$$\lim_{h \rightarrow 0} \frac{|\phi(h)|}{|h|} = 0 \quad (2)$$

$$\lim_{k \rightarrow 0} \frac{|\psi(k)|}{|k|} = 0. \quad (3)$$

Our goal is to show that  $D(g \circ f)_p = Dg_{f(p)} \circ Df_p$ , in other words,

$$\lim_{h \rightarrow 0} \frac{|\rho(h)|}{|h|} = 0. \quad (4)$$

Using the definition of  $\phi(h)$  and the linearity of  $Dg_{f(p)}$ ,

$$\begin{aligned} \rho(h) &= g(f(p+h)) - g(f(p)) - Dg_{f(p)}(f(p+h) - f(p) - \phi(h)) \\ &= g(f(p+h)) - g(f(p)) - Dg_{f(p)}(f(p+h) - f(p)) + Dg_{f(p)}(\phi(h)). \end{aligned}$$

Letting  $k = f(p+h) - f(p)$ , we get

$$\begin{aligned} \rho(h) &= g(f(p)+k) - g(f(p)) - Dg_{f(p)}(k) + Dg_{f(p)}(\phi(h)) \\ &= \psi(k) + Dg_{f(p)}(\phi(h)). \end{aligned}$$

Hence, using the triangle inequality,

$$\frac{|\rho(h)|}{|h|} = \frac{|\psi(k) + Dg_{f(p)}(\phi(h))|}{|h|} \leq \frac{|\psi(k)|}{|h|} + \frac{|Dg_{f(p)}(\phi(h))|}{|h|}.$$

Thus, to show 4 it suffices to verify the following two statements:

- (a)  $\lim_{h \rightarrow 0} \frac{|\psi(k)|}{|h|} = 0$ .
- (b)  $\lim_{h \rightarrow 0} \frac{|Dg_{f(p)}(\phi(h))|}{|h|} = 0$ .

We first check (b). Since  $Dg_{f(p)}$  is a linear function, according to Lemma 6.4 there is a constant  $c \geq 0$  such that  $|Dg_{f(p)}(x)| \leq c|x|$  for all  $x \in \mathbb{R}^k$ . Therefore, using 2,

$$\begin{aligned} \lim_{h \rightarrow 0} \frac{|Dg_{f(p)}(\phi(h))|}{|h|} &\leq \lim_{h \rightarrow 0} \frac{c|\phi(h)|}{|h|} \\ &= c \lim_{h \rightarrow 0} \frac{|\phi(h)|}{|h|} \\ &= 0. \end{aligned}$$

The proof of (a) is a little tougher. Given  $\varepsilon > 0$ , equation 3 says that there exists a  $\delta_1 > 0$  such that

$$0 < |k| < \delta_1 \Rightarrow |\psi(k)| < \varepsilon|k|.$$

Since  $f$  is differentiable at  $p$ , it is continuous at  $p$  (exercise). Therefore, given  $\delta_1$ , there exists a  $\delta_2 > 0$  such that

$$|h| < \delta_2 \Rightarrow |f(p+h) - f(p)| < \delta_1.$$

Since  $k = f(p+h) - f(p)$ , this means that

$$0 < |h| < \delta_2 \Rightarrow |k| < \delta_1 \Rightarrow |\psi(k)| \leq \varepsilon|k|.$$

Thus, assuming  $0 < |h| < \delta_2$ ,

$$\begin{aligned} \frac{|\psi(k)|}{|h|} &\leq \frac{\varepsilon|k|}{|h|} \\ &= \frac{\varepsilon|f(p+h) - f(p)|}{|h|} \\ &= \frac{\varepsilon|\phi(h) + Df_p(h)|}{|h|} \\ &\leq \frac{\varepsilon|\phi(h)|}{|h|} + \frac{\varepsilon|Df_p(h)|}{|h|}. \end{aligned}$$

Equation 2 says that  $\lim_{h \rightarrow 0} |\phi(h)|/|h| = 0$ ; therefore, by taking  $\delta_2$  smaller, if necessary, we may assume that  $|\phi(h)|/|h| < 1$ . In addition, by the lemma, there is a  $c' \geq 0$  such that  $|Df_p(h)| \leq c'|h|$  for all  $h$ . It follows that for  $0 < |h| < \delta_2$ ,

$$\frac{|\psi(k)|}{|h|} \leq \varepsilon \frac{|\phi(h)|}{|h|} + \varepsilon \frac{|Df_p(h)|}{|h|} \leq \varepsilon + \varepsilon c' = \varepsilon(1 + c').$$

Since  $\varepsilon$  can be made arbitrarily small, we have shown that

$$\lim_{h \rightarrow 0} \frac{|\psi(k)|}{|h|} = 0,$$

as desired.

We know that every linear function corresponds to a matrix and that under this correspondence composition of functions corresponds to products of matrices. We have just shown that  $D(g \circ f)_p = Dg_{f(p)} \circ Df_p$ ; so there is a corresponding result for the matrices representing the derivatives. You already know that the matrix corresponding to a derivative is the Jacobian matrix, but we have not proved that yet. Once we prove that result in Section 8, we will have established that  $J(g \circ f)(p) = Jg(f(p))Jf(p)$ , as well.  $\square$

## 7. Partial derivatives

In Chapter 1 we quickly introduced partial derivatives as just ordinary derivatives with respect to one variable, pretending the other variables were constants. We now formalize this idea.

**Definition 7.1.** Let  $f: U \rightarrow \mathbb{R}^m$  with  $U$  an open subset of  $\mathbb{R}^n$ , and let  $e_i$  be the  $i$ -th standard basis vector for  $\mathbb{R}^n$ . The partial of  $f$  at  $p \in U$  with respect to the  $i$ -th variable is the vector in  $\mathbb{R}^m$

$$\frac{\partial f}{\partial x_i}(p) := \lim_{t \rightarrow 0} \frac{f(p + te_i) - f(p)}{t}$$

provided this limit exists.

If  $n = m = 1$ , we recover the usual definition of the derivative from one variable calculus. So in that case,  $\partial f / \partial x = df / dx$ . In any case, you can think of the partial derivative of a function  $f$  with respect to the  $i$ -th variable as the instantaneous change per unit time of  $f$  at  $p$  in the direction of  $e_i$ .

**Example 7.2.** Let  $f(x, y, z) = (x + y^2, yz)$ . The partial derivative of  $f$  at the point  $p = (1, 2, 3)$  with respect to the second variable is

$$\begin{aligned} \frac{\partial f}{\partial y}(p) &= \lim_{t \rightarrow 0} \frac{f((1, 2, 3) + t(0, 1, 0)) - f(1, 2, 3)}{t} \\ &= \lim_{t \rightarrow 0} \frac{f(1, 2 + t, 3) - f(1, 2, 3)}{t} \\ &= \lim_{t \rightarrow 0} \frac{(1 + (2 + t)^2, (2 + t)3) - (5, 6)}{t} \\ &= \lim_{t \rightarrow 0} \frac{(4t + t^2, 3t)}{t} \\ &= \lim_{t \rightarrow 0} (4 + t, 3) \\ &= (4, 3). \end{aligned}$$

It is important to understand that the definition really does just say to pretend that every variable except the  $i$ -th variable is constant, then take the derivative of each component function as if it were a function from one variable calculus. Thus, in the previous example, where  $f(x, y, z) = (x + y^2, yz)$ , the partial derivative with respect to  $y$  at a general point  $(x, y, z)$  is  $\partial f / \partial y = (2y, z)$ , without an explicit calculation of a limit. At  $(1, 2, 3)$ , we have  $\partial f / \partial y(1, 2, 3) = (4, 3)$ . To see this in detail for an arbitrary function  $f$ , first note we can take the partial of  $f$  with respect to the  $i$ -th variable by taking the corresponding partial of each component of  $f$ :

$$\begin{aligned} \lim_{t \rightarrow 0} \frac{f(p + te_i) - f(p)}{t} &= \lim_{t \rightarrow 0} \frac{(f_1(p + te_i) - f_1(p), \dots, f_m(p + te_i) - f_m(p))}{t} \\ &= \lim_{t \rightarrow 0} \left( \frac{(f_1(p + te_i) - f_1(p))}{t}, \dots, \frac{(f_m(p + te_i) - f_m(p))}{t} \right) \\ &= \left( \lim_{t \rightarrow 0} \frac{(f_1(p + te_i) - f_1(p))}{t}, \dots, \lim_{t \rightarrow 0} \frac{(f_m(p + te_i) - f_m(p))}{t} \right) \\ &= \left( \frac{\partial f_1}{\partial x_i}(p), \dots, \frac{\partial f_m}{\partial x_i}(p) \right). \end{aligned}$$

The third equality is Theorem 3.2, part 2. Now pick a component  $f_j$ , and define

$$g(x) = f_j(p_1, \dots, p_{i-1}, x, p_{i+1}, \dots, p_n)$$

for  $x$  sufficiently close to  $p_i$  (so that  $f_j$  makes sense at the given point). The function  $g$  is the ordinary function from one variable calculus you get by setting all variables of  $f_j$  equal to constants except the  $i$ -th. The derivative of  $g$  at  $x = p_i$  in the sense of one variable calculus is

$$\lim_{t \rightarrow 0} \frac{g(p_i + t) - g(p_i)}{t} = \lim_{t \rightarrow 0} \frac{f_j(p + te_i) - f_j(p)}{t} = \frac{\partial f_j}{\partial x_i}(p).$$

**7.1. Higher order partial derivatives.** Let  $f: U \rightarrow \mathbb{R}^m$  be a function with  $U$  an open subset of  $\mathbb{R}^n$ . Suppose that each partial derivative  $\partial f / \partial x_i$  exists at each point of  $U$ . The partial derivatives are then again functions with domain  $U$  and codomain  $\mathbb{R}^m$ ; so it makes sense to try to take their partial derivatives. If they exist, we could try to take their partials, and so on. Let  $x_{i_1}, \dots, x_{i_k}$  be a list of variables of  $f$ , possibly with repeats. Recursively define the  $k$ -th order partial derivative of  $f$  with respect to this list of variables by

$$\frac{\partial^k f}{\partial x_{i_1} \dots \partial x_{i_k}} := \frac{\partial^{k-1}}{\partial x_{i_1} \dots \partial x_{i_{k-1}}} \left( \frac{\partial f}{\partial x_{i_k}} \right).$$

In other words, first take the partial of  $f$  with respect to  $x_k$ , then take the partial of the resulting function with respect to  $x_{k-1}$ , etc.

**Example 7.3.** Let  $f(x, y) = (x^3 - 2xy^4 + y^2, xy^2)$ . To find the third order partial of  $f$  with respect to the list  $y, y, x$  we calculate

$$\frac{\partial f}{\partial x} = (3x^2 - 2y^4, y^2), \quad \frac{\partial^2 f}{\partial y \partial x} = (-8y^3, 2y), \quad \frac{\partial^3 f}{\partial y \partial y \partial x} = (-24y^2, 2).$$

Does the order in which you take the partial derivatives matter? For instance, in the previous example, does  $\partial^2 f / \partial y \partial x = \partial^2 f / \partial x \partial y$ ? If you do the calculation, you will see that in fact they are equal even though the two paths to the final result look quite different along the way.

Despite this example, in general, the order in which you take the derivatives *does* matter. The exercises at the end of the chapter will provide an example. However, it turns out that if the partial derivatives of order  $k$  exist and are continuous, then the order does not matter for any partial derivative of order less than or equal to  $k$ . The precise result is stated below.

**Theorem 7.4.** Let  $f: U \rightarrow \mathbb{R}^m$  with  $U$  an open subset of  $\mathbb{R}^n$ . Let  $x$  and  $y$  be any two variables for  $f$ . Suppose that  $\partial f / \partial x$ ,  $\partial f / \partial y$ , and  $\partial^2 f / \partial x \partial y$  exist on  $U$  and that  $\partial^2 f / \partial x \partial y$  is continuous at  $p \in U$ . Then  $\partial^2 f / \partial y \partial x$  exists at  $p$  and

$$\frac{\partial^2 f}{\partial x \partial y}(p) = \frac{\partial^2 f}{\partial y \partial x}(p).$$

PROOF: See Rudin's *Principles of Mathematical Analysis*, (Theorem 9.41 in the third edition). The key step in the proof is the mean value theorem. The proof is not very difficult, but is not central to what we are covering in these notes, so I won't include it here.  $\square$

For the most part, we will consider *smooth* functions whose mixed partials of all orders

exist. In that case, the order of taking partials is never important. To denote the partial derivative which is of order  $i_\ell$  with respect to  $x_\ell$  for each  $\ell = 1, \dots, n$ , we write

$$\frac{\partial^k f}{\partial x_1^{i_1} \dots \partial x_n^{i_n}}$$

where  $k = i_1 + \dots + i_n$ . For instance, in Example 7.3, we could write

$$\frac{\partial^3 f}{\partial y^2 \partial x} = (-24y^2, 2).$$

## 8. The derivative is given by the Jacobian matrix

In this section, we prove one of the key results of differential calculus: that the derivative of a function is the linear function associated with the matrix of partial derivatives known as the Jacobian matrix.

**Theorem 8.1.** *Let  $f: U \rightarrow \mathbb{R}^m$  with  $U$  an open subset of  $\mathbb{R}^n$ . If  $f$  is differentiable at  $p \in U$ , then each of the first partial derivatives of  $f$  exists, and  $Df_p$  is the linear function associated with the Jacobian matrix,*

$$Jf(p) := \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(p) & \frac{\partial f_1}{\partial x_2}(p) & \dots & \frac{\partial f_1}{\partial x_n}(p) \\ \frac{\partial f_2}{\partial x_1}(p) & \frac{\partial f_2}{\partial x_2}(p) & \dots & \frac{\partial f_2}{\partial x_n}(p) \\ \vdots & & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(p) & \frac{\partial f_m}{\partial x_2}(p) & \dots & \frac{\partial f_m}{\partial x_n}(p) \end{pmatrix}.$$

The  $j$ -th column of  $Jf_p$  is  $\partial f(p)/\partial x_j$ , the  $j$ -th partial derivative of  $f$ , thought of as a column vector.

You have been using this result since the beginning of the notes. For instance, if  $f: \mathbb{R}^2 \rightarrow \mathbb{R}^3$  with  $f(x, y) = (x^2 + y, 2xy, 3x^3 - x^2y)$ , then

$$Jf(x, y) = \begin{pmatrix} 2x & 1 \\ 2y & 2x \\ 9x^2 - 2xy & -x^2 \end{pmatrix}.$$

At the point  $p = (1, 1)$ ,

$$Jf(1, 1) = \begin{pmatrix} 2 & 1 \\ 2 & 2 \\ 7 & -1 \end{pmatrix}.$$

Our theorem then says that the derivative at  $(1, 1)$  is the associated linear function:  $Df_p(x, y) = (2x + y, 2x + 2y, 7x - y)$ .

Since the derivative is a linear function, we know it has a corresponding matrix. Our strategy for proving Theorem 8.1 is to isolate the  $i, j$ -th entry of this matrix by using a composition of functions, and use the chain rule to show that its value is  $\partial f_i(p)/\partial x_j$ .

Note that in our example  $Df_p(1, 0) = (2, 2, 7)$  and  $Df_p(0, 1) = (1, 2, -1)$ . In general, you know that if  $L: \mathbb{R}^n \rightarrow \mathbb{R}^m$  is any linear function, then  $L(e_j)$  is the  $j$ -th column of the matrix corresponding to  $L$ . To isolate the  $i$ -th entry in the  $j$ -th column, define the  $i$ -th projection mapping

$$\begin{aligned} \pi_i: \mathbb{R}^m &\rightarrow \mathbb{R} \\ (x_1, \dots, x_m) &\mapsto x_i \end{aligned}$$

Then  $(\pi_i \circ L)(e_j)$  is the  $i, j$ -th entry of the matrix corresponding to  $L$ . For instance, returning to our example,  $\pi_3(Df_p(e_2)) = \pi_3(Df_p(0, 1)) = \pi_3(1, 2, -1) = -1$ .

Thus, Theorem 8.1 may be rephrased as claiming

$$\pi_i(Df_p(e_j)) = \frac{\partial f_i}{\partial x_j}(p).$$

We are now ready to prove that result.

PROOF: Fix  $1 \leq i \leq m$  and  $1 \leq j \leq n$ , and define

$$\begin{aligned} g: (-\varepsilon, \varepsilon) &\rightarrow \mathbb{R}^n \\ t &\mapsto p + te_j \end{aligned}$$

where  $\varepsilon$  is chosen small enough so that the image of  $g$  is always in  $U$ . So  $g$  is a parametrized small line segment passing through  $p$  at time  $t = 0$  in the direction of  $e_j$ . Let  $\pi_i: \mathbb{R}^m \rightarrow \mathbb{R}$  be the  $i$ -th projection mapping,  $\pi_i(x_1, \dots, x_m) = x_i$ . Consider the composition  $h := \pi_i \circ f \circ g$ :

$$h: (-\varepsilon, \varepsilon) \xrightarrow{g} U \xrightarrow{f} \mathbb{R}^m \xrightarrow{\pi_i} \mathbb{R}.$$

Thus,  $h(t) = \pi_i(f(g(t))) = \pi_i(f(p + te_j)) = f_i(p + te_j)$ , and directly from the definition of partial derivatives,

$$\begin{aligned} h'(0) &= \lim_{t \rightarrow 0} \frac{h(t) - h(0)}{t} \\ &= \lim_{t \rightarrow 0} \frac{f_i(p + te_j) - f_i(p)}{t} \\ &= \frac{\partial f_i}{\partial x_j}(p). \end{aligned}$$

On the other hand, since  $h$  is a function from one variable calculus, we have seen that  $Dh_0$  is the linear function consisting of multiplication by  $h'(0)$  (cf. Section 4.1). Combining this with the chain rule gives

$$\begin{aligned} \left( \frac{\partial f_i}{\partial x_j}(p) \right) t &= h'(0)t = Dh_0(t) \\ &= D(\pi_i \circ f \circ g)_0(t) = (D(\pi_i \circ f)_{g(0)} \circ Dg_0)(t) \\ &= (D\pi_{i, f(g(0))} \circ Df_{g(0)} \circ Dg_0)(t) = (D\pi_{i, p_i} \circ Df_p \circ Dg_0)(t) \\ &= (\pi_i \circ Df_p \circ Dg_0)(t) \\ &= \pi_i(Df_p(te_j)) = t\pi_i(Df_p(e_j)) \\ &= t(i, j\text{-th entry of the matrix representing } Df_p). \end{aligned}$$

On the fourth and fifth lines, we have used the fact that the derivative of any affine function at any point is the linear part of the original function itself (exercise). In particular,  $D\pi_{i, p_i} = \pi_i$  and  $Dg_0(t) = te_j$ . Letting  $t = 1$  in the above calculation finishes the proof of the theorem.  $\square$

We now prove an important partial converse to Theorem 8.1. The proof is included mainly for completeness. The reader is encouraged to understand the statement of the theorem but can safely skip its proof for now.

**Theorem 8.2.** *Let  $f: U \rightarrow \mathbb{R}^m$  with  $U$  an open subset of  $\mathbb{R}^n$ . Let  $p \in U$ . Suppose that each partial derivative  $\partial f_i / \partial x_j$  exists and is continuous in an open set containing  $p$ . Then  $f$  is differentiable at  $p$ .*

PROOF: (adapted from Rudin's *Principles of Mathematical Analysis*). A function is differentiable if and only if each of its component functions is differentiable (exercise). Hence, we may assume that  $m = 1$ , i.e., that  $f$  is a real-valued function. In that case, the Jacobian matrix has only one row:

$$Jf(p) = \left( \frac{\partial f}{\partial x_1}(p) \quad \dots \quad \frac{\partial f}{\partial x_n}(p) \right).$$

From Theorem 8.1, we know that if  $f$  is differentiable at  $p$ , then its derivative must be the linear function determined by this Jacobian matrix:

$$(x_1, \dots, x_n) \mapsto \frac{\partial f}{\partial x_1}(p)x_1 + \dots + \frac{\partial f}{\partial x_n}(p)x_n.$$

Hence, our task is to show

$$\lim_{h \rightarrow 0} \frac{|f(p+h) - f(p) - \sum_{i=1}^n \frac{\partial f}{\partial x_i}(p)h_i|}{|h|} = 0. \quad (5)$$

Given  $\varepsilon > 0$ , first choose  $\delta > 0$  such that

$$\left| \frac{\partial f}{\partial x_i}(p+h) - \frac{\partial f}{\partial x_i}(p) \right| < \frac{\varepsilon}{n} \quad (6)$$

whenever  $|h| < \delta$ . We may do this since the first partials of  $f$  are continuous. Since  $U$  is open, we may choose  $\delta$  small enough so that  $p+h \in U$  when  $|h| < \delta$ . Also we may choose  $\delta$  small enough so that it works simultaneously for  $i = 1, \dots, n$ .

To establish 5, we must relate the difference  $f(p+h) - f(p)$  to an expression involving the partial derivatives of  $f$ . We do this by taking a path from  $p$  to  $p+h$  along line segments parallel to the coordinate axes. The function  $f$  restricted to any one of these paths will be a real-valued function of one variable, call it  $g$ . Applying the mean value theorem to  $g$  brings  $g'$  into play, and since the path in question is parallel to a coordinate axis,  $g'$  will be related to the partial derivative of  $f$  in that direction. We now spell out the details.

Given any  $|h| < \delta$ , define  $h^{(0)} := \vec{0}$  and  $h^{(i)} := \sum_{j=1}^i h_j e_j \in \mathbb{R}^n$  for  $1 \leq i \leq n$ . Hence,

$$\begin{aligned} h^{(0)} &:= \vec{0} \\ h^{(1)} &:= h_1 e_1 = (h_1, 0, \dots, 0) \\ h^{(2)} &:= h_1 e_1 + h_2 e_2 = (h_1, h_2, 0, \dots, 0) \\ &\vdots \\ h^{(n)} &:= h_1 e_1 + \dots + h_n e_n = (h_1, h_2, h_3, \dots, h_n) = h. \end{aligned}$$

The path we take from  $p$  to  $p+h$  starts at  $p = p + h^{(0)}$  and travels in the direction of  $e_1$  out to  $p + h^{(1)}$ . From there, it travels in the direction of  $e_2$  out to  $p + h^{(2)}$ , and so on. The  $i$ -th line segment that makes up this path can be parametrized by

$$\ell^{(i)}(t) := p + h^{(i-1)} + t(h^{(i)} - h^{(i-1)}) = p + h^{(i-1)} + th_i e_i = p + (h_1, \dots, h_{i-1}, th_i, 0, \dots, 0),$$

for  $0 \leq t \leq 1$ .

The image of each  $\ell^{(i)}$  is contained in  $U$ ; in fact each point in the image is within a distance of  $\delta$  of  $p$  (exercise). Thus, it makes sense to define  $g^{(i)}$  to be the restriction of  $f$  to the  $i$ -th line segment:

$$g^{(i)}(t) := f \circ \ell^{(i)}(t) = f(p + h^{(i-1)} + t(h^{(i)} - h^{(i-1)})),$$

for  $0 \leq t \leq 1$ . So as  $t$  goes from 0 to 1, the function  $g^{(i)}$  takes on the values of  $f$  along the  $i$ -th line segment. The change in  $f$  from  $p$  to  $p+h$  can be broken into a sum of changes along the line segments:

$$\begin{aligned} f(p+h) - f(p) &= \sum_{i=1}^n f(p+h^{(i)}) - f(p+h^{(i-1)}) \\ &= \sum_{i=1}^n g^{(i)}(1) - g^{(i)}(0). \end{aligned}$$

If you write out the first sum, you will see that almost every term cancels, leaving  $f(p+h) - f(p)$ , as claimed.

Each  $g^{(i)}$  is an ordinary function from one variable calculus. Since it is the composition of differentiable functions, the chain rule says that  $g$  is differentiable and

$$\begin{aligned} \left( \frac{dg^{(i)}}{dt}(t) \right) &= Jg^{(i)}(t) \\ &= J(f \circ \ell^{(i)})(t) \\ &= Jf(\ell^{(i)}(t))J\ell^{(i)}(t) \\ &= \left( \frac{\partial f}{\partial x_1}(\ell^{(i)}(t)) \quad \dots \quad \frac{\partial f}{\partial x_n}(\ell^{(i)}(t)) \right) \begin{pmatrix} 0 \\ \vdots \\ 0 \\ h_i \\ 0 \\ \vdots \\ 0 \end{pmatrix} \\ &= \left( \frac{\partial f}{\partial x_i}(\ell^{(i)}(t))h_i \right). \end{aligned}$$

Apply the mean value theorem to find  $c_i \in (0, 1)$  such that  $dg^{(i)}(c_i)/dt = g^{(i)}(1) - g^{(i)}(0)/(1-0) = g^{(i)}(1) - g^{(i)}(0)$ . Thus, we have

$$\begin{aligned} f(p+h) - f(p) &= \sum_{i=1}^n f(p+h^{(i)}) - f(p+h^{(i-1)}) \\ &= \sum_{i=1}^n g^{(i)}(1) - g^{(i)}(0) \\ &= \sum_{i=1}^n \frac{dg^{(i)}}{dt}(c_i) \\ &= \sum_{i=1}^n \frac{\partial f}{\partial x_i}(\ell^{(i)}(c_i))h_i. \end{aligned}$$

Each  $\ell^{(i)}(c_i)$  is within a distance of  $\delta$  from  $p$ , so multiplying condition 6, above, through by  $|h_i|$  gives that

$$\left| \frac{\partial f}{\partial x_i}(\ell^{(i)}(c_i))h_i - \frac{\partial f}{\partial x_i}(p)h_i \right| < \frac{\varepsilon|h_i|}{n}.$$

Therefore,

$$\begin{aligned} \left| f(p+h) - f(p) - \sum_{i=1}^n \frac{\partial f}{\partial x_i}(p)h_i \right| &= \left| \sum_{i=1}^n \frac{\partial f}{\partial x_i}(\ell^{(i)}(c_i))h_i - \sum_{i=1}^n \frac{\partial f}{\partial x_i}h_i \right| \\ &< \sum_{i=1}^n \frac{\varepsilon|h_i|}{n} \\ &\leq \varepsilon|h|. \end{aligned}$$

This establishes equation 5 and finishes the proof.  $\square$

If the partial derivatives of  $f$  exist at  $p$  but are not continuous in an open set about  $p$ , Theorem 8.2 leaves open the possibility that although the Jacobian matrix exists,  $f$  may not be differentiable. This is indeed possible although pathological. An example is given in the exercises.

## 9. Conclusion

After reading this chapter, you are in a position to thoroughly understand Chapter 1. You may want to review that chapter now. In particular, we have completed everything listed in the “To Do” section of that chapter. We have given a precise definition of the derivative. From that definition we could see in what sense the derivative (more precisely, the best affine approximation) gives a good approximation of a function. We have given a precise definition of a partial derivative, and we have proved that the derivative, if it exists, is the linear function associated with the Jacobian matrix. In addition, we proved the chain rule, which states that the composition of differentiable functions is again differentiable, the derivative of the composition being the composition of the corresponding derivatives. This result is surprisingly important. For instance, it was the key step in showing the relation between the derivative and the Jacobian matrix. In the world of physics, it implies the law of conservation of energy, as we will see in the exercises.

## EXERCISES

- (1) Let  $p \in \mathbb{R}^n$ . Show that the set  $U = \mathbb{R}^n \setminus \{p\}$ , obtained by removing the single point  $p$  from  $\mathbb{R}^n$ , is open.
- (2) Which of the following subsets are open? (No proofs are necessary.)
- $\{x \in \mathbb{R} \mid x \neq 2 \text{ and } x \neq 7\} \subset \mathbb{R}$ .
  - $\mathbb{R}^2 \setminus \{(m, n) \in \mathbb{R}^2 \mid m \text{ and } n \text{ are integers}\}$
  - $\{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 < 1 \text{ and } z \leq 0\}$ .
  - $\mathbb{R} \setminus \{1/n \mid n \text{ a nonzero integer}\}$ .
- (3) Let  $S := \{a, b, c\}$  be an arbitrary set consisting of three elements. Describe five different topologies on  $S$  (cf. Definition 2.5).
- (4) (a) Let  $C_\alpha$  be a closed subset of  $\mathbb{R}^n$  for each  $\alpha$  in some index set  $I$ . Prove that  $\bigcap_{\alpha \in I} C_\alpha$  is closed.  
 (b) Give an example of a collection of closed subsets  $\{C_\alpha\}_{\alpha \in I}$  of  $\mathbb{R}^n$  such that  $\bigcup_{\alpha \in I} C_\alpha$  is not closed.  
 (c) Let  $C_1, \dots, C_k$  be closed subsets of  $\mathbb{R}^n$ . Prove that  $\bigcup_{i=1}^k C_i$  is closed.

- (5) Let  $f(x) = 4x + 7$ .
- Prove that  $\lim_{x \rightarrow 2} f(x) = 15$  directly from the definition of a limit.
  - Prove that  $f$  is continuous directly from the definition of continuity.
- (6) Prove the remaining parts of Theorem 3.2. To show that  $\lim_{x \rightarrow s} fg = ab$  in part 4, use the identity

$$f(x)g(x) - ab = (f(x) - a)(g(x) - b) + b(f(x) - a) + a(g(x) - b).$$

Given this result about products, in order to prove  $\lim_{x \rightarrow s} f/g = a/b$  it suffices to prove  $\lim_{x \rightarrow s} 1/g = 1/b$ . First argue that given  $\varepsilon > 0$ , you can take  $x$  close enough to  $s$  so that both  $|g(x)| > |b|/2$  and  $|g(x) - b| < \varepsilon b^2/2$ . Then bound  $|(1/g(x)) - (1/b)| = |(b - g(x))/g(x)b|$ .

- (7) (a) Prove parts 1–5 of Theorem 3.4.  
 (b) Let  $p \in \mathbb{R}^m$ . Use an  $\varepsilon$ - $\delta$  argument to prove that the constant function

$$\begin{aligned} f: \mathbb{R}^n &\rightarrow \mathbb{R}^m \\ v &\mapsto p \end{aligned}$$

is continuous.

- (c) Let  $1 \leq i \leq n$  and define  $\rho_i: \mathbb{R}^n \rightarrow \mathbb{R}^{n-1}$  by dropping the  $i$ -th coordinate of a point:

$$\rho_i(x_1, \dots, x_n) = (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n).$$

Use an  $\varepsilon$ - $\delta$  argument to prove that  $\rho_i$  is continuous.

- (d) Define the  $i$ -th projection mapping,  $\pi_i: \mathbb{R}^m \rightarrow \mathbb{R}$ , by  $\pi_i(x_1, \dots, x_m) = x_i$ . Use an  $\varepsilon$ - $\delta$  argument to prove that  $\pi_i$  is continuous.

- (8) Prove part 6 of Theorem 3.4.
- (9) Let  $x_1, x_2, \dots$  be a sequence of points in  $\mathbb{R}^n$ , denoted  $\{x_t\}$ . We say that *the limit of the sequence is  $x$*  or that *the sequence converges to  $x$* , denoted  $\lim_{t \rightarrow \infty} x_t = x \in \mathbb{R}^n$ , if for all  $\varepsilon > 0$  there is an integer  $T$  such that  $|x - x_t| < \varepsilon$  whenever  $t > T$ . If each  $x_t$  is an element of a subset  $S \subseteq \mathbb{R}^n$ , we'll say that  $\{x_t\}$  is a *sequence in  $S$*  or that it belongs to  $S$ .

- (a) Prove that if  $x$  is a limit point of a set  $S \subseteq \mathbb{R}^n$ , then there is a sequence  $\{x_t\}$  of points in  $S$  such that  $\lim_{t \rightarrow \infty} x_t = x$ . (Note that the converse is not true. For instance, let  $S = [0, 1] \cup \{2\}$ . Then 2 is an isolated point of  $S$ , but if we let  $x_t = 2$  for all  $t$ , then  $\lim_{t \rightarrow \infty} x_t = 2$ .)
- (b) True or false: Coupled with Proposition 2.10, the previous exercise shows that  $S$  is closed if and only if it contains the limits of all its convergent sequences. In other words,  $S$  is closed if and only if  $x \in S$  whenever  $\lim_{t \rightarrow \infty} x_t = x$  for some sequence  $\{x_t\}$  in  $S$ . If true, provide a proof. If false, give a counterexample.
- (c) Let  $S \subseteq \mathbb{R}^n$ , and let  $f: S \rightarrow \mathbb{R}^m$  be a function. Let  $s$  be a limit point of  $S$ . Prove that  $\lim_{x \rightarrow s} f(x) = p$  if and only if  $\lim_{t \rightarrow \infty} f(x_t) = p$  for all sequences  $\{x_t\}$  in  $S$  converging to  $s$ .
- (d) With  $f$  as above, suppose that  $s \in S$ . Prove that  $f$  is continuous at  $s$  if and only if  $\lim_{t \rightarrow \infty} f(x_t) = f(s)$  for all sequences  $\{x_t\}$  in  $S$  converging to  $s$ . (Thus, you could say that a continuous function is one that “commutes” with the process of taking limits:  $\lim_{t \rightarrow \infty} f(x_t) = f(\lim_{t \rightarrow \infty} x_t)$ .)
- (10) Let  $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$ . Prove that  $\lim_{x \rightarrow s} f(x) = p$  if and only if  $\lim_{x \rightarrow s} |f(x) - p| = 0$ .
- (11) Let  $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$ , and suppose that  $\lim_{h \rightarrow \vec{0}} f(h) = \vec{0}$ . Let  $x$  be a nonzero vector in  $\mathbb{R}^n$ . Define a new function  $g: \mathbb{R} \rightarrow \mathbb{R}^m$  by  $g(t) := f(tx)$ . Prove that  $\lim_{t \rightarrow 0} g(t) = \vec{0}$ .
- (12) Prove Theorem 5.1.
- (13) Let  $f: U \rightarrow \mathbb{R}^m$  with  $U$  an open subset of  $\mathbb{R}^n$ . Suppose that  $f$  is differentiable at  $p \in U$ . Show that  $f$  is continuous at  $p$  by completing the following steps:
- (a) Use the definition of the derivative to show that there is a  $\delta' > 0$  such that  $0 < |h| < \delta' \Rightarrow |f(h+p) - f(p) - Df_p(h)| < |h|$ . (Let  $\epsilon = 1$  in the definition of the limit.)
- (b) From Lemma 6.4, we know there is a constant  $c$  such that  $|Df_p(h)| \leq c|h|$  for all  $h \in \mathbb{R}^n$ . Use this fact and the reverse triangle inequality (cf. exercises to Chapter 3) to show that  $0 < |h| < \delta' \Rightarrow |f(h+p) - f(p)| < (1+c)|h|$ .
- (c) Now use the definition of continuity to prove that  $f$  is continuous at  $p$ . (Given  $\epsilon > 0$ , let  $\delta = \min\{\epsilon/(1+c), \delta'\}$ .)
- (14) Let  $f(x, y) = (x^2 - 2y, xy)$ . Prove directly from the definition of the derivative, Definition 4.1, that  $Df_{(2,1)}(x, y) = (4x - 2y, x + 2y)$ .
- (15) Let  $A: \mathbb{R}^n \rightarrow \mathbb{R}^m$  be an affine function, so  $A(x) = q + L(x)$  for some  $q \in \mathbb{R}^m$  and some linear function  $L: \mathbb{R}^n \rightarrow \mathbb{R}^m$ . Prove directly from the definition of the derivative that the derivative of  $A$  at every point  $p \in \mathbb{R}^n$  is its linear part,  $L$ , i.e.,  $DA_p = L$  for all  $p$ . (As a special case, when  $q = 0$ , this shows that the derivative of a linear function at any point is just the linear function, itself.)
- (16) Prove that a function is differentiable at a point if and only if each of its component functions is differentiable at that point.
- (17) For each of the functions  $f$  and  $g$  and points  $p$ , below, verify the chain rule by completing the following steps.
- (i) Calculate  $Jg(f(p))$ ,  $Jf(p)$ , and the product  $Jg(f(p))Jf(p)$ .
- (ii) Calculate  $(g \circ f)$ .
- (iii) Calculate  $J(g \circ f)(p)$  and verify that it equals  $Jg(f(p))Jf(p)$ .
- (iv) Calculate  $Dg_{f(p)}$ ,  $Df_p$ ,  $D(g \circ f)_p$ , and  $Dg_{f(p)} \circ Df_p$ . Show the calculation of  $Dg_{f(p)} \circ Df_p$  even though it is implied by the corresponding result for the Jacobian matrices.

- (a)  $f(x, y) = (x^2 + y, xy - y^2)$ ,  $g(x, y) = x^2 - 3y$ , and  $p = (2, 3)$ .  
 (b)  $f(x, y) = (xy, x^2 + 2)$ ,  $g(x, y) = (x^2 - 3y, x - y^2)$ , and  $p = (2, -2)$ .  
 (c)  $f(t) = (t, t^2, t^3)$ ,  $g(x, y, z) = (x^2 - 4y, xy + z^2)$ , and  $p = 1$ .  
 (d)  $f(x, y, z) = x + 2y + z^3$ ,  $g(t) = (t, t + t^2)$ , and  $p = (1, 0, -1)$ .  
 (e)  $f(x) = \cos(x)$ ,  $g(x) = e^x$ , and  $p = \pi/2$ .
- (18) By the chain rule if  $f: \mathbb{R}^n \rightarrow \mathbb{R}^k$  and  $g: \mathbb{R}^k \rightarrow \mathbb{R}^m$  are differentiable functions and  $p \in \mathbb{R}^n$ , then  $D(g \circ f)_p = Dg_{f(p)} \circ Df_p$ , or in terms of Jacobian matrices,  $J(g \circ f)(p) = Jg(f(p))Jf(p)$ . You may sometimes see a version of the chain rule stated as follows. Suppose  $z$  is a function of two variables,  $x$  and  $y$ , both of which are a function of a third variable:  $z = h(x, y)$ ,  $x = m(t)$ , and  $y = n(t)$ . Then

$$\frac{dz}{dt} = \frac{\partial z}{\partial x} \frac{dx}{dt} + \frac{\partial z}{\partial y} \frac{dy}{dt}. \quad (7)$$

What is the connection between our version and this other version?

- (a) What are  $f$  and  $g$  in this situation?  
 (b) How does our chain rule give equation 7?
- (19) Prove that each point in the image of the parametrized line segment  $\ell^{(i)}$  in Theorem 8.2 is within a distance of  $\delta$  of the point  $p$  given that  $|h| < \delta$ .
- (20) Suppose that  $f$  is a differentiable function with  $f(1, 2) = 5$  and  $f(1.01, 2) = 5.1$ . What is your best guess for the value of the partial derivative of  $f$  with respect to  $x$  at  $(1, 2)$ ?
- (21) This exercise shows that the continuity hypothesis in Theorem 7.4 is necessary. Define

$$f(x, y) = \begin{cases} \frac{xy(x^2 - y^2)}{x^2 + y^2} & \text{if } (x, y) \neq (0, 0); \\ 0 & \text{if } (x, y) = (0, 0). \end{cases}$$

- (a) Show that  $f$ ,  $\partial f/\partial x$ , and  $\partial f/\partial y$  are continuous on  $\mathbb{R}^2$ .  
 (b) Show that  $\partial^2 f/\partial x \partial y$  and  $\partial^2 f/\partial y \partial x$  exist on  $\mathbb{R}^2$  and are continuous on  $\mathbb{R}^2 \setminus \{(0, 0)\}$ .  
 (c) Show that
- $$\frac{\partial^2 f}{\partial x \partial y}(0, 0) = 1 \neq -1 = \frac{\partial^2 f}{\partial y \partial x}(0, 0).$$
- (d) Extra credit: carefully describe the graph of  $f$  and explain geometrically why the second order partials that you just calculated are not equal.
- (22) This exercise shows that the continuity hypothesis in Theorem 8.2 is necessary. Define

$$f(x, y) = \begin{cases} \frac{xy}{x^2 + y^2} & \text{if } (x, y) \neq (0, 0); \\ 0 & \text{if } (x, y) = (0, 0). \end{cases}$$

Prove that  $\partial f/\partial x$  and  $\partial f/\partial y$  exist at every point in  $\mathbb{R}^2$  but that  $f$  is not even continuous at  $(0, 0)$ , hence  $f$  is not differentiable at  $(0, 0)$  (by a previous exercise). (Extra credit: carefully describe the graph of this function.)

- (23) **Conservation of energy.** Let  $f: U \rightarrow \mathbb{R}^n$  be a differentiable function on  $U$ , an open subset of  $\mathbb{R}^n$ . We think of  $f$  as a vector field (attaching a vector  $f(p)$  to each point  $p$  in  $U$ ). Suppose that  $f = \text{grad } \phi$  for some function  $\phi: U \rightarrow \mathbb{R}$ . In that case,

we say that  $f$  is a *conservative* vector field with *potential (energy)*  $\psi := -\phi$ . Hence  $f = -\text{grad } \psi$ .

Let  $I$  be an interval in  $\mathbb{R}$ , and let  $h: I \rightarrow U$  be a twice differentiable curve in  $U$ . Suppose that  $h$  describes the motion of a particle of mass  $m$  satisfying *Newton's law* with respect to  $f$ ; this means that  $f(h(t)) = mh''(t)$ . It is interpreted as saying that the force exerted on the particle at time  $t$  is a constant times the particle's acceleration ( $F = ma$ ). Finally, define the *kinetic energy* of the particle to be  $K(t) := \frac{1}{2}m|h'(t)|^2$ .

Prove that the sum of the potential and kinetic energy is a constant:  $K(t) + \psi(h(t)) = \text{constant}$ . You can do this by showing the derivative of the left-hand side of this equation is zero (using the chain rule and Exercise 16 of Chapter 3.)



# Optimization

## 1. Introduction

Let  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  be a real-valued function. In this chapter, we focus on the problem of finding the local and global minima and maxima for  $f$ . In Section 2.2 of Chapter 2 we defined the gradient of  $f$  at any point  $p \in \mathbb{R}^n$  and claimed that it is a vector pointing in the direction one must move to make  $f$  increase as quickly as possible. As  $p$  varies, so does the gradient. Points at which the gradient is zero are called *critical points* of  $f$ . At a local maximum, there is no direction in which the function increases, and the gradient at the corresponding point is zero, i.e., every point at which  $f$  assumes a local maximum (or minimum, it turns out) is a critical point. The converse is not true. Critical points which do not give local maxima or minima are called *saddle points*, which suggests the behavior that occurs at these points. To tell the difference, we need to look at higher order approximations of  $f$ , generalizations of the derivative, called *Taylor polynomials*.

Another approach towards finding maxima and minima is to start at any point and go with the flow of the gradient. Move in the direction that the gradient is pointing, and you will arrive at a point which is closer to being a local maximum. Move in the opposite direction to look for local minima. This point of view, which lends itself to numerical calculations on a computer, is especially useful when the domain of  $f$  is restricted, e.g., when trying to optimize  $f$  over a region in  $\mathbb{R}^n$  which has a boundary. It is called the method of *Lagrange multipliers*, which we cover near the end of the chapter. In that case, one can imagine moving in the direction of the gradient until hitting the boundary of the region. Moving further in the direction of the gradient would take one outside of the region. Hence, in the case where the domain of  $f$  is not all of  $\mathbb{R}^n$ , a maximum may occur at a point in the region even though that point is not a critical point.

## 2. Directional derivatives and the gradient

**Definition 2.1.** Let  $f: U \rightarrow \mathbb{R}$  where  $U$  is an open subset of  $\mathbb{R}^n$ , and assume that  $f$  is differentiable at  $p \in U$ . Let  $v \in \mathbb{R}^n$  be any unit vector. The directional derivative of  $f$  at  $p$  in the direction of  $v$  is the real number

$$f_v(p) := \lim_{t \rightarrow 0} \frac{f(p + tv) - f(p)}{t}$$

The directional derivative is visibly the rate of change of  $f$  in the direction of  $v$ . A special case is when  $v$  is one of the standard basis vectors,  $e_i$ , and in that case we get an ordinary partial derivative

$$f_{e_i}(p) = \frac{\partial f}{\partial x_i}(p).$$

Another way of thinking of the directional derivative is to define the curve  $c(t) = p + tv$  in  $\mathbb{R}^n$ . The directional derivative is the rate of change of  $f$  restricted to points along this curve. More formally, define  $g(t) = (f \circ c)(t) = f(p + tv)$ . Then  $g$  is a function from one variable calculus whose ordinary derivative at zero is the directional derivative:

$$g'(0) = \lim_{t \rightarrow 0} \frac{g(t) - g(0)}{t} = f_v(p).$$

In fact we can let  $c$  be any differentiable curve passing through  $p$  with velocity  $v$  at time  $t = 0$ , and get the same result (exercise).

**Example 2.2.** Let  $f(x, y) = x^2 - y^2$ . The directional derivative of  $f$  at the point  $p = (1, 2)$  in the direction of the unit vector  $v = (2, 3)/\sqrt{13}$  is

$$\begin{aligned} f_v(p) &= \lim_{t \rightarrow 0} \frac{f(p + tv) - f(p)}{t} \\ &= \lim_{t \rightarrow 0} \frac{f(1 + 2t/\sqrt{13}, 2 + 3t/\sqrt{13}) - f(1, 2)}{t} \\ &= \lim_{t \rightarrow 0} \frac{((1 + 2t/\sqrt{13})^2 - (2 + 3t/\sqrt{13})^2) + 3}{t} \\ &= -8/\sqrt{13}. \end{aligned}$$

As you might guess, there is an easier way to calculate the directional derivative. If  $f$  is differentiable at a point  $p \in U$ , the Jacobian matrix consists of a single row. In Chapter 2, we defined the *gradient* of  $f$  at  $p$  to be this single row, thought of as a vector in  $\mathbb{R}^n$ :

$$\text{grad } f(p) := \nabla f(p) := \left( \frac{\partial f}{\partial x_1}(p), \dots, \frac{\partial f}{\partial x_n}(p) \right).$$

We can use the gradient to calculate any directional derivative at  $p$ .

**Proposition 2.3.** *Let  $f: U \rightarrow \mathbb{R}$  where  $U$  is an open subset of  $\mathbb{R}^n$ , and assume that  $f$  is differentiable at  $p \in U$ . Let  $v \in \mathbb{R}^n$  be any unit vector. Then*

$$f_v(p) = \nabla f(p) \cdot v.$$

PROOF: Define  $c(t) = p + tv$  for  $t$  in a small enough interval about 0 so that the image of  $c$  is contained in  $U$ . Define  $g = f \circ c$ . We have already seen that  $g'(0) = f_v(p)$ . By the chain rule

$$f_v(p) = g'(0) = (f \circ c)'(0) = \nabla f(c(0)) \cdot c'(0) = \nabla f(p) \cdot v.$$

In making this calculation, recall that  $\nabla f$  and  $c'$  are essentially the Jacobian matrices of  $f$  and  $c$ , respectively. Their dot product is essentially the product of the Jacobian matrices arising from the chain rule.  $\square$

**Important note.** With notation as in the previous proposition, check that

$$f_v(p) = Df_p(v).$$

If  $w \in U \subseteq \mathbb{R}^n$  is any nonzero vector—not necessarily of unit length—we can let  $v = w/|w|$  and write  $Df_p(w/|w|) = f_{w/|w|}(p)$ , or  $Df_p(w) = |w|f_{w/|w|}(p)$ . In other words,  $Df_p(w)$  is the rate of change of  $f$  in the direction of  $w$  at  $p$ , scaled by the length of  $w$ . This gives a nice interpretation of the derivative in the case where the codomain is  $\mathbb{R}$ .

**Example 2.4.** Let's try Example 2.2 again, this time using the gradient. The gradient of  $f(x, y) = x^2 - y^2$  is  $\nabla f = (2x, -2y)$ . Hence the directional derivative of  $f$  at  $p = (1, 2)$  in the direction of  $v = (2, 3)/\sqrt{13}$  is

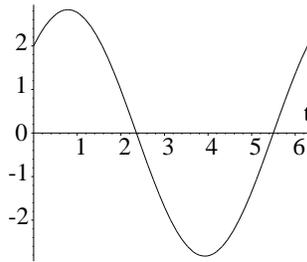
$$\nabla f(p) \cdot v = (2, -4) \cdot (2, 3)/\sqrt{13} = -8/\sqrt{13},$$

as before.

**Example 2.5.** Let  $f(x, y) = x^2 + y^2$ . Its graph is a paraboloid. It is clear that at any point  $(x, y) \in \mathbb{R}^2$ , one should move radially outward in order to make the function increase most rapidly, and this is confirmed by the gradient,  $\nabla f(x, y) = (2x, 2y)$ . Fix the point  $p = (1, 1)$ , and let  $v = (\cos \theta, \sin \theta)$  be an arbitrary unit vector. The directional derivative of  $f$  at the point  $p$  and in the direction of  $v$  is

$$f_v(p) = \nabla f(p) \cdot v = (2, 2) \cdot (\cos \theta, \sin \theta) = 2 \cos \theta + 2 \sin \theta.$$

Here is a graph of the directional derivative as a function of  $\theta$ :



Note how this graph agrees with geometric intuition.

**2.1. Main properties of the gradient.** In Chapter 2, you were asked to take three properties of the gradient on faith:

- The gradient points in the direction of quickest increase of the function.
- The length of the gradient gives the rate of increase of the function in the direction of its quickest increase.
- The gradient is perpendicular to its corresponding level set.

We are now ready to rigorously establish those properties.

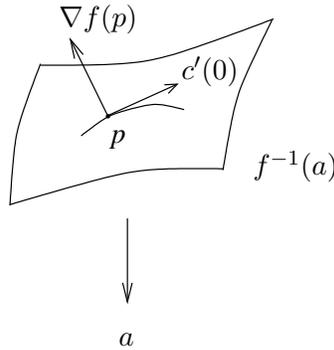
The rate of increase of a real-valued function  $f$  at a point  $p$  and in the direction of the unit vector  $v$  is given by the corresponding directional derivative. By the proposition just proved,

$$f_v(p) = \nabla f(p) \cdot v = |\nabla f(p)||v| \cos(\theta) = |\nabla f(p)| \cos(\theta),$$

where  $\theta$  is the angle (cf. Definition 2.11) between the gradient  $\nabla f(p)$  and  $v$ . (Since  $v$  is just a direction—a unit vector—we have  $|v| = 1$ ; so  $v$  disappears in the calculation. Also, we are assuming that  $\nabla f(p) \neq 0$ ; otherwise, the gradient does not point anywhere.) As  $v$  varies, the directional derivative reaches its maximum when  $\theta = 0$ , i.e., when it is pointing in the direction of the gradient, as claimed. In that case,  $\cos(\theta) = 1$  and  $f_v(p) = |\nabla f(p)|$ ; in other words, the rate of quickest increase of  $f$  at  $p$  is the length of the gradient vector.

The minimum occurs when  $\theta = \pi$  radians, and  $f_v(p) = -|\nabla f(p)|$ . Of course, if  $\nabla f(p) = \vec{0}$ , then  $f_v(p) = 0$  for all  $v$ . That takes care of the first two properties.

To prove the third property, we need to decide what is meant by saying the gradient is perpendicular to its corresponding level set. Recall that a level set of  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  is just the inverse image of a point (cf. Section 2.2). For instance, if we think of  $f$  as assigning a temperature to each point in space, a level set is an isotherm: the set of all points in space having the same given temperature. Pick a point  $p \in \mathbb{R}^n$ , and let  $a = f(p)$ . So  $p$  lies on the level set  $f^{-1}(a)$ . Pick any differentiable curve  $c$  lying in the level set and passing through the point  $p$  at time  $t = 0$ . Thus,  $f(c(t)) = a$  for all  $t$ , and  $c(0) = p$ . We will show that the  $\nabla f(p)$  is perpendicular to the tangent to  $c$  at time 0, i.e.,  $\nabla f(p) \cdot c'(0) = 0$ . This is precisely what we mean by saying that the gradient is perpendicular to its corresponding level set: it is perpendicular to all curves in the level set passing through the point in question. Here is a picture for a typical function  $f$  with domain  $\mathbb{R}^3$  (so its level sets are typically surfaces):



To verify the claim that the gradient is perpendicular to the curve at  $p$ , use the chain rule:

$$f(c(t)) = a = \text{constant} \implies f(c(t))'(p) = 0 \implies \nabla f(p) \cdot c'(0) = 0.$$

As mentioned earlier, the product of the Jacobians  $Jf(p)$  and  $Jc(0)$  arising from the chain rule is given by the dot product of  $\nabla f(p)$  with  $c'(0)$  since  $Jf(p)$  is essentially the gradient of  $f$  and  $Jc(0)$  is the tangent to  $c$ .

As an example, suppose  $f(x, y, z) = x^2 + y^2 + z^2$  and  $p = (0, 0, 1)$ . Then  $f(p) = 1$ , and the level set through  $p$  is a sphere of radius 1 centered at the origin:  $f^{-1}(1) = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\}$ . Our calculation has shown that if you take any curve lying in this sphere and passing through the north pole,  $(0, 0, 1)$ , its velocity will be perpendicular to the gradient at that point,  $\nabla f(p) = (0, 0, 2)$ .

### 3. Taylor's theorem

In Chapter 4, we saw that the best affine approximation gives the best first order approximation of a function. It allows us to approximate a function near a point with a function whose components are polynomials of degree at most 1. In this section, we learn how to make best approximations whose components are polynomials of quadratic and higher degrees. At the expense of using more complicated functions, we will get better approximations.

You may already be familiar with the one-variable version of this idea. Let  $f: \mathbb{R} \rightarrow \mathbb{R}$ , and let  $p \in \mathbb{R}$ . We would like to find a polynomial  $T_p^k f(x)$  of degree  $k$  that approximates  $f$  near  $p$ . The polynomial will have the form

$$T_p^k f(x) = a_0 + a_1(x - p) + a_2(x - p)^2 + a_3(x - p)^3 + \cdots + a_k(x - p)^k.$$

Requiring that the  $i$ -th order derivatives of  $T_p^k f$  and  $f$  agree at  $p$  for  $i = 0, 1, \dots, k$  determines the coefficients  $a_0, a_1, \dots, a_k$  and forces  $T_p^k f$  to closely approximate  $f$  near  $p$ :

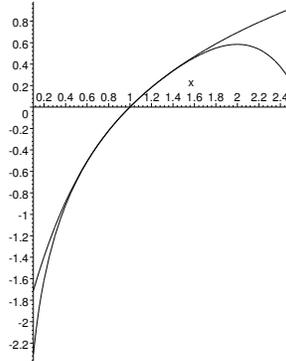
$$\frac{d^i}{dx^i} T_p^k f(p) = f^{(i)}(p) \implies a_i = \frac{1}{i!} f^{(i)}(p),$$

where the factorial,  $i!$ , is defined by  $i! = 1 \cdot 2 \cdot \dots \cdot i$  if  $i > 0$ , and  $0! = 1$ .

**Example 3.1.** Let  $f(x) = \ln(x)$  and let  $p = 1$ . To calculate the 4-th order Taylor polynomial of  $f$  at  $p$ , calculate the derivatives of  $\ln(x)$  up to order 4:  $\ln'(x) = 1/x$ ,  $\ln''(x) = -1/x^2$ ,  $\ln'''(x) = 2/x^3$ ,  $\ln^{(4)}(x) = -3 \cdot 2/x^4$ . Therefore, “up to order 4,” the best approximation of  $\ln(x)$  near  $x = 1$  is

$$\begin{aligned} \ln(x) &\approx \ln(1) + \ln'(1)(x-1) + \frac{1}{2!} \ln''(1)(x-1)^2 \\ &\quad + \frac{1}{3!} \ln'''(1)(x-1)^3 + \frac{1}{4!} \ln^{(4)}(1)(x-1)^4 \\ &= (x-1) - \frac{1}{2!}(x-1)^2 + \frac{2}{3!}(x-1)^3 - \frac{3 \cdot 2}{4!}(x-1)^4 \\ &= (x-1) - \frac{1}{2}(x-1)^2 + \frac{1}{3}(x-1)^3 - \frac{1}{4}(x-1)^4. \end{aligned}$$

From the graphs of the two functions near  $x = 1$ , we can see that the Taylor polynomial is a good approximation, but only near  $x = 1$ . Can you tell which function is  $\ln(x)$  and which is the Taylor polynomial?



The idea for Taylor polynomials in higher dimensions is similar. For instance, suppose  $f(x, y)$  is a real-valued function of two variables, and let  $p = (p_1, p_2) \in \mathbb{R}^2$ . To find a polynomial of degree  $k$  closely approximating  $f$  near  $p$ , we write

$$\begin{aligned} T_p^k f(x, y) &= a_{0,0} + a_{1,0}(x-p_1) + a_{0,1}(y-p_2) + a_{2,0}(x-p_1)^2 \\ &\quad + a_{1,1}(x-p_1)(y-p_2) + a_{0,2}(y-p_2)^2 + \dots + a_{0,k}(y-p_2)^k \\ &= \sum_{i,j: i+j \leq k} a_{i,j}(x-p_1)^i (y-p_2)^j. \end{aligned}$$

Just as before, we can determine all of the coefficients  $a_{i,j}$  and force  $T_p^k f$  to be a good approximation of  $f$  near  $p$  by requiring all mixed partial derivatives of  $T_p^k f$  and  $f$  to agree at  $p$  up to order  $k$ . Check that

$$\frac{\partial^{i+j} T_p^k f}{\partial x^i \partial y^j}(p) = i! j! a_{i,j}.$$

So we let

$$a_{i,j} = \frac{1}{i!j!} \frac{\partial^{i+j} f}{\partial x^i \partial y^j}(p).$$

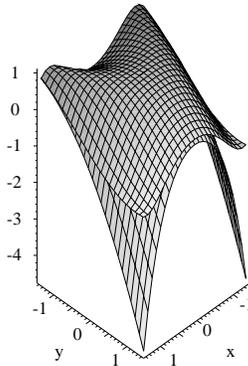
**Example 3.2.** We will calculate the fourth order Taylor polynomial for  $f(x, y) = \cos(x^2 + y)$  at the origin,  $p = (0, 0)$ . First, calculate the values of the derivatives for  $f$  up to order 4 at  $(0, 0)$ . It turns out that they are all zero except as listed below:

$$\begin{aligned} f(0, 0) &= 1, & \frac{1}{0!2!} \frac{\partial^2 f}{\partial y^2}(0, 0) &= -\frac{1}{2}, & \frac{1}{2!1!} \frac{\partial^3 f}{\partial x^2 \partial y}(0, 0) &= -1, \\ \frac{1}{4!} \frac{\partial^4 f}{\partial x^4}(0, 0) &= -\frac{1}{2}, & \frac{1}{4!} \frac{\partial^4 f}{\partial y^4}(0, 0) &= \frac{1}{24}. \end{aligned}$$

So the best fourth order approximation to  $f$  at  $(0, 0)$  is

$$T_p^4 f(x, y) = 1 - \frac{1}{2}y^2 - x^2y - \frac{1}{2}x^4 + \frac{1}{24}y^4.$$

Here is a picture of a piece of the graph of  $f$  sitting on top of a piece of the graph of the Taylor polynomial. Note the close fit in the middle, near  $(0, 0)$ .



**Example 3.3.** Let  $f(x, y) = 3 + x + x^2 + 3xy - y^2$ . To find the second order Taylor polynomial of  $f$  at  $p = (1, 2)$ , first calculate the relevant derivatives:

$$\begin{aligned} f(1, 2) &= 7, & \frac{1}{1!0!} \frac{\partial f}{\partial x}(1, 2) &= 9, & \frac{1}{0!1!} \frac{\partial f}{\partial y}(1, 2) &= -1 \\ \frac{1}{2!0!} \frac{\partial^2 f}{\partial x^2}(1, 2) &= 1, & \frac{1}{1!1!} \frac{\partial^2 f}{\partial x \partial y}(1, 2) &= 3, & \frac{1}{0!2!} \frac{\partial^2 f}{\partial y^2}(1, 2) &= -1. \end{aligned}$$

Thus,

$$T_p^2 f(x, y) = 7 + 9(x - 1) - (y - 2) + (x - 1)^2 + 3(x - 1)(y - 2) - (y - 2)^2.$$

Since  $f$  is a polynomial of degree 2, it would make sense that the best approximation of  $f$  by a polynomial of degree 2 would be  $f$ , itself, no matter what point we are looking at. This turns out to be the case: if you expand  $T_p^2 f(x, y)$ , you will see that it exactly equals  $f$ .

We now formalize the definition of a Taylor polynomial and prove Taylor's theorem which describes in what sense Taylor polynomials give good approximations.

**Definition 3.4.** Let  $f: S \rightarrow \mathbb{R}^m$  with  $S \subseteq \mathbb{R}^n$ . Suppose that all partial derivatives of  $f$  of order less than or equal to  $k$  exist at  $p \in S$ . The Taylor polynomial of order  $k$  for  $f$  at  $p$  is

$$T_p^k f(x_1, \dots, x_n) := \sum_{i_1 + \dots + i_n \leq k} \frac{1}{i_1! \dots i_n!} \frac{\partial^{i_1 + \dots + i_n} f}{\partial x_1^{i_1} \dots \partial x_n^{i_n}}(p) (x_1 - p_1)^{i_1} \dots (x_n - p_n)^{i_n}.$$

If the partial derivatives of every order exist at  $p$ , one defines the Taylor series for  $f$  at  $p$ ,  $T_p^\infty f$ , by summing over all non-negative indices,  $i_1, \dots, i_n$ .

In general  $f$  will be vector-valued, e.g.,  $m > 1$ . In that case, each mixed partial of  $f$  evaluated at  $p$  will be a vector. The definition then says to multiply that vector by the scalar,  $(x_1 - p_1)^{i_1} \dots (x_n - p_n)^{i_n}$ . In the end, one may find the Taylor polynomial of a vector-valued function by calculating the Taylor polynomials of the components of  $f$  separately. The first order Taylor polynomial is the best affine approximation of  $f$  at  $p$ , i.e.,  $T_p^1 f = T_p f$ .

We prove Taylor's theorem by first proving the special case of real-valued functions of one variable. The full theorem then can be reduced to this special case.

**Theorem 3.5.** (Taylor's Theorem in one variable) Let  $f: S \rightarrow \mathbb{R}$  with  $S \subseteq \mathbb{R}$ . Suppose that  $S$  contains an open interval containing the closed interval  $[a, b]$ . Also suppose that all the derivatives up to order  $k$  exist and are continuous on  $[a, b]$  and the  $(k + 1)$ -th derivative exists on  $(a, b)$ . Let  $x, p \in [a, b]$ . Then there exists a number  $c$  between  $x$  and  $p$  such that

$$f(x) = T_p^k f(x) + \frac{1}{(k + 1)!} f^{(k+1)}(c) (x - p)^{k+1}$$

where  $f^{(i)}$  denotes the  $i$ -th derivative of  $f$  and  $T_p^k f$  is the  $k$ -th order Taylor polynomial for  $f$  at  $p$ :

$$\begin{aligned} T_p^k f(x) &= \sum_{i=0}^k \frac{1}{i!} f^{(i)}(p) (x - p)^i \\ &= f(p) + f'(p)(x - p) + \frac{1}{2} f''(p)(x - p)^2 + \dots + \frac{1}{k!} f^{(k)}(p)(x - p)^k. \end{aligned}$$

PROOF: The proof is left as an exercise (with hints!) at the end of the chapter.  $\square$

**Theorem 3.6.** (Taylor's Theorem in several variables.) Let  $f: U \rightarrow \mathbb{R}$  where  $U$  is an open subset of  $\mathbb{R}^n$ . Suppose that the partial derivatives of  $f$  up to order  $k + 1$  exist and are continuous on  $U$ . Let  $p, x \in U$  and suppose that the line segment,  $p + tx$ ,  $0 \leq t \leq 1$ , is contained in  $U$ . Then

$$f(x) = T_p^k f(x) + r(x)$$

where  $T_p^k f$  is the  $k$ -th order Taylor polynomial for  $f$  at  $p$  and  $r$  is a function such that  $\lim_{x \rightarrow p} r(x)/|x - p|^k = \vec{0}$ .

PROOF: The details of this proof are not so important for your first pass through multivariable calculus. The basic idea is to restrict  $f$  to a line, thus reducing to the case of Taylor's theorem in one variable, then appeal to the chain rule. A sketch of the details appears below.

Define  $g(t) = p + tx$  for  $0 \leq t \leq 1$ . The function  $f \circ g$  is a real-valued function of one variable, and Taylor's theorem in one variable applies. We get that

$$(f \circ g)(t) = \sum_{i=0}^k \frac{1}{i!} \frac{d^i(f \circ g)}{dt^i}(0) t^i + \frac{1}{(k+1)!} \frac{d^{k+1}(f \circ g)}{dt^{k+1}}(c) t^{k+1}. \quad (8)$$

for some  $c \in (0, 1)$ . Now use the chain rule to calculate the derivatives. For instance, we have

$$\begin{aligned} \frac{d(f \circ g)}{dt}(t) &= \sum_{s_1=1}^n \frac{\partial f}{\partial x_{s_1}}(g(t)) x_{s_1} \\ \frac{d^2(f \circ g)}{dt^2}(t) &= \sum_{s_1=1}^n \sum_{s_2=1}^n \frac{\partial^2 f}{\partial x_{s_2}^2 \partial x_{s_1}^{s_1}}(g(t)) x_{s_2} x_{s_1} \\ \frac{d^3(f \circ g)}{dt^3}(t) &= \sum_{s_1=1}^n \sum_{s_2=1}^n \sum_{s_3=1}^n \frac{\partial^3 f}{\partial x_{s_3}^3 \partial x_{s_2}^2 \partial x_{s_1}^{s_1}}(g(t)) x_{s_3} x_{s_2} x_{s_1}. \end{aligned}$$

Notice that in the third derivative, for instance, the term  $x_1 x_2 x_3$  will occur  $3! = 6$  times: once for each way of setting  $s_1, s_2$ , and  $s_3$  equal to 1, 2, and 3 in some order. In Taylor's formula for one variable, displayed above as equation 8, we divide by  $3!$  for the third derivative, and this accounts for the fact that the term involving  $x_1 x_2 x_3$  in Taylor's formula for several variables occurs once. On the other hand, the term involving  $x_1^3$  only occurs once in the third derivative, when  $s_1 = s_2 = s_3 = 1$ . Thus, the corresponding term in the Taylor's formula for several variables occurs with a  $1/3!$  in front. Careful counting reveals that the number of times  $x_1^{i_1} \cdots x_n^{i_n}$  occurs in the formula for the  $i$ -th derivative is  $i!/(i_1! \cdots i_n!)$ , accounting for the factor of  $1/(i_1! \cdots i_n!)$  in Taylor's formula for several variables. (Technical detail: note that we have used that  $\partial^2 f / \partial x_i \partial x_j = \partial^2 f / \partial x_j \partial x_i$ , Theorem 7.4, in this reasoning.)

Evaluating equation 8 at  $t = 1$  gives a formula for  $f(g(1)) = f(p+x)$ . Substituting  $x-p$  for  $x$  then gives  $f(x) = T_p^k f(x) + r(x)$  where the remainder,  $r$ , involves the  $k+1$ -th derivative of  $f \circ g$ . To explain why  $\lim_{x \rightarrow \bar{0}} r(x)/|x-p|^k = 0$ , you can expand the derivative and note that the resulting expression involves only terms of total degree  $k+1$  in the  $x_i$ . You also need to note that the  $(k+1)$ -st partials of  $f$  are bounded near  $p$  since they are continuous.  $\square$

**Example 3.7.** Let  $f(x, y) = (e^{2x+y}, -3y + x^2 + xy)$  and  $p = (0, 0)$ . To find the second order Taylor polynomial, we calculate the necessary derivatives:

$$\begin{aligned} f(0, 0) &= (1, 0), & \frac{\partial f}{\partial x}(0, 0) &= (2, 0), & \frac{\partial f}{\partial y}(0, 0) &= (1, -3) \\ \frac{\partial^2 f}{\partial x^2}(0, 0) &= (4, 2), & \frac{\partial^2 f}{\partial x \partial y}(0, 0) &= (2, 1), & \frac{\partial^2 f}{\partial y^2}(0, 0) &= (1, 0). \end{aligned}$$

Therefore,

$$\begin{aligned} T_p^2 f(x, y) &= (1, 0) + (2, 0)x + (1, -3)y + \frac{1}{2!}(4, 2)x^2 + (2, 1)xy + \frac{1}{2!}(1, 0)y^2 \\ &= (1 + 2x + y + 2x^2 + 2xy + y^2/2, -3y + x^2 + xy). \end{aligned}$$

The second component of  $f$  is the same as that of  $T_p^2 f$ , as might be expected.

The first order Taylor polynomial is

$$T_p^1 f(x, y) = (1, 0) + (2, 0)x + (1, -3)y = (1 + 2x + y, -3y),$$

the best affine approximation of  $f$  at  $(0, 0)$ .

## 4. Maxima and minima

### 4.1. Basic definitions, first results.

**Definition 4.1.** A point  $s \in S \subseteq \mathbb{R}^n$  is in the interior of  $S$  if there is a nonempty open ball centered at  $s$  contained entirely in  $S$ . Otherwise, if  $s$  is a limit point of both  $S$  and the complement of  $S$ , then it is called a boundary point of  $S$ .

**Definition 4.2.** Let  $f: S \rightarrow \mathbb{R}$  where  $S \subset \mathbb{R}^n$ . Let  $s \in S$ .

- (1)  $f(s)$  is a (global) maximum for  $f$  if  $f(s) \geq f(s')$  for all  $s' \in S$ .
- (2)  $f(s)$  is a (global) minimum for  $f$  if  $f(s) \leq f(s')$  for all  $s' \in S$ .
- (3)  $f(s)$  is a local maximum for  $f$  if  $s$  is an interior point and  $f(s) \geq f(s')$  for all  $s'$  in some nonempty open ball centered at  $s$ .
- (4)  $f(s)$  is a local minimum for  $f$  if  $s$  is an interior point and  $f(s) \leq f(s')$  for all  $s'$  in some nonempty open ball centered at  $s$ .
- (5) A global extremum for  $f$  is a global maximum or global minimum for  $f$ . A local extremum for  $f$  is a local maximum or local minimum for  $f$ .

Since the gradient of a real-valued function points in the direction of quickest increase of the function (and its negative points in the direction of quickest decrease) one would expect that the gradient would vanish when it reaches a local extrema. This turns out to be the case, as we prove below.

**Definition 4.3.** Let  $f: S \rightarrow \mathbb{R}$  where  $S \subseteq \mathbb{R}^n$ . A point  $s \in S$  is a critical or stationary point of  $f$  if it is an interior point,  $f$  is differentiable at  $s$ , and  $\nabla f(s) = 0$ , i.e., all the partial derivatives of  $f$  vanish at  $s$ .

**Theorem 4.4.** Let  $f: U \rightarrow \mathbb{R}$  be a differentiable function on an open set  $U \subseteq \mathbb{R}^n$ , and let  $p \in U$ . If  $f(p)$  is a local extremum for  $f$ , then  $p$  is a critical point for  $f$ .

PROOF: Let  $v$  be any vector in  $\mathbb{R}^n$  and define  $g(t) = p + tv$  for  $t$  small enough so that the image of  $g$  is contained in  $U$ . Since  $f(p)$  is a local extremum for  $f$ , it follows that  $f(g(0))$  is a local extremum for  $f \circ g$ . By one variable calculus\*,  $(f \circ g)'(0) = 0$ . It follows from the chain rule that  $\nabla f(p) \cdot v = 0$ . Since  $v$  is arbitrary,  $\nabla f(p) = \vec{0}$ . For instance, if  $v = e_i$ , the  $i$ -th standard basis vector, then  $\nabla f(p) \cdot v = 0$  implies that the  $i$ -component of  $\nabla f(p)$  is zero, i.e.,  $\partial f(p)/\partial x_i = 0$ .  $\square$

The converse to this theorem is not true, as you already know from one variable calculus. For example, the function  $f(x) = x^3$  has a critical point at  $x = 0$  even though  $f(0) = 0$  is not a local extremum.

**Definition 4.5.** A critical point whose value is not a local extremum is called a saddle point.

\*The Newtonian quotient will be non-negative on one side of  $t = 0$  and non-positive on the other, so the only possibility for the limit is zero.

To determine the shape of a function near a critical point, we need to look at higher order derivatives.

**4.2. Analyzing critical points using Taylor's theorem.** By definition, at a critical point, the first partials of a function vanish. Hence, the best affine approximation (the first order Taylor polynomial) is just a constant function, indicating that the function has “flattened out.” We need to use the higher order derivatives to see how the function flexes nearby.

**Example 4.6.** Let  $f(x, y) = x^2 - 3xy + y^3$  for  $(x, y) \in \mathbb{R}^2$ . We would like to find all local extrema of  $f$ . Since each point in the domain of  $f$  is an interior point, Theorem 4.4 says that the extrema must occur at critical points. So we should solve the equation(s)  $\nabla f(x, y) = \vec{0}$ :

$$\left. \begin{array}{l} \frac{\partial f}{\partial x} = 2x - 3y = 0 \\ \frac{\partial f}{\partial y} = -3x + 3y^2 = 0 \end{array} \right\} \implies \left. \begin{array}{l} y = \frac{2}{3}x \\ x = y^2 \end{array} \right\} \implies (x, y) = (0, 0) \text{ or } \left(\frac{9}{4}, \frac{3}{2}\right).$$

To analyze the behavior near these points, calculate the second order Taylor polynomials. The first partials vanish at the critical points (that's how we just found them); the second order partials are

$$\frac{\partial^2 f}{\partial x^2} = 2, \frac{\partial^2 f}{\partial x \partial y} = -3, \frac{\partial^2 f}{\partial y^2} = 6y,$$

At  $(0, 0)$  we get

$$\begin{aligned} f(x, y) &\approx T_{(0,0)}^2 f(x, y) \\ &= f(0, 0) + \frac{\partial f}{\partial x}(0, 0)x + \frac{\partial f}{\partial y}(0, 0)y \\ &\quad + \frac{1}{2!0!} \frac{\partial^2 f}{\partial x^2}(0, 0)x^2 + \frac{1}{1!1!} \frac{\partial^2 f}{\partial x \partial y}(0, 0)xy + \frac{1}{0!2!} \frac{\partial^2 f}{\partial y^2}(0, 0)y^2 \\ &= x^2 - 3xy. \end{aligned}$$

Now complete the square:

$$T_{(0,0)}^2 f(x, y) = x^2 - 3xy = x^2 - 3xy + \left(\frac{3}{2}\right)^2 y^2 - \left(\frac{3}{2}\right)^2 y^2 = \left(x + \frac{3}{2}y\right)^2 - \frac{9}{4}y^2.$$

After the linear change of coordinates  $x \leftrightarrow x - \frac{3}{2}y$ ,  $y \leftrightarrow \frac{3}{2}y$ , the Taylor polynomial takes the form  $(x, y) \mapsto x^2 - y^2$ , a saddle. It turns out that this linear change of coordinates does not change the nature of the critical point; hence,  $(0, 0)$  is a saddle point.

Now consider the remaining critical point,  $\left(\frac{9}{4}, \frac{3}{2}\right)$ :

$$\begin{aligned} f(x, y) &\approx T_{\left(\frac{9}{4}, \frac{3}{2}\right)}^2 f(x, y) \\ &= -\frac{27}{16} + \left(x - \frac{9}{4}\right)^2 - 3\left(x - \frac{9}{4}\right)\left(y - \frac{3}{2}\right) + \frac{9}{2}\left(y - \frac{3}{2}\right)^2. \end{aligned}$$

To determine the shape of  $f$  near the critical point, consider the behavior of the following function near  $(0, 0)$ :

$$Q(x, y) = x^2 - 3xy + \frac{9}{2}y^2.$$

We have dropped the constant  $-27/16$  and made a translation to get  $Q$ , neither of which changes the nature of the critical point. Now complete the square:

$$\begin{aligned} Q(x, y) &= x^2 - 3xy + \frac{9}{2}y^2 \\ &= x^2 - 3xy + \left(\frac{3}{2}\right)^2 y^2 - \left(\frac{3}{2}\right)^2 y^2 + \frac{9}{2}y^2 \\ &= \left(x - \frac{3}{2}y\right)^2 + \frac{9}{4}y^2. \end{aligned}$$

After a linear change of coordinates, as before, this function becomes  $(x, y) \mapsto x^2 + y^2$ , from which it is easy to see that  $(0, 0)$  is a local minimum. Therefore, the critical point  $(\frac{9}{4}, \frac{3}{2})$  yields a local minimum of  $f$ .

As in the previous example, we will in general be able to analyze the behavior of a function at a critical point by looking at an associated quadratic function.

**Definition 4.7.** Let  $f: U \rightarrow \mathbb{R}$  be a function on an open set  $U \subseteq \mathbb{R}^n$  with partial derivatives up to order 2. Let  $p \in U$  be a critical point of  $f$ . The quadratic form for  $f$  at  $p$  is defined by  $Q_p f(x) := T_p^2 f(x + p)$ , where  $T_p^2 f$  is the second order Taylor polynomial for  $f$  at  $p$ . Hence,

$$Q_p f(x_1, \dots, x_n) := \sum_{i_1 + \dots + i_n = 2} \frac{1}{i_1! \dots i_n!} \frac{\partial^2 f}{\partial x_1^{i_1} \dots \partial x_n^{i_n}}(p) x_1^{i_1} \dots x_n^{i_n}.$$

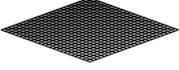
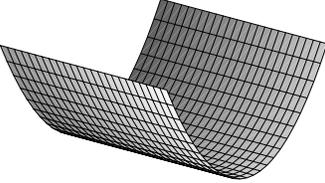
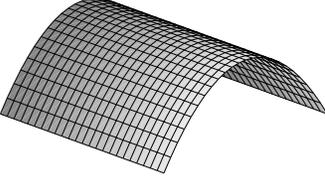
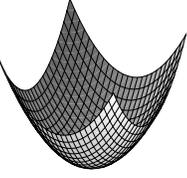
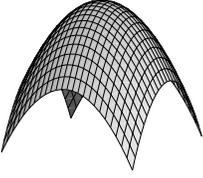
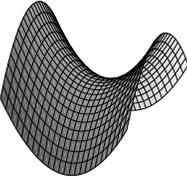
For instance, in the previous example,  $Q_{(0,0)} f(x, y) = x^2 - 3xy$  and  $Q_{(9/4, 3/2)} f(x, y) = x^2 - 3xy + \frac{9}{2}y^2$ .

4.2.1. *Algorithm for analyzing critical points.* Let  $f: U \rightarrow \mathbb{R}$  with  $U$  an open subset of  $\mathbb{R}^n$ . Let  $p \in U$  and suppose that  $f$  has continuous partial derivatives up to order three in an open set about  $p$ . Without proof, we state the following algorithm for finding the local minima and maxima of  $f$  on  $U$ . (Note that  $U$  is an open set, so each of its points is an interior point. If  $U$  were not open, we will see later in this chapter that it would be possible for  $f$  to have extrema that do not occur at critical points.)

- (1) Calculate the critical points by setting  $\nabla f = \vec{0}$ .
- (2) For each critical point  $p$ , calculate the second order Taylor polynomial  $T_p^2 f$  for  $f$  at  $p$ . (Because  $p$  is a critical point, it will have no linear part.)
- (3) After translating, i.e., forgetting the constant term, and making an affine change of variables in  $T_p^2 f$ , we get the quadratic form,  $Q_p f$ , for  $f$  at  $p$ .
- (4) Completing the square in the quadratic form, if necessary, and making the obvious change of variables,  $Q_p f$  can be transformed into a function  $\tilde{Q}_p f$  having one of the forms listed below. The consequence for the nature of  $p$  is also listed.

$\tilde{Q}_p f$	behavior of $f$ at $p$
0	inconclusive
$x_1^2 + x_2^2 + \dots + x_k^2, \quad k < n$	inconclusive
$-x_1^2 - x_2^2 - \dots - x_k^2, \quad k < n$	inconclusive
$x_1^2 + \dots + x_s^2 - x_{s+1}^2 - \dots - x_k^2, \quad 1 < s < k \leq n$	saddle
$x_1^2 + \dots + x_n^2$	local minimum
$-x_1^2 - \dots - x_n^2$	local maximum

In the two dimensional case, the possibilities for  $\tilde{Q}_p f$  are summarized in the following table:

$\tilde{Q}_p f$	CONCLUSION	GRAPH
0	need higher order terms	
$x^2$	need higher order terms	
$-x^2$	need higher order terms	
$x^2 + y^2$	local min.	
$-x^2 - y^2$	local max.	
$x^2 - y^2$	saddle	

**Note:** The quadratic form  $Q(x, y) = xy$  can be put into standard form by making the change of variables  $x \leftrightarrow x - y$ ,  $y \leftrightarrow x + y$ , to get  $\tilde{Q}(x, y) = (x - y)(x + y) = x^2 - y^2$ , a

saddle. Similarly, factor the  $y$  out of  $Q(x, y) = xy - 3y^2$  to get  $(x - 3y)y \rightsquigarrow xy \rightsquigarrow x^2 - y^2$ , a saddle again (the  $\rightsquigarrow$  denotes a suitable change of variables).

#### 4.2.2. More examples.

**Example 4.8.** Define  $f(x, y) = x^2 + y^3$  and  $g(x, y) = x^2 + y^4$ . The point  $(0, 0)$  is a critical point for both of these functions, and both have the same second order Taylor polynomial there, namely  $Q(x, y) = x^2$ . However, for  $f$ , the point  $(0, 0)$  is a saddle point and for  $g$ , it gives a minimum. Thus, this is an example where the quadratic form does not determine the behavior of the critical point. On the other hand, if  $f(x, y) = x^2 + y^2 + y^3$  and  $g(x, y) = x^2 + y^2 + y^4$ , both  $f$  and  $g$  would have the same quadratic form at the critical point  $(0, 0)$ , namely  $Q(x, y) = x^2 + y^2$ , but this time, we could conclude that  $(0, 0)$  gives a local minimum for both functions.

**Example 4.9.** Let  $f(x, y) = \frac{xye^x}{1+y^2}$ . The critical points are found by solving the equations

$$\frac{\partial f}{\partial x}(x, y) = \frac{(y + xy)e^x}{1 + y^2} = 0 \quad \text{and} \quad \frac{\partial f}{\partial y}(x, y) = -\frac{(-1 + y^2)xe^x}{(1 + y^2)^2} = 0.$$

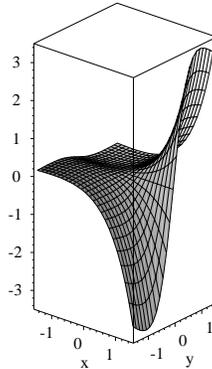
The solutions are  $(0, 0)$ ,  $(1, -1)$ , and  $(-1, -1)$ . We will analyze each.

At  $p = (0, 0)$ , we have  $\frac{\partial^2 f}{\partial x^2}(p) = \frac{\partial^2 f}{\partial y^2}(p) = 0$ , and  $\frac{\partial^2 f}{\partial x \partial y}(p) = 1$ . Hence,  $Q_p f = xy$ ; so  $(0, 0)$  is a saddle.

At  $p = (1, -1)$ , we have  $\frac{\partial^2 f}{\partial x^2}(p) = \frac{1}{2e}$ ,  $\frac{\partial^2 f}{\partial x \partial y}(p) = 0$ , and  $\frac{\partial^2 f}{\partial y^2}(p) = \frac{1}{2e}$ . Hence,  $Q_p f = \frac{1}{4e}x^2 + \frac{1}{4e}y^2$ ; so  $(1, -1)$  yields a local minimum for  $f$ .

At  $p = (-1, -1)$ , we have  $\frac{\partial^2 f}{\partial x^2}(p) = -\frac{1}{2e}$ ,  $\frac{\partial^2 f}{\partial x \partial y}(p) = 0$ , and  $\frac{\partial^2 f}{\partial y^2}(p) = -\frac{1}{2e}$ . Hence,  $Q_p f = -\frac{1}{4e}x^2 - \frac{1}{4e}y^2$ ; so  $(-1, -1)$  gives a local maximum for  $f$ .

Here is a piece of the graph, which at least does not contradict our findings:



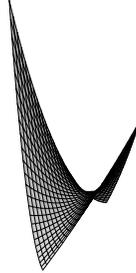
**Example 4.10.** Let  $f(x, y) = 3 - 2x + 2y + x^2y^2$ . There is only one critical point:

$$\left. \begin{aligned} \frac{\partial f}{\partial x} &= -2 + 2xy^2 = 0 \\ \frac{\partial f}{\partial y} &= 2 + 2x^2y = 0 \end{aligned} \right\} \implies x = 1, y = -1.$$

The quadratic form at the critical point is

$$Q(x, y) = x^2 - 4xy + y^2 = (x - 2y)^2 - 3y^2,$$

which looks like  $(x, y) \mapsto x^2 - y^2$ , a saddle. Here is a piece of the graph near  $(1, -1)$ :



### 4.3. An existence theorem.

**Definition 4.11.** A subset  $S \subset \mathbb{R}^n$  is bounded if there is a constant  $M$  such that  $|s| < M$  for all  $s \in S$ ; in other words  $S$  is contained in  $B_M(\vec{0})$ , the open ball of radius  $M$  centered at  $\vec{0}$ .

The following theorem is proved in a course in advanced calculus (e.g., see Rudin's *Principles of Mathematical Analysis*, Theorem 4.14.) Recall that a set is *closed* if its complement is open (Definition 2.8). Equivalently, a set is closed if it contains all of its limit points.

**Theorem 4.12.** (*Extreme value theorem.*) Let  $S \subset \mathbb{R}^n$  be a closed and bounded set, and let  $f: S \rightarrow \mathbb{R}$  be a continuous function. Then the image of  $f$  is closed and bounded. In particular,  $f$  has a global maximum and a global minimum.

Now, for the first time, we have a theorem that guarantees the existence of extrema a priori. Our previous result, Theorem 4.4, says that interior points which give extrema must be critical points. Note two problems with that theorem if we are trying to establish the existence of extrema. First, the theorem only applies to interior points, not to points that might be on the boundary of the function's domain. Second, even if the critical points exist, they could all turn out to be saddles. By combining the extreme value theorem with the previous theorem on critical points, we can refine our search for extrema.

4.3.1. *Strategy for finding extrema.* Let  $f: S \rightarrow \mathbb{R}$  be a continuous function, differentiable on its interior.

- (1) By some means, determine a closed and bounded set in  $S$ —perhaps a box or a sphere—which contains points yielding the extrema of  $f$ , at least those extrema you find interesting. Call this set  $C$ . For instance, if you are looking for a global minimum, you could try to find  $C$  such that  $f$  is negative at some points in  $C$  and  $f$  is positive outside of  $C$ . You are then assured that  $f$  will have a global minimum and that it occurs at a point in  $C$ . If  $S$  itself is closed and bounded, you might choose  $C = S$ .
- (2) Theorem 4.12 guarantees the existence of a global extrema of the function  $f$  with its domain restricted to  $C$ . In other words, consider  $f$  now as a function  $f: C \rightarrow \mathbb{R}$  and forget about the rest of the domain. Now look for extrema in two places: in the interior of  $C$  and on the boundary of  $C$ .
- (3) In the interior of  $C$ , any extremum must occur at a critical point. So apply our previous algorithm to analyze those points. The local extrema you find are potential global extrema. Save them for now.
- (4) Now consider the boundary of  $C$ . It will usually consist of pieces, each of which has smaller “dimension.” For example, if  $C$  is a box in  $\mathbb{R}^3$ , the boundary will consist

of six squares. By restricting  $f$  to each piece of the boundary, we end up with functions of one fewer variable, each of which can be analyzed just like  $f$  itself. For example, if  $f$  is restricted to a square, look for critical points in the interior of the square, then restrict  $f$  further to the sides of the square. Restricting  $f$  to each side of the square, look for critical points in the interior of these line segments, then check the endpoints. In this way, we get a sort of recursive search for the extrema of  $f$ .

- (5) In the end, we get a list of points which potentially yield global extrema. Just evaluate  $f$  at each of these points to see which give the actual extrema. If you are only interested in global extrema, as opposed to local extrema, these will be evident without having to compute Taylor polynomials.

**Example 4.13.** Let  $f(x, y) = xy$ . The graph of  $f$  is a saddle surface, so it clearly has no extrema, even local extrema. However, if we restrict  $f$  to a closed and bounded set, it must have extrema. For instance, let  $S = \{(x, y) \in \mathbb{R}^2 \mid x, y \in [0, 1]\}$ , a closed unit square. To find the extrema of  $f$  on this square, first note that there are no critical points for  $f$  in the interior of the box. The gradient is zero only at the origin, which is on the boundary of the box. So we now restrict  $f$  to the boundary. The boundary consists of the four sides of the square. First, consider the side  $\{(x, 0) \in \mathbb{R}^2 \mid x \in [0, 1]\}$ . The function  $f$  restricted to this side is a function of the single variable  $x$  and in fact is just the zero function:  $f(x, 0) = 0$ . Similarly,  $f$  restricted to the side for which the first coordinate is 0 is also the zero function:  $f(0, y) = 0$ . Now consider the side of the square for which the first coordinate is 1. We have  $f(1, y) = y$ , which is minimized when  $y = 0$  and maximized when  $y = 1$ . Setting the second coordinate equal to 1 yields  $f(x, 1) = x$ , with the same result. In the end, we find that there are no local extrema in the interior, and that the extrema on the boundary occur at  $(0, 0)$  and  $(1, 1)$ . At these points,  $f$  is 0 and 1 respectively. The global maximum of  $f$  on  $S$  is thus 1, and occurs at  $(1, 1)$ . The global minimum on  $S$  is 0 and occurs at  $(0, 0)$ .

**Example 4.14.** Let  $f(x, y) = 144x^3y^2(1 - x - y)$  for  $x \geq 0$  and  $y \geq 0$ . So the domain of  $f$  is just one quadrant of the plane. Consider the triangle  $T$  bounded by the three lines  $x = 0$ ,  $y = 0$ , and  $x + y = 1$ . Outside of this triangle (but in the given quadrant), the function is never positive since at these points  $1 - x - y < 0$  while  $x$  and  $y$  are non-negative. Along the edges of the triangle, the function is 0. In the interior of the triangle the function is positive. Since  $f$  is continuous, the extreme value theorem guarantees the existence of a global maximum of  $f$  on  $T$ . To find it, we first look for critical points of  $f$  on  $T$

$$\begin{aligned}\nabla f(x, y) &= (3 \cdot 144x^2y^2(1 - x - y) - 144x^3y^2, 2 \cdot 144x^3y(1 - x - y) - 144x^3y^2) \\ &= (0, 0)\end{aligned}$$

$$\implies 3 \cdot 144x^2y^2(1 - x - y) = 144x^3y^2 \quad \text{and} \quad 2 \cdot 144x^3y(1 - x - y) = 144x^3y^2.$$

Since critical points are by definition interior points, we may assume that  $x \neq 0$  and  $y \neq 0$ . (At any rate, we know that the global maximum is going to be positive, so we can exclude points along the axes.) Thus, we find the critical points on  $T$  by solving a system of linear equations:

$$\begin{aligned}3(1 - x - y) &= x &\implies 4x + 3y &= 3 \\ 2(1 - x - y) &= y &\implies 2x + 3y &= 2\end{aligned} \implies (x, y) = \left(\frac{1}{2}, \frac{1}{3}\right).$$

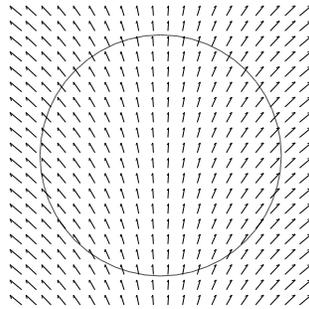
Since we know that  $f$  has a global maximum in the interior of  $T$ , and since  $(\frac{1}{2}, \frac{1}{3})$  is the sole critical point of  $f$  inside  $T$ , we know it gives the global maximum of  $f$  without having to analyze the Taylor function for  $f$  there. The global maximum is thus  $f(\frac{1}{2}, \frac{1}{3})$ . (Note:

even if  $f$  had had several critical points inside  $T$ , we would only need to evaluate  $f$  at these points to determine the global maximum; we would not need to calculate Taylor series.)

**4.4. Lagrange multipliers.** Let  $f: S \rightarrow \mathbb{R}$  with  $S \subseteq \mathbb{R}^n$ . Imagine that  $S$  is an ocean and that you are floating in a boat in  $S$ , trying to find the extrema of  $f$ . Luckily, there is a current in the ocean, given by the gradient of  $f$ . Thus, if you just go with the flow, your boat will naturally glide in the direction of quickest increase of the function  $f$ . If you are lucky, the boat will stop at some point, and you will have found a critical point of  $f$ . You jot down the coordinates, then paddle away from the critical point, hoping to catch a current leading you to another critical point.

Now imagine another scenario. Suppose  $S$  is a swimming pool. As you follow the flow of the gradient there are now two possibilities. You may be swept to a critical point, just as before. The other possibility is that you crash into the side of the pool. What happens then? Well, the flow is pushing you into the wall, but it may have some component along the wall which causes your boat to slide. You continue to slide until the current is pushing perpendicularly into the wall. There is now no direction in which you can flow (without leaving the pool) in which the function can increase. This is the main idea behind the method of *Lagrange multipliers*.

**Example 4.15.** Let  $f(x, y) = x^2 + 2y$  constrained to the unit disc  $x^2 + y^2 \leq 1$ . So in this case, the swimming pool is circular. Since  $f$  is continuous and the domain is closed and bounded, we know that  $f$  attains its maximum someplace on the disc. The gradient of  $f$  is  $\nabla f(x, y) = (2x, 2) = 2(x, 1)$ :



The function has no critical points, so the maximum must occur someplace on the boundary. From the picture, it looks like a boat starting at any point in this flow would be pushed against the wall then would slide up to the point  $(0, 1)$ . To prove this, define  $g(x, y) = x^2 + y^2$ . The boundary of the pool is the level set  $g = 1$ . The gradient of  $g$  is  $\nabla g(x, y) = 2(x, y)$ , and is perpendicular to the level set through  $(x, y)$ . So at which points does our flow push straight into the wall? Since  $\nabla g$  points away from the origin, the answer is: at exactly those points along the boundary at which  $\nabla g(x, y)$  points in the same direction as the  $\nabla f(x, y)$ . To “point in the same direction” means that there is a *positive* scalar  $\lambda$  such that  $\nabla g(x, y) = \lambda \nabla f(x, y)$ . So we need to solve a system of equations:

$$x^2 + y^2 = 1, \quad \nabla g(x, y) = (2x, 2y) = \lambda \nabla f(x, y) = \lambda(2x, 2).$$

The only solution for which  $\lambda > 0$  is  $(0, 1)$ , as expected. Thus, the maximum value for  $f$  is  $f(0, 1) = 2$ .

To find the minimum of  $f$ , we would follow the flow of  $-\nabla f(x, y)$ . A similar analysis would lead to the same set of equations:  $x^2 + y^2 = 1$  and  $\nabla g(x, y) = \lambda \nabla f(x, y)$ , but this time we would be looking for a solution for which  $\lambda$  is negative. It turns out that the

unique minimum is  $f(0, -1) = -2$ . If it were not obvious from the start that  $\nabla g$  pointed away from the region rather than into it, we could have just solved the system of equations without worrying about the sign of  $\lambda$ . In the end, we would find both the maximum and the minimum, and they would be easily distinguishable.

Of course, an entirely different way to proceed would be to parametrize the boundary and substitute into  $f$ . For instance, letting  $x = \cos \theta$  and  $y = \sin \theta$ , we get  $f(\cos \theta, \sin \theta) = \cos^2 \theta + 2 \sin \theta$ . We have reduced the problem to maximizing a function of one variable. (Check that this gives the same result.)

To summarize the previous example, to find the extrema of a function  $f(x)$  as  $x \in \mathbb{R}^n$  varies over a level set of some constraining function, say  $g(x) = c$ , solve the system of equations

$$g(x) = c, \quad \nabla g(x) = \lambda \nabla f(x).$$

The local extremum for  $f$  will appear among the solutions. The  $\lambda$  that appear are called *Lagrange multipliers*. Just as was the case with saddle points before, not all solutions need correspond to extrema.

**Example 4.16.** Maximize the function  $f(x, y, z) = x^2 + 2y^2 + 3z^2$  subject to the constraint  $x^2 + y^2 + z^2 = 4$ .

SOLUTION: Define  $g(x, y, z) = x^2 + y^2 + z^2$ . To maximize  $f$  along the level set  $g = 4$ , solve the equations

$$x^2 + y^2 + z^2 = 4, \quad \nabla g(x, y, z) = (2x, 2y, 2z) = \lambda \nabla f(x, y, z) = \lambda(2x, 4y, 6z).$$

A bit of algebra gives six solutions:  $(\pm 2, 0, 0)$ ,  $(0, \pm 2, 0)$ , and  $(0, 0, \pm 2)$ . Since  $g = 4$  describes a closed and bounded set (a sphere of radius 2),  $f$  has a global maximum and global minimum when restricted to that set. The maximum is  $f(0, 0, \pm 2) = 12$  and the minimum is  $f(\pm 2, 0, 0) = 4$ .

The method of Lagrange multipliers can be generalized to deal with several simultaneous constraints. Consider the points  $x \in \mathbb{R}^n$  satisfying the system of equations  $g_1(x) = c_1, \dots, g_k(x) = c_k$ . Then if  $s \in \mathbb{R}^n$  is an extremum for  $f$  restricted by these constraints there are constants  $\lambda_1, \dots, \lambda_k$  such that

$$\nabla f(s) = \lambda_1 \nabla g_1(s) + \dots + \lambda_k \nabla g_k(s)$$

provided the vectors  $\nabla g_1(s), \dots, \nabla g_k(s)$  are linearly independent. For instance, suppose there are just two constraints  $g_1 = c_1$  and  $g_2 = c_2$ , each defining a surface in  $\mathbb{R}^3$ . The points satisfying both constraints lie along the intersection of these surfaces, which we suppose to be a curve. Using the swimming pool analogy again, the method relies on the fact that a boat pushed by the gradient of  $f$  will come to rest at a point on the side of the pool (the curve) provided the gradient of  $f$  lies in the plane spanned by the gradients of  $g_1$  and  $g_2$  at that point. The reason this works is that this plane is perpendicular to the curve. If interested, try drawing the picture.

**Example 4.17.** Find the minimum distance between the ellipse  $\frac{x^2}{4} + \frac{y^2}{9} = 1$  and the circle  $(x - 3)^2 + (y + 5)^2 = 1$ .

SOLUTION: Consider the function

$$f(x, y, u, v) = d((x, y), (u, v))^2 = (x - u)^2 + (y - v)^2.$$

Our problem is to minimize  $f$  subject to the constraints

$$g_1(x, y, u, v) = \frac{x^2}{4} + \frac{y^2}{9} = 1 \quad \text{and} \quad g_2(x, y, u, v) = (u - 3)^2 + (v + 5)^2 = 1.$$

The method of Lagrange multipliers says to look for solutions to the system

$$\begin{aligned} \nabla f(x, y, u, v) &= 2(x - u, y - v, u - x, v - y) \\ &= \lambda_1 \nabla g_1(x, y, u, v) + \lambda_2 \nabla g_2(x, y, u, v) \\ &= \lambda_1 \left( \frac{x}{2}, \frac{2y}{9}, 0, 0 \right) + \lambda_2 (0, 0, 2(u - 3), 2(v + 5)), \end{aligned}$$

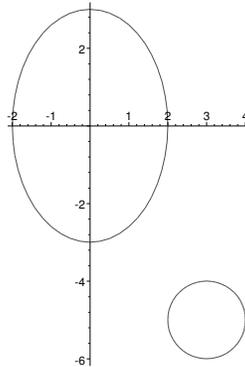
satisfying the constraints  $g_1 = 1$  and  $g_2 = 1$ . Altogether, we get a system of six equations:

$$\begin{aligned} x - u &= \lambda_1 x/4 & y - v &= \lambda_1 y/9 \\ u - x &= \lambda_2(u - 3) & v - y &= \lambda_2(v + 5) \\ x^2/4 + y^2/9 &= 1 & (u - 3)^2 + (v + 5)^2 &= 1. \end{aligned}$$

The function `NSolve` in the program Mathematica says there are four solutions

$$\begin{aligned} (x, y) &\approx (-0.58, 2.9), & (u, v) &\approx (3.4, -5.9) \\ (x, y) &\approx (-0.58, 2.9), & (u, v) &\approx (2.6, -4.1) \\ (x, y) &\approx (0.98, -2.6), & (u, v) &\approx (3.6, -5.9) \\ (x, y) &\approx (0.98, -2.6), & (u, v) &\approx (2.4, -4.2) \end{aligned}$$

Can you determine the geometrical meaning of these pairs of points?



4.4.1. *A final note.* All of the methods we have presented for finding extrema involve, at some point, solving a system of equations in several variables. For instance, to find critical points, you need to set each of the first partials of a function equal to zero. Another example appears just above, where the method of Lagrange multipliers yields a system of six equations. Unless you are very lucky, you will need to solve the system of equations using some type of numerical method on a computer; in other words, you will have to be content with approximations to the solutions.

## 5. The Hessian

In this section, we present the Hessian test for analyzing critical points for functions of two variables.

The *transpose* of an  $m \times n$  matrix  $A$  is the  $n \times m$  matrix  $A^T$  determined by

$$A_{ij}^T := A_{ji}.$$

For instance,

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} = \begin{pmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{pmatrix}.$$

A square matrix is *symmetric* if it is equal to its own transpose, as in the following example:

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 5 & 6 \\ 3 & 6 & 9 \end{pmatrix}.$$

With each *quadratic form*

$$Q(x_1, \dots, x_n) = \sum_{i=1}^n a_{ii}x_i^2 + \sum_{1 \leq i < j \leq n} a_{ij}x_i x_j,$$

there is an associated symmetric matrix,  $M = M_Q$ , whose  $ij$ -th entry is

$$M_{ij} = \begin{cases} a_{ii} & \text{if } i = j \\ a_{ij}/2 & \text{if } i < j \\ a_{ji}/2 & \text{if } j < i. \end{cases}$$

Thus, roughly, to form this matrix put the coefficients of the squared terms of  $Q$  along the diagonal, and put half the coefficients of the mixed terms in the corresponding off-diagonal positions. Equating  $x = (x_1, \dots, x_n)$  with the column matrix  $(x_1 \cdots x_n)^T$ , it is easily checked that

$$Q(x) = x^T M x.$$

For instance, the quadratic form  $Q(x, y) = x^2 + 5xy - 7y^2$  can be written

$$Q(x, y) = \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} 1 & 5/2 \\ 5/2 & -7 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix},$$

and  $Q(x, y, z) = x^2 - 2xy + xz + 7y^2 - 10yz + 3z^2$  can be written

$$Q(x, y, z) = \begin{pmatrix} x & y & z \end{pmatrix} \begin{pmatrix} 1 & -1 & 1/2 \\ -1 & 7 & -5 \\ 1/2 & -5 & 3 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

The so-called *spectral theorem* from linear algebra simplifies the study of quadratic forms

**Theorem 5.1.** *If  $M$  is a real symmetric matrix, there is a real orthogonal matrix  $O$  such that*

$$O^T M O = D,$$

where  $D$  is a diagonal matrix.

We pause for a few words of explanation. A *real matrix* is just a matrix with real entries. An *orthogonal matrix* is a matrix whose columns have unit length and are pair-wise perpendicular. (Thus,  $O^T$  is the inverse of  $O$ , i.e.,  $O^T O$  is the  $n \times n$  identity matrix.) A *diagonal matrix* is a matrix with nonzero entries only along the diagonal.

If  $O^T M O = D$ , then  $M = O D O^T$ , whence

$$Q(x) = x^T M x = x^T (O D O^T) x = (O^T x)^T M (O^T x).$$

(We have used the fact that if  $AB$  is a product of matrices, then  $(AB)^T = B^T A^T$ .) The important point here is that by replacing  $x$  by  $O^T x$ , i.e., making a certain linear change of coordinates, we have removed the cross terms of  $Q$ , which makes it easy to analyze the behavior of  $Q$ .

The diagonal entries of  $D$  are the *eigenvalues* of  $M$  and the number of negative eigenvalues is called the *index* of  $M$ . Let  $Q(x) = x^T M x$  be the quadratic form associated with  $M$ . If  $M$  has only positive eigenvalues, then  $Q(x) > 0$  for all non-zero vectors  $x$ . We say  $Q$  is *positive definite*. After changing coordinates in accord with the spectral theorem,  $Q$  takes the form  $\sum_{i=1}^n d_i x_i^2$  where the  $d_i$  are the diagonal entries of  $D$ . Thus, letting  $u_i = \sqrt{d_i} x_i$ , we get the quadratic form  $u_1^2 + u_2^2 + \cdots + u_n^2$ . If this quadratic form arose from analyzing a critical point of a function at some point, the point would give a local minimum for the function. Similarly, if  $M$  has only negative eigenvalues, then  $Q(x) < 0$  for all non-zero vectors  $x$ , and  $Q$  is called *negative definite*. After a change of coordinates,  $Q$  takes the form  $\sum_{i=1}^n d_i x_i^2$ , as before, but now all the  $d_i$  are negative. Letting  $u_i = -\sqrt{|d_i|}$ , we get the quadratic form  $-u_1^2 - u_2^2 - \cdots - u_n^2$ . In the context of analyzing a critical point, we would now be looking at a relative maximum. If  $D$  has some positive and some negative entries, then  $Q$  assumes positive and negative values, and after changing coordinates as above, we would get the quadratic form  $u_1^2 + \cdots + u_s^2 - u_{s+1}^2 - \cdots - u_{s+t}^2$  for some  $s$  and  $t$  such that  $1 < s < s+t \leq n$ . In this case,  $Q$  is called *indefinite*. If this quadratic form arises from analyzing a critical point, that point is a saddle (neither a local minimum nor a local maximum). Finally, if  $M$  has some zero eigenvalues (which is the case exactly when the determinant of  $M$  is zero), but the nonzero eigenvalues all have the same sign, we say  $Q$  is a *degenerate* quadratic form. After changing coordinates,  $Q$  has the form  $\pm(u_1^2 + \cdots + u_k^2)$  with  $k < n$ . If we were analyzing a critical point, we would need to look at higher order terms of the Taylor series of the function. We say the matrix  $M$  is positive definite if  $Q$  is positive definite, and similarly for the terms “negative definite”, “indefinite”, and “degenerate.”

Given a function  $f: \mathbb{R}^n \rightarrow \mathbb{R}$ , one defines the *Hessian* matrix for  $f$  to be the  $n \times n$  matrix,  $H = Hf$ , with  $ij$ -th entry

$$H_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}.$$

We assume  $f$  is sufficiently smooth so that  $H$  is symmetric. Thus, at any point  $p \in \mathbb{R}^n$ , we can evaluate  $H$  to get a real symmetric matrix, i.e., a symmetric matrix whose entries are real numbers. We denote this matrix by  $Hf(p)$ .

To analyze a critical point  $p$  of  $f$ , we start by calculating the quadratic form  $Q_p f$  for  $f$  at  $p$ , which is the second order term of the Taylor polynomial for  $f$  at  $p$  after translating to the origin. Let  $H = Hf(p)$  be the Hessian matrix for  $f$  at  $p$ . The reader should ascertain that the symmetric matrix associated with the quadratic form for  $f$  at  $p$  is  $\frac{1}{2}H$ :

$$Q_p f(x) = \frac{1}{2} x^T H x.$$

(Thus, one may write

$$T_p^2 f(x) = f(p) + \frac{1}{2} (x-p)^T H (x-p),$$

at  $p$ , and in fact,

$$T_q^2 f(x) = f(q) + \nabla f(q) \cdot (x-q) + \frac{1}{2} (x-q)^T H f(q) (x-q)$$

at an arbitrary point  $q \in \mathbb{R}^n$ .) To analyze the quadratic form it suffices to apply the spectral theorem. The critical point is a minimum if  $H$  is positive definite or a maximum if  $H$  is negative definite. If  $H$  is degenerate, one needs more terms from the Taylor series at  $p$ .

There is a simple test for determining the behavior of the Hessian matrix in the case  $n = 2$ .

**Theorem 5.2.** (Second derivative test) *Let  $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ , let  $p$  be a critical point of  $f$ , and let*

$$H = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$$

*be the Hessian for  $f$  at  $p$ . (Thus  $a = \partial^2 f(p)/\partial x^2$ ,  $b = \partial^2 f(p)/\partial x\partial y$ , and  $c = \partial^2 f(p)/\partial y^2$ .) Then*

- (1) *If  $ac - b^2 > 0$ , then (i)  $H$  is positive definite if  $a > 0$  ( $p$  is a local minimum) or (ii)  $H$  is negative definite if  $a < 0$  ( $p$  is a local maximum).*
- (2) *If  $ac - b^2 < 0$ , then  $H$  is indefinite ( $p$  is a saddle point).*
- (3) *If  $ac - b^2 = 0$ , then  $H$  is degenerate (more terms of the Taylor series are needed to determine the behavior of  $f$  at  $p$ ).*

In general, the determinant of the Hessian is called the *discriminant* the quadratic form for  $f$  at  $p$ . In the case of two variables, this theorem says that the discriminant often determines the nature of a critical point.

**Example 5.3.** We previously considered the polynomial  $f(x, y) = x^2 - 3xy + y^3$ . It has two critical points:  $(0, 0)$  and  $(9/4, 3/2)$ . The second partials of  $f$  are

$$\frac{\partial^2 f}{\partial x^2}(x, y) = 2, \quad \frac{\partial^2 f}{\partial x\partial y}(x, y) = -3, \quad \frac{\partial^2 f}{\partial y^2}(x, y) = 6y.$$

Thus, the Hessian for  $f$  is

$$Hf(x, y) = \begin{pmatrix} 2 & -3 \\ -3 & 6y \end{pmatrix}$$

and the discriminant is  $\det Hf(x, y) = 12y - 9$ .

At the critical point  $(0, 0)$ , the discriminant is  $-9$ . Hence,  $(0, 0)$  is a saddle. At  $(9/4, 3/2)$ , the discriminant is  $12(3/2) - 9 = 9 > 0$ . Since  $\partial^2 f/\partial x^2(9/4, 3/2) = 2 > 0$ , we conclude that  $(9/4, 3/2)$  is a local minimum.

## EXERCISES

- (1) Let  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  be a differentiable function, and let  $p \in \mathbb{R}^n$ . Let  $v \in \mathbb{R}^n$  be a unit vector in  $\mathbb{R}^n$ . Let  $c: \mathbb{R} \rightarrow \mathbb{R}^n$  be any differentiable curve satisfying  $c(0) = p$  and  $c'(0) = v$ . Define  $g = f \circ c$ , the restriction of  $f$  to  $c$  (the function  $g$  takes on exactly the values that  $f$  takes at points along the curve  $c$ ). Prove that  $f_v(p) = g'(0)$ .
- (2) Prove Theorem 3.5. Using the notation in Theorem 3.5, let  $x$  and  $p$  be fixed points in the interval  $[a, b]$ . Define the number  $M$  by

$$f(x) = T_p^k f(x) + M(x - p)^{k+1},$$

where

$$\begin{aligned} T_p^k f(x) &= \sum_{i=0}^k \frac{1}{i!} f^{(i)}(p)(x - p)^i \\ &= f(p) + f'(p)(x - p) + \frac{1}{2} f''(p)(x - p)^2 + \frac{1}{6} f'''(p)(x - p)^3 + \cdots \end{aligned}$$

is the  $k$ -th order Taylor polynomial for  $f$  at  $p$ . Note that since  $x$  and  $p$  are fixed,  $M$  is merely a number, not a function. To prove Taylor's theorem, we need to show that  $(k + 1)! M = f^{(k+1)}(c)$  for some  $c$  between  $x$  and  $p$ .

Now define the function

$$g(t) = f(t) - T_p^k f(t) - M(t - p)^{k+1}.$$

- (a) Explain why  $g^{(k+1)}(t) = f^{(k+1)}(t) - (k + 1)! M$ . (Thus, we are done if we show that  $g^{(k+1)}(c) = 0$  for some  $c$  between  $x$  and  $p$ .)
- (b) Show that  $g(p) = g'(p) = g''(p) = \cdots = g^{(k)}(p) = 0$ .
- (c) The mean value theorem states that if  $h(t)$  is a real-valued function, continuous on  $[\alpha, \beta]$  and differentiable on  $(\alpha, \beta)$ , then there is a point  $\gamma \in (\alpha, \beta)$  such that  $h'(\gamma)(\beta - \alpha) = h(\beta) - h(\alpha)$ . Use the mean value theorem to argue that  $g'(x_1) = 0$  for some  $x_1$  between  $x$  and  $p$ . Repeat the argument to show that  $g''(x_2) = 0$  for some  $x_2$  between  $p$  and  $x_1$ , and so on. Finally, we get that  $g^{(k+1)}(x_{k+1}) = 0$  for some  $x_{k+1}$  between  $p$  and  $x_k$ , concluding the proof.
- (3) Suppose  $f: \mathbb{R}^3 \rightarrow \mathbb{R}$ , and let  $p = (1, 2, 3)$ . Also suppose that

$$f(p) = 7, \quad \frac{\partial f}{\partial x}(p) = 2, \quad \frac{\partial f}{\partial z}(p) = -5$$

$$\frac{\partial^2 f}{\partial x^2}(p) = 6, \quad \frac{\partial^2 f}{\partial x \partial y}(p) = 7$$

$$\frac{\partial^3 f}{\partial x^2 \partial y}(p) = 8, \quad \frac{\partial^3 f}{\partial y^3}(p) = 24$$

$$\frac{\partial^4 f}{\partial x^2 \partial y^2}(p) = 14$$

$$\frac{\partial^5 f}{\partial x^5}(p) = 12, \quad \frac{\partial^5 f}{\partial x^3 \partial y^2}(p) = 48,$$

and that all other partial derivatives of  $f$  of order at most five vanish at  $p$ . What is,  $T_p^5 f$ , the fifth order Taylor polynomial for  $f$  at  $p = (1, 2, 3)$ ?

- (4) Calculate the  $k$ -th order Taylor polynomials,  $T_p^k f$ , for each of the following:

- (a)  $f(x) = \ln(x)$ ,  $p = 1$ ,  $k = \infty$ . Let  $x = 2$  to get a nice series for  $\ln(2)$ .
- (b)  $f(x) = x - 4x^2 + x^3$ ,  $p = 2$ ,  $k = 1, 2, 3, 4, 5$ . What happens if you multiply out these expressions, i.e. replace  $(x - 2)^2$  by  $x^2 - 4x + 4$ , etc?
- (c)  $f(x, y) = \sin(xy)$ ,  $p = (0, 0)$ ,  $k = 3$ . (Not all of the higher order partials vanish at  $p$ ).
- (d)  $f(x, y) = e^{x+y}$ ,  $p = (0, 0)$ ,  $k = \infty$ . What is the relation between this Taylor series and those for  $e^x$  and  $e^y$ ?
- (e)  $f(x, y) = ye^{xy}$ ,  $p = (1, 1)$ ,  $k = 2$ .
- (f)  $f(x, y) = 3 + x - 2xy + y^2$ ,  $p = (-1, 3)$ ,  $k = 2$ . As above, multiply out the expression you get to see that the Taylor series gives you the original function written in a different form (in a form which is useful for studying the function near the point  $(-1, 3)$ ).
- (5) Extended binomial theorem.
- (a) Find the Taylor series at 0 for  $(1 + x)^\alpha$ , where  $\alpha$  is any real number.
- (b) Use the  $n$ th-order Taylor polynomial to find fractions approximating  $\sqrt{2}$  for  $n = 1, \dots, 6$ . (Express your answers as fractions, not decimals.)
- (c) Show that the Taylor series in (a) reduces to the normal binomial theorem when  $\alpha$  is a positive integer.
- (6) In each of the following, suppose  $Q$  is the quadratic form for a function at a critical point. For each, (i) complete the square and change coordinates as necessary to express  $Q$  equivalently in the form  $\sum_{i=1}^n b_i x_i^2$  where each  $b_i = \pm 1$ ; and (ii) state whether the original critical point is a local minimum, a local maximum, or a saddle point.
- (a)  $x^2 - 5xy + 2y^2$
- (b)  $2x^2 - 6xy + 5y^2$
- (c)  $5xy$
- (d)  $x^2 + 3xy - y^2 + 4yz + z^2$
- (7) Let  $f(x, y) = x^2y^2 - 4x^2y + 4x^2 - 2xy^2 + 8xy - 8x + y^2 - 4y + 4$
- (a) Show that  $(1, 2)$  is a critical point of  $f$ .
- (b) Is  $(1, 2)$  a local minimum, a local maximum, or a saddle?
- (8) Define  $f(x, y) = x^2 + y^3 + 3xy^2 - 2x$ . Find the three critical points of  $f$  and analyze each by looking at the quadratic term in the Taylor expansions.
- (9) Let  $S := \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 4\}$ , a unit disc of radius 2. Define  $f: S \rightarrow \mathbb{R}$  by  $f(x, y) = xy$ .
- (a) How do you know, without calculating, that  $f$  has global extrema on  $S$ ?
- (b) Find the global extrema.
- (10) Find the extrema of the function  $f(x, y) = xy(4x^2 + y^2 - 16)$  on the quarter ellipse
- $$E = \{(x, y) \in \mathbb{R}^2 \mid x \geq 0, y \geq 0, 4x^2 + y^2 \leq 16\}.$$
- (11) Find and analyze the local extrema of  $f(x, y) = x^2 + xy - 4x + \frac{3}{2}y^2 - 7y$ .
- (12) Let  $f(x, y) = \frac{1}{3}x^3 + \frac{1}{3}y^3 + (x - \frac{3}{2})^2 - (y + 4)^2$ .
- (a) Find and analyze each critical point.
- (b) Does the function have global extrema?
- (13) Find and analyze the critical points of each of the following functions
- (a)  $f(x, y) = x^2y + xy^2$ .
- (b)  $g(x, y) = e^{x+y}$ .

- (c)  $h(x, y) = x^5y + xy^5 + xy$ .
- (14) The function  $h(x, y, z, t) = x^2 - 3xy + 2xt + y^2 + 2yz + 3z^2 - zt$  has a critical point at the origin. Does it give a local minimum, a local maximum, or is it a saddle point?
- (15) Discuss the local and global extrema of  $f(x, y) = \frac{1}{x-1} - \frac{1}{y-1}$  on the open unit ball  $B_1(0, 0)$  in  $\mathbb{R}^2$ .
- (16) The graph of the function  $m(x, y) = 6xy^2 - 2x^3 - 3y^4$  is called a *monkey saddle*. (You may want to have a computer draw it for you.) Find and analyze each of its three critical points.
- (17) (a) Use the method of Lagrange multipliers to minimize  $f(x, y) = x^2 + y^2$  subject to the constraint  $3x + 5y = 8$ .  
 (b) Draw a picture of the line and the gradient field for  $f$ , indicating the point where  $f$  is minimized.
- (18) Use the method of Lagrange multipliers to maximize  $f(x, y, z) = x - 2y + 2z$  on the sphere of radius 3 centered at the origin.
- (19) Use the method of Lagrange multipliers to maximize  $f(x, y, z) = xy + yz$  subject to the constraints  $x^2 + y^2 = 2$  and  $yz = 2$ .
- (20) (a) Use the method of Lagrange multipliers to find the minimum distance between the ellipse  $x^2 + 2y^2 = 1$  and the line  $x + y = 4$ , i.e., how close can a point on the ellipse get to a point on the line?  
 (b) Draw a picture illustrating your solution.
- (21) Let  $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ , and let  $p \in \mathbb{R}^2$ . Let

$$a = \frac{\partial^2 f}{\partial x^2}(p), \quad b = \frac{\partial^2 f}{\partial x \partial y}(p), \quad c = \frac{\partial^2 f}{\partial y^2}(p),$$

so that the Hessian for  $f$  at  $p$  is

$$H = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$$

and the quadratic form for  $f$  at  $p$  is

$$Q(x, y) = \frac{1}{2}ax^2 + bxy + \frac{1}{2}cy^2.$$

Verify Theorem 5.2 in the case where  $a \neq 0$  by completing the square. (You will be using the fact that showing  $Q$  is positive definite, negative definite, indefinite, or degenerate is equivalent to showing that its Hessian has the corresponding property. Also, the case  $a = 0$  is easy to work out, but is omitted to avoid special cases.)

- (22) Find the two critical points of

$$f(x, y) = \frac{1}{3}x^3 - 3x^2 + \frac{1}{4}y^2 + xy + 13x - y + 2$$

(they have integer coordinates), and analyze them using the second derivative test (Hessian).

- (23) Let  $f(x, y) = xe^{-x^2-y^2}$ .
- (a) Use the second derivative test to analyze the critical points of  $f$ .
- (b) Does  $f$  have a global maximum? A global minimum? Justify your answer by appealing to the extreme value theorem. (Hint: What does  $f$  look like outside of a large disk centered at the origin?)

(24) Use the second derivative test to determine the behavior of the given function at the given point.

(a)  $f(x, y) = -5 - 8x - 5y + 4x^2 - 2xy - y^2 + x^2y$  at  $p = (1, -3)$ .

(b)  $f(x, y) = 4 + 5x^2 - 4y - 4x^2y + y^2 + x^2y^2$  at  $p = (0, 2)$ .

(c)  $f(x, y) = 2 - 8x + 5x^2 - x^3 + 5y - 4y^2 + y^3$  at  $p = (2, 1)$ .

(d)  $f(x, y) = -\cos(x) - 1 + \sin^2(y)$  at  $p = (0, 0)$ .



# Some Differential Geometry

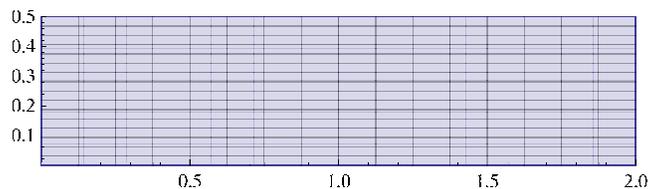
## 1. Stretching

Let  $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$  with  $n \leq m$ . Think of  $\mathbb{R}^n$  as a piece of putty and of  $f$  as a set of instructions for deforming that putty and placing it in  $\mathbb{R}^m$ .

**Example 1.1.** Define

$$\begin{aligned} f: \mathbb{R}^2 &\rightarrow \mathbb{R}^2 \\ (x, y) &\mapsto (2x, y/2) \end{aligned}$$

The image of a unit grid is pictured below.

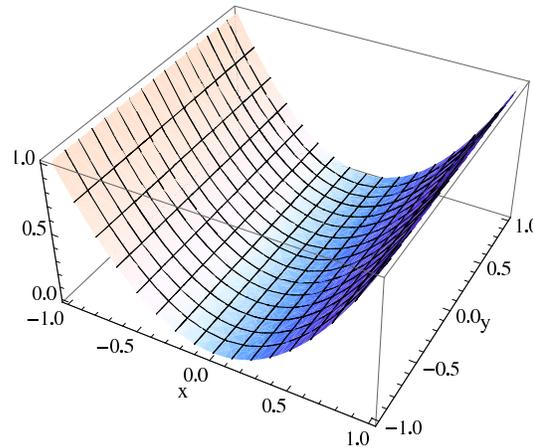


The function  $f$  stretches in one direction by a factor of 2 and in another by a factor of  $\frac{1}{2}$ . The stretching is uniform, unlike in the following examples.

**Example 1.2.** Let

$$\begin{aligned} g: \mathbb{R}^2 &\rightarrow \mathbb{R}^3 \\ (x, y) &\mapsto (x, y, x^2) \end{aligned}$$

The image of a unit grid is pictured below.

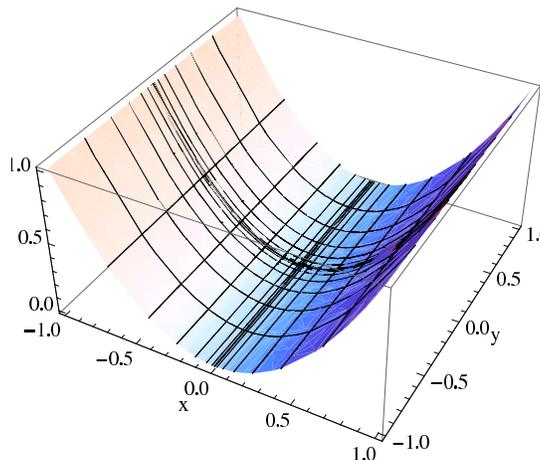


Distances along lines parallel to the  $x$ -axis in  $\mathbb{R}^2$  are unchanged, but distances along lines parallel to the  $y$ -axis are stretched.

**Example 1.3.**

$$\begin{aligned} h: \mathbb{R}^2 &\rightarrow \mathbb{R}^3 \\ (x, y) &\mapsto (x^3, y^3, x^6) \end{aligned}$$

What happens here? The image is the same as in the previous example, but the stretching has changed. The image of a unit grid:



## 2. First Fundamental Form

If we think of  $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$  as stretching its domain,  $\mathbb{R}^n$ , then the distances between points in  $\mathbb{R}^n$  are also stretched. Points that were close together may end up far apart. Further, since the stretching is not necessarily uniform, the way distances are warped can change from point to point.

In order to measure the way a function stretches its domain, we start with the definition of the tangent space of  $\mathbb{R}^n$ . Earlier, we have thought of the tangent space at a point on a parametrized surface as the set of velocity vectors of parametrized curves lying on the

surface as they pass through the point. Applying this same idea to  $\mathbb{R}^n$  itself instead of a surface in space leads to the following definition.

**Definition 2.1.** *The tangent space to  $\mathbb{R}^n$  at  $p \in \mathbb{R}^n$  is*

$$T_p\mathbb{R}^n := \{p\} \times \mathbb{R}^n = \{(p, v) \mid v \in \mathbb{R}^n\}.$$

If  $p$  is clear from context, we will usually abuse notation and write  $v$  instead of  $(p, v)$  for a point in  $T_p\mathbb{R}^n$ . Again, it may be helpful to think of  $v$  as the velocity of some curve passing through  $p$ . In the end, we have essentially attached a copy of  $\mathbb{R}^n$  to each point in  $\mathbb{R}^n$ .

Recall that earlier, in order to make measurements in  $\mathbb{R}^n$ , the key was to define an inner product. We now define an inner product in each tangent space,  $T_p\mathbb{R}^n$ , varying with the point  $p$ .

**Definition 2.2.** *Let  $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$  be a differentiable function. The first fundamental form for  $f$  at  $p \in \mathbb{R}^n$  is the function  $\langle \cdot, \cdot \rangle_p: T_p\mathbb{R}^n \times T_p\mathbb{R}^n \rightarrow \mathbb{R}$  given by*

$$\langle u, v \rangle_p := Df_p(u) \cdot Df_p(v).$$

This defines a dot product in the tangent space at  $p$  in terms of the normal dot product after applying the derivative map.

**Example 2.3.** Consider the function in example 1.2:  $g: \mathbb{R}^2 \rightarrow \mathbb{R}^3$  where  $g(x, y) = (x, y, x^2)$ . Its Jacobian matrix is

$$Jg = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 2x & 0 \end{pmatrix}.$$

Consider  $u = (1, 0)$  and  $v = (3, 2)$  as vectors in  $\mathbb{R}_p^2$ , and take their inner product. By considering the picture of  $g$  exhibited in example 1.2, how do think this inner product will change as  $p$  varies?

First, letting  $p = (0, 0)$ , we have  $Dg_p(x, y) = (x, y, 0)$ , so

$$\langle u, v \rangle_p = Dg_p(1, 0) \cdot Dg_p(3, 2) = (1, 0, 0) \cdot (3, 2, 0) = 3.$$

Thus, the inner product of  $u$  and  $v$  as elements of the tangent space  $\mathbb{R}_{(0,0)}^2$  is the same as the ordinary inner product of  $u$  and  $v$  as elements of  $\mathbb{R}^2$ . (In fact, the same could be said for any two vectors  $u$  and  $v$  at the point  $p = (0, 0)$ .) In this way, one may say that at the origin, the function  $g$  does not stretch  $\mathbb{R}^2$  at all.

We repeat the calculation of the inner product at a few other points. If  $q = (1, 2)$ , then  $Dg_q(x, y) = (x, y, 2x)$ , so

$$\langle u, v \rangle_q = Dg_q(1, 0) \cdot Dg_q(3, 2) = (1, 0, 2) \cdot (3, 2, 6) = 15.$$

If  $q' = (1, 5)$ , then  $Dg_{q'}(x, y) = (x, y, 2x)$ , so

$$\langle u, v \rangle_{q'} = Dg_{q'}(1, 0) \cdot Dg_{q'}(3, 2) = (1, 0, 2) \cdot (3, 2, 6) = 15.$$

The inner product in  $\mathbb{R}_q^2$  and in  $\mathbb{R}_{q'}^2$  have the same value. In fact, the inner product would be the same for any point of the form  $(1, t)$ , i.e., it does not depend on the second coordinate of the point. By looking at the picture in example 1.2, this result should not be surprising. On the other hand, by varying the first coordinate of the point, the inner product changes. For example, at the point  $r = (4, 2)$ , we have  $Dg_r(x, y) = (x, y, 8x)$ , so

$$\langle u, v \rangle_r = Dg_r(1, 0) \cdot Dg_r(3, 2) = (1, 0, 8) \cdot (3, 2, 24) = 192.$$

**Proposition 2.4.** (Properties of the first fundamental form.) Let  $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$  be a differentiable function, and let  $p \in \mathbb{R}^n$ . For all  $u, v, w \in T_p\mathbb{R}^n$  and  $s \in \mathbb{R}$ ,

1.  $\langle u, v \rangle_p = \langle v, u \rangle_p$ .
2.  $\langle u + v, w \rangle_p = \langle u, w \rangle_p + \langle v, w \rangle_p$ .
3.  $\langle su, v \rangle_p = s\langle u, v \rangle_p$ .
4.  $\langle u, u \rangle_p \geq 0$ , and if  $Df_p$  is 1-1, then  $\langle u, u \rangle_p = 0$  if and only if  $u = 0$ .

In sum, we say  $\langle \cdot, \cdot \rangle_p$  is a symmetric bilinear form on  $\mathbb{R}^n$ . If  $Df_p$  is 1-1, then the form is positive definite, so we can use it just like the ordinary inner product to make the following definitions.

**Definition 2.5.** With notation as in the previous proposition, assume that  $Df_p$  is 1-1. Then

1. The length of  $u \in T_p\mathbb{R}^n$  is  $|u|_p := \sqrt{\langle u, u \rangle_p}$ .
2. Vectors  $u, v \in T_p\mathbb{R}^n$  are perpendicular if  $\langle u, v \rangle_p = 0$ .
3. The angle between nonzero  $u, v \in T_p\mathbb{R}^n$  is

$$\arccos \frac{\langle u, v \rangle_p}{|u|_p |v|_p}.$$

4. The component of  $u \in T_p\mathbb{R}^n$  along nonzero  $v \in T_p\mathbb{R}^n$  is

$$\frac{\langle u, v \rangle_p}{\langle v, v \rangle_p}.$$

**Example 2.6.** Define  $f: \mathbb{R}^2 \rightarrow \mathbb{R}^3$  by  $f(x, y) = (x, y, xy)$ . At the point  $p = (1, 2)$ , we have  $Df_p(x, y) = (x, y, 2x + y)$ . Thus, the lengths of the standard basis vectors in  $\mathbb{R}_p^2$  are

$$|(1, 0)|_p = \sqrt{Df_p(1, 0) \cdot Df_p(1, 0)} = \sqrt{(1, 0, 2) \cdot (1, 0, 2)} = \sqrt{5}$$

$$|(0, 1)|_p = \sqrt{Df_p(0, 1) \cdot Df_p(0, 1)} = \sqrt{(0, 1, 1) \cdot (0, 1, 1)} = \sqrt{2}.$$

The angle between these vectors is

$$\arccos \left( \frac{(1, 0, 2) \cdot (0, 1, 1)}{\sqrt{5}\sqrt{2}} \right) = \arccos \left( \frac{2}{\sqrt{10}} \right) \approx 50.8^\circ.$$

The first fundamental form can be encoded in a matrix.

**Definition 2.7.** Let  $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$  be a differentiable function. The first fundamental form matrix is the  $n \times n$  symmetric matrix,  $\mathbb{I}$  with  $i, j$ -th entry  $f_{x_i} \cdot f_{x_j}$ ,

$$\mathbb{I} = \left( \frac{\partial f}{\partial x_i} \cdot \frac{\partial f}{\partial x_j} \right).$$

**Example 2.8.** In the case  $f: \mathbb{R}^2 \rightarrow \mathbb{R}^m$ , the first fundamental form matrix is

$$\mathbb{I} = \begin{pmatrix} E & F \\ F & G \end{pmatrix},$$

where

$$E = f_x \cdot f_x, \quad F = f_x \cdot f_y, \quad \text{and} \quad G = f_y \cdot f_y.$$

In particular, if  $f(x, y) = (x, y, xy)$ , then  $f_x = (1, 0, y)$  and  $f_y = (0, 1, x)$ , so

$$\mathbb{I} = \begin{pmatrix} 1 + y^2 & xy \\ xy & 1 + x^2 \end{pmatrix}.$$

At the point  $p = (1, 2)$ ,

$$\mathbb{I}(p) = \begin{pmatrix} 5 & 2 \\ 2 & 2 \end{pmatrix}.$$

**Proposition 2.9.** *Let  $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$  be a differentiable function, and let  $p \in \mathbb{R}^n$ . For  $u, v \in T_p\mathbb{R}^n$ ,*

$$\langle u, v \rangle_p = u^T \mathbb{I}(p) v = \begin{pmatrix} u_1 & \dots & u_n \end{pmatrix} \begin{pmatrix} f_{x_1}(p) \cdot f_{x_1}(p) & \dots & f_{x_1}(p) \cdot f_{x_n}(p) \\ \vdots & \ddots & \vdots \\ f_{x_n}(p) \cdot f_{x_1}(p) & \dots & f_{x_n}(p) \cdot f_{x_n}(p) \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}.$$

PROOF: By straightforward calculation,

$$\begin{aligned} \langle u, v \rangle_p &= Df_p(u) \cdot Df_p(v) \\ &= \left( \sum u_i f_{x_i}(p) \right) \cdot \left( \sum v_i f_{x_i}(p) \right) \\ &= \sum_{i,j} u_i v_j f_{x_i}(p) \cdot f_{x_j}(p) \\ &= \sum_i u_i \left( \sum_j f_{x_i}(p) \cdot f_{x_j}(p) v_j \right) \\ &= u^T \mathbb{I}(p) v. \end{aligned}$$

□

**Example 2.10.** Continuing examples 2.6 and 2.8, for  $f(x, y) = (x, y, xy)$  at  $p = (1, 2)$  we use the proposition to recalculate

$$\langle (1, 0), (1, 0) \rangle_p = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} 5 & 2 \\ 2 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} 5 \\ 2 \end{pmatrix} = 5.$$

Thus,  $|(1, 0)|_p = \sqrt{5}$ .

### 3. Metrics

Proposition 2.9 opens up an interesting possibility: in order to define an inner product on each tangent space of  $\mathbb{R}^n$ , we do not really need a function  $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$ . All that is needed is a matrix of a certain form.

**Definition 3.1.** *A pseudo-metric on  $\mathbb{R}^n$  is a symmetric  $n \times n$  matrix  $\mathbb{I}$  whose entries are real-valued functions on  $\mathbb{R}^n$ . Given any such  $\mathbb{I}$ , define for each  $p \in \mathbb{R}^n$  and for all  $u, v \in T_p\mathbb{R}^n$*

$$\langle u, v \rangle_p := u^T \mathbb{I}(p) v.$$

*The matrix  $\mathbb{I}$  is called a metric on  $\mathbb{R}^n$  if  $\langle \cdot, \cdot \rangle_p$  is positive definite for each  $p \in \mathbb{R}^n$ , i.e., for all  $u \in T_p\mathbb{R}^n$ ,  $\langle u, u \rangle_p \geq 0$  with equality exactly when  $u = 0$ .*

With this definition, it is easy to check that a pseudo-metric yields a symmetric, bilinear form on  $T_p\mathbb{R}^n$  that is positive definite if the pseudo-metric is a metric, (cf. the properties listed in Proposition 2.4). Thus, given a metric, we will use Definition 2.5 to define lengths, perpendicularity, angles, and components in each tangent space.

In the case  $n = 2$ , there is an easy criterion by which to judge whether a pseudo-metric is a metric.

**Proposition 3.2.** *The pseudo-metric given by the matrix  $\begin{pmatrix} E & F \\ F & G \end{pmatrix}$  with  $E, F, G$  functions of  $x, y$  is a metric if and only if*

$$E(p) > 0, \quad \text{and} \quad E(p)G(p) - F^2(p) > 0$$

for all  $p \in \mathbb{R}^2$ .

PROOF: Let  $p \in \mathbb{R}^2$ . For  $(x, y) \in \mathbb{R}_p^2$ ,

$$\langle (x, y), (x, y) \rangle_p = \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} E(p) & F(p) \\ F(p) & G(p) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = E(p)x^2 + 2F(p)xy + G(p)y^2.$$

We have a metric if and only if  $\langle x, y \rangle_p \geq 0$  with equality exactly when  $(x, y) = (0, 0)$ . For points of the form  $(x, 0)$ , where  $y = 0$ , the requirement is just that  $E(p) > 0$ . So we now assume  $E(p) > 0$  and  $y \neq 0$ . Divide the inequality  $E(p)x^2 + 2F(p)xy + G(p)y^2 \geq 0$  through by  $y^2$  to get the equivalent inequality  $E(p)z^2 + 2F(p)z + G(p) \geq 0$  where  $z = x/y$ . Considering the quadratic equation, this inequality holds for all  $z \in \mathbb{R}$  if and only if the discriminant,  $(2F(p))^2 - 4E(p)G(p) < 0$ , guaranteeing that  $E(p)z^2 + 2F(p)z + G(p)$  has no real zeros. The result follows.  $\square$

The first fundamental form of any function  $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$  gives a pseudo-metric that is a metric if and only if  $Df_p$  is 1-1 for each  $p$ . We say the pseudo-metric is *induced* by the function  $f$ . It is natural to ask if every pseudo-metric or metric is induced by some function.

**Example 3.3.** (The hyperbolic plane.) Consider the upper half-plane,  $H = \{(x, y) \in \mathbb{R}^2 \mid y > 0\}$ . In 1901, David Hilbert showed that the following metric on  $H$  is not induced by any function of  $H$  into  $\mathbb{R}^3$ :

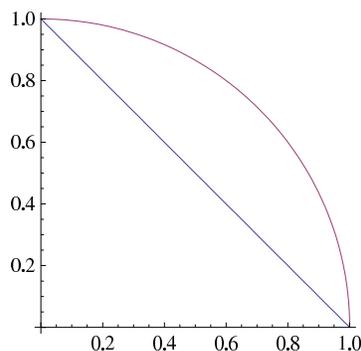
$$\begin{pmatrix} \frac{1}{y^2} & 0 \\ 0 & \frac{1}{y^2} \end{pmatrix}$$

In 1955, Blanuša showed that this metric on  $H$  is induced by a function of  $H$  into  $\mathbb{R}^6$ . I believe it has since been shown that it can be induced by a function into  $\mathbb{R}^5$ , and the question of whether  $\mathbb{R}^4$  would suffice is open.

More generally, the Nash embedding theorem says that a metric is always induced by a function into  $\mathbb{R}^n$  provided  $n$  is large enough.

#### 4. Lengths of curves

Consider  $\mathbb{R}^2$  with the metric induced by  $f(x, y) = (x, y, x^2 + y^2)$ . Which of the two paths between  $(0, 1)$  and  $(1, 0)$  pictured below is shorter, the straight line or the circle arc?



**Definition 4.1.** Let  $c: [a, b] \rightarrow \mathbb{R}^n$  be a parametrized curve. The length of  $c$  is

$$\text{length}(c) = \int_a^b |c'(t)| dt.$$

The definition says that the length of the curve is the integral of its speed.

**Example 4.2.** Parametrize a circle of radius  $r$  by  $c(t) = (r \cos(t), r \sin(t))$  for  $t \in [0, 2\pi]$ . The length of  $c$  is

$$\text{length}(c) = \int_0^{2\pi} |(-r \sin(t), r \cos(t))| dt = \int_0^{2\pi} r dt = 2\pi r.$$

If a curve is a 1 – 1 mapping, there is a good case to be made for thinking of the length of the curve as the length of its image.

**Definition 4.3.** Let  $c: [a, b] \rightarrow \mathbb{R}^n$  be a parametrized curve. Suppose  $\alpha: [a', b'] \rightarrow [a, b]$  is a function such that  $\alpha'$  is continuous,  $\alpha'(t) > 0$  for all  $t$ ,  $\alpha(a') = a$ , and  $\alpha(b') = b$ . Then the curve  $\tilde{c} = c \circ \alpha: [a', b'] \rightarrow \mathbb{R}^n$  is called a reparametrization of  $c$ .

Think of  $\alpha$  as a parametrization of  $[a, b]$  that does not change directions. (For curves  $c$  and  $\tilde{c}$ , write  $c \sim \tilde{c}$  if  $\tilde{c}$  is a reparametrization of  $c$ . Then  $\sim$  is an equivalence relation. If we like, we could define an *unparametrized* curve to be an equivalence class under this relation.)

**Proposition 4.4.** With notation as in the previous definition,

$$\text{length}(c) = \text{length}(\tilde{c}).$$

PROOF: Exercise. □

Let  $u: [a, b] \rightarrow \mathbb{R}^n$  be a parametrized curve in  $\mathbb{R}^n$ , and  $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$  any mapping. Letting  $c = f \circ u$  gives a parametrized curve lying on the image of  $f$ . The length of  $c$  can be calculated using the chain rule

$$\begin{aligned} \text{length}(c) &= \int_a^b |c'(t)| dt = \int |(f \circ u)'(t)| dt \\ &= \int_a^b |Df_{u(t)}(u'(t))| dt \\ &= \int_a^b \sqrt{Df_{u(t)}(u'(t)) \cdot Df_{u(t)}(u'(t))} dt \\ &= \int_a^b \sqrt{\langle u'(t), u'(t) \rangle_{u(t)}} dt \\ &= \int_a^b |u'(t)|_{u(t)} dt. \end{aligned}$$

The calculation motivates the following definition.

**Definition 4.5.** Let  $u: [a, b] \rightarrow \mathbb{R}^n$  be a parametrized curve, and let  $\mathbb{I}$  be a pseudo-metric on  $\mathbb{R}^n$ . The length of  $u$  with respect to  $\mathbb{I}$  is

$$\text{length}(u) = \int_a^b |u'(t)|_{u(t)} dt,$$

where the length of the vector  $u'(t)$  is its length with respect to  $\mathbb{I}$  as an element of the tangent space  $T_{u(t)}\mathbb{R}^n$ .

**Example 4.6.** We answer the question with which we began this section. Let  $f(x, y) = (x, y, x^2 + y^2)$ .

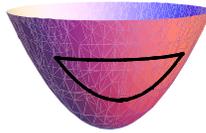
- (1) First consider the straight-line path from  $(0, 1)$  to  $(1, 0)$ . Parametrize the line by  $u(t) = (t, 1-t)$  with  $t \in [0, 1]$ . It follows that  $c(t) := f(u(t)) = (t, 1-t, 2t^2 - 2t + 1)$ , and  $c'(t) = (1, -1, 4t - 2)$ . Thus,

$$\text{length}(c) = \int_0^1 |c'(t)| dt = \int_0^1 \sqrt{16t^2 - 16t + 6} dt \approx 1.80.$$

- (2) Next, consider the circular arc,  $u(t) = (\sin t, \cos t)$  for  $t \in [0, \pi/2]$ . The corresponding curve on the paraboloid is  $c(t) = f(u(t)) = (\sin t, \cos t, 1)$ . Thus,  $c'(t) = (\cos t, -\sin t, 0)$ , and

$$\text{length}(c) = \int_0^{\pi/2} |c'(t)| dt = \int_0^{\pi/2} 1 dt = \frac{\pi}{2} \approx 1.57.$$

Thus, the circular arc gives a shorter path than the straight line with respect to the metric induced by  $f$ .



## 5. Geodesics

Given two points in  $\mathbb{R}^n$  with a given metric, what is the shortest path between the points? We start our discussion of this problem with the case of a metric induced by a function.

**Definition 5.1.** Let  $u: [a, b] \rightarrow \mathbb{R}^n$  and  $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$ , so  $c = f \circ u$  is a parametrized curve on the image of  $f$ . Then  $c$  is a geodesic if

$$c''(t) \cdot f_{x_i}(u(t)) = 0$$

for  $i = 1, \dots, n$ .

Roughly,  $c$  is a geodesic if at each point its acceleration has no component lying in the tangent space of  $f$ . Imagine what it would feel like to ride a roller coaster along a geodesic: you would only feel acceleration straight up or straight down in your seat, not to the side.

**Example 5.2.** With notation as above, suppose  $c$  is a curve whose components are linear functions. Then  $c''(t) = 0$ ; hence,  $c$  is a geodesic. The image of  $c$  is a line. Compare this with the following situation:  $f(x, y) = (x, y)$  and  $c(t) = u(t) = (t^2, 0)$  (a curve in the plane with the ordinary metric). In this case,

$$c''(t) \cdot f_x(u(t)) = (2, 0) \cdot (1, 0) = 2 \neq 0.$$

Hence, even though the image of  $c$  is a line, it is not a geodesic. The curve  $c$  is an example of what is called a *pre-geodesic*. It fails to be a geodesic because it has a component along its direction of motion  $c'(t)$ . For any such curve, there is a reparametrization that is a geodesic, as is clear in this example. Thinking of roller coasters again, riding a pre-geodesic, you might

feel acceleration straight up or down but also directly forward or backward. However, you would not feel acceleration to the sides.

**Example 5.3.** Consider a curve in the plane with the ordinary metric. Take  $f(x, y) = (x, y)$  and  $u(t) = (u_1(t), u_2(t))$  and  $c = f \circ u = u$ . When is  $c$  a geodesic? Calculating, we need

$$\begin{aligned} c''(t) \cdot f_x(u(t)) &= (u_1''(t), u_2''(t)) \cdot (1, 0) = u_1''(t) = 0 \\ c''(t) \cdot f_y(u(t)) &= (u_1''(t), u_2''(t)) \cdot (0, 1) = u_2''(t) = 0. \end{aligned}$$

This implies both  $u_1$  and  $u_2$  are linear functions. So the only geodesics in the plane endowed with the usual metric are lines. Similarly, in  $\mathbb{R}^n$  with the usual metric, curves whose components are linear functions are geodesics.

**Example 5.4.** Suppose that  $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$ ,  $u: [a, b] \rightarrow \mathbb{R}^n$ , and  $c = f \circ u$ . If  $c$  has linear components, then  $c''(t) = 0$ , and hence  $c$  is a geodesic. Thus, every line on the image of  $f$  (appropriately parametrized) is a geodesic.

**Example 5.5.** Let  $f(x, y) = (x, y, x^2 + y^2)$  and  $u(t) = (\sin t, \cos t)$  for  $t \in [0, \pi/2]$ . We saw earlier that  $u$  is a shorter path from  $(0, 1)$  to  $(1, 0)$  than the straight line path in  $\mathbb{R}^2$  with the metric coming from  $f$ . Is  $c = f \circ u$  a geodesic? We have  $c(t) = (\sin t, \cos t, 1)$ ,  $f_x = (1, 0, 2x)$ , and  $f_y = (0, 1, 2y)$ . Hence,

$$\begin{aligned} c''(t) \cdot f_x(u(t)) &= (-\sin t, -\cos t, 0) \cdot (1, 0, 2 \sin t) = -\sin t \\ c''(t) \cdot f_y(u(t)) &= (-\sin t, -\cos t, 0) \cdot (0, 1, 2 \cos t) = -\cos t. \end{aligned}$$

Thus,  $c$  is not a geodesic since otherwise both of these dot products would have to be 0 for all  $t$ .

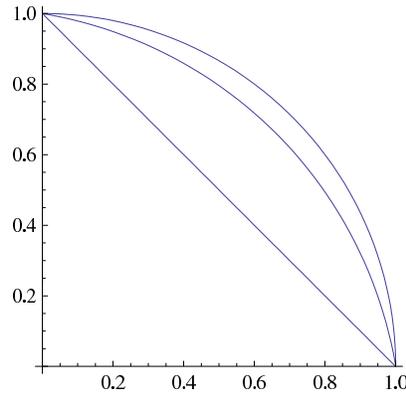
How would one calculate a geodesic on this surface? Let  $u(t) = (x(t), y(t))$  be an arbitrary curve, and consider  $c(t) = (x(t), y(t), x^2(t) + y^2(t))$ . Dropping the argument  $t$ , we have

$$c''(t) = (x'', y'', 2xx'' + 2(x')^2 + 2yy'' + 2(y')^2).$$

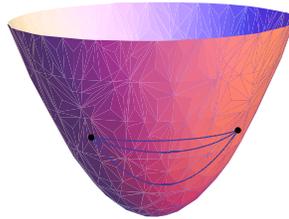
Thus,  $c$  is a geodesic if and only if  $x(t)$  and  $y(t)$  satisfy the following system of differential equations:

$$\begin{aligned} c''(t) \cdot f_x(u(t)) &= x'' + 2x(2xx'' + 2(x')^2 + 2yy'' + 2(y')^2) = 0 \\ c''(t) \cdot f_y(u(t)) &= y'' + 2y(2xx'' + 2(x')^2 + 2yy'' + 2(y')^2) = 0. \end{aligned}$$

Playing with a numerical differential equations solver, one finds that the geodesic with the initial conditions  $u(0) = (0, 1)$  and  $u'(0) = (1, -0.17)$  will come close to traveling from  $(0, 1)$  for  $(1, 0)$  as  $t$  goes from 0 to about 1.43. The length of this curve is approximately 1.53, which is shorter than the straight-line path and circular arc considered earlier. Here are pictures of all three paths in the plane:



and on the paraboloid:



The geodesic lies between the straight-line path and the circular arc.

### Remarks.

- It turns out that if a curve gives the shortest path between two points, then it is a geodesic. The converse is not true however. For example, geodesics on a sphere turn out to be great circles, like the equator. Given two points on the equator, there are two directions one may follow the equator in order to travel from one point to the other, and usually one of these will be shorter than the other.
- In this section and in the next geodesics are defined via second-order differential equations. Existence and uniqueness theorems from the theory of differential equations then come into play to show that given an initial point and velocity, there is a unique geodesic starting at that point with the given velocity.

**5.1. The Notion of a Geodesic is Intrinsic.** We have seen that although a function  $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$  can be used to induce a metric on  $\mathbb{R}^n$ , all information about the metric is contained in the first fundamental form matrix. So in general we defined a metric on  $\mathbb{R}^n$  to just be an  $n \times n$  symmetric matrix whose coefficients are real-valued functions on  $\mathbb{R}^n$ . The purpose of this section is to show that geodesics can be defined “intrinsically”, that is, in terms of the metric, without appeal to an external function  $f$ .

Fix a metric  $\mathbb{I}$ , and denote its  $ij$ -th entry by  $g_{ij}$ . Since  $\mathbb{I}$  is positive definite, it turns out that it must be an invertible matrix, and we denote the entries of the inverse by  $g^{ij}$ :

$$\mathbb{I} = (g_{ij}), \quad \mathbb{I}^{-1} = (g^{ij}).$$

For each  $i, j, \ell \in \{1, \dots, n\}$  define a *Christoffel symbol*

$$\Gamma_{ij}^{\ell} = \frac{1}{2} \sum_{k=1}^n \left( \frac{\partial g_{jk}}{\partial x_i} - \frac{\partial g_{ij}}{\partial x_k} + \frac{\partial g_{ki}}{\partial x_j} \right) g^{k\ell}.$$

The three terms in parentheses can be remembered by “cycling indices” and alternating signs. Note that  $\Gamma_{ij}^\ell = \Gamma_{ji}^\ell$  since both  $\mathbb{I}$  and  $\mathbb{I}^{-1}$  are symmetric.

**Example 5.6.** In the case  $n = 2$  we can write

$$\mathbb{I} = \begin{pmatrix} E & F \\ F & G \end{pmatrix}, \quad \mathbb{I}^{-1} = \frac{1}{\Delta} \begin{pmatrix} G & -F \\ -F & E \end{pmatrix}$$

where  $\Delta = \det \mathbb{I} = EG - F^2$ . The reader should verify the following:

$$\begin{aligned} \Gamma_{11}^1 &= \frac{1}{2\Delta}(GE_x - 2FF_x + FE_y) & \Gamma_{11}^2 &= \frac{1}{2\Delta}(2EF_x - EE_y - FE_x) \\ \Gamma_{12}^1 &= \frac{1}{2\Delta}(GE_y - FG_x) & \Gamma_{12}^2 &= \frac{1}{2\Delta}(EG_x - FE_y) \\ \Gamma_{22}^1 &= \frac{1}{2\Delta}(2GF_y - GG_x - FG_y) & \Gamma_{22}^2 &= \frac{1}{2\Delta}(EG_y - 2FF_y + FG_x). \end{aligned}$$

As an attempt to explain the geometrical significance of the Christoffel symbols, let  $f: \mathbb{R}^n \rightarrow \mathbb{R}^{n+1}$  and suppose that the image of  $Df_p$  has dimension  $n$  for each  $p \in \mathbb{R}^n$ . We call such an  $f$  a regular hypersurface in  $\mathbb{R}^{n+1}$ . Define the *normal vector*,  $n$ , to the image of  $f$  at each point using the natural generalization of the cross product in 3 dimensions by formally taking the determinant of the matrix whose first row consists of the standard basis vectors for  $\mathbb{R}^{n+1}$  and whose remaining rows are the first partials of  $f$ . It turns out  $n(p)$  is perpendicular to each  $f_{x_i}(p)$ .

At each point  $p$ , we can then form a basis for  $\mathbb{R}^{n+1}$  consisting of the normal vector,  $n(p)$ , and a basis for the tangent space,  $f_{x_1}(p), \dots, f_{x_n}(p)$ . The second partials of  $f$  should measure of how the surface is moving away from its tangent space. It turns out that the Christoffel symbols encode the tangential components of these second partials. We will just state the theorem without proof.

**Theorem 5.7.** *With notation as above, for each  $i, j \in \{1, \dots, n\}$ , and for each  $p \in \mathbb{R}^n$ ,*

$$f_{x_i x_j}(p) = r_{ij}(p) n(p) + \sum_k \Gamma_{ij}^k(p) f_{x_k}(p),$$

for some  $r_{ij}$  is a real-valued function on  $\mathbb{R}^n$ .

So we may consider the Christoffel symbols as given “tangential coordinates” of the second partials of  $f$ .

We now give the intrinsic definition of a geodesic. We use the standard dot notation for derivatives with respect to  $t$ : one dot signifies to a single derivative and two dots signifies to a second derivative.

**Definition 5.8.** *Let  $\mathbb{R}^n$  be endowed with a metric having Christoffel symbols  $\Gamma_{ij}^k$ , and let  $x(t) = (x_1(t), \dots, x_n(t))$  be a curve in  $\mathbb{R}^n$ . Then  $x(t)$  is a geodesic if it satisfies the system of differential equations*

$$\ddot{x}_k + \sum_{i,j} \Gamma_{i,j}^k \dot{x}_i \dot{x}_j = 0, \quad k = 1, \dots, n.$$

For  $i \neq j$ , both  $\Gamma_{i,j}^k$  and  $\Gamma_{j,i}^k$  appear separately in the above sum. Since,  $\Gamma_{i,j}^k = \Gamma_{j,i}^k$  these can be combined to give  $2\Gamma_{i,j}^k$ .

For instance, in the case  $n = 2$  the condition is that

$$\begin{aligned} \ddot{x}_1 + \dot{x}_1^2 \Gamma_{11}^1 + 2\dot{x}_1 \dot{x}_2 \Gamma_{12}^1 + \dot{x}_2^2 \Gamma_{22}^1 &= 0 \\ \ddot{x}_2 + \dot{x}_1^2 \Gamma_{11}^2 + 2\dot{x}_1 \dot{x}_2 \Gamma_{12}^2 + \dot{x}_2^2 \Gamma_{22}^2 &= 0 \end{aligned}$$

Admittedly, this is a grungy definition. In a proper course in differential geometry, one develops the notion of a “covariant derivative” and gives a cleaner, more conceptual definition of a geodesic. Nevertheless, armed with a numerical differential equation solver, we can use this definition to make calculations. The following theorem states that our new notion of a geodesic generalizes the old one.

**Theorem 5.9.** *Let  $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$  be a smooth mapping (i.e., all its partials exist of all orders), and assume its derivative is 1-1 at all points. Let  $u: I \rightarrow \mathbb{R}^n$  be a smooth curve on some interval  $I$ , and let  $c = f \circ u$ . Consider  $\mathbb{R}^n$  with the metric induced by  $f$ . Then  $u$  is a geodesic under the definition just given if and only if*

$$c''(t) \cdot f_{x_i}(u(t)) = 0, \quad i = 1, \dots, n.$$

PROOF: We only provide a proof in the case  $f: \mathbb{R}^2 \rightarrow \mathbb{R}^3$ .

Since  $c = f \circ u$ , the chain rule says that  $c' = u'_1 f_x + u'_2 f_y$ . Use the chain rule again to find the acceleration of  $c$  and use the Theorem 5.7 to get

$$\begin{aligned} c'' &= u''_1 f_x + u''_2 f_y + (u'_1)^2 f_{xx} + 2u'_1 u'_2 f_{xy} + (u'_2)^2 f_{yy} \\ &= u''_1 f_x + u''_2 f_y + (u'_1)^2 (\Gamma_{11}^1 f_x + \Gamma_{11}^2 f_y + r_{11} n) + \\ &\quad 2u'_1 u'_2 (\Gamma_{12}^1 f_x + \Gamma_{12}^2 f_y + r_{12} n) + (u'_2)^2 (\Gamma_{22}^1 f_x + \Gamma_{22}^2 f_y + r_{22} n) \end{aligned}$$

Take the dot product of  $c''$  with  $f_x$  and  $f_y$ , remembering that each of these latter vectors is perpendicular to the normal and that  $E = f_x \cdot f_y$ , etc., to get

$$c'' \cdot f_x = u''_1 E + u''_2 F + (u'_1)^2 (\Gamma_{11}^1 E + \Gamma_{11}^2 F) + 2u'_1 u'_2 (\Gamma_{12}^1 E + \Gamma_{12}^2 F) + (u'_2)^2 (\Gamma_{22}^1 E + \Gamma_{22}^2 F)$$

$$c'' \cdot f_y = u''_1 F + u''_2 G + (u'_1)^2 (\Gamma_{11}^1 F + \Gamma_{11}^2 G) + 2u'_1 u'_2 (\Gamma_{12}^1 F + \Gamma_{12}^2 G) + (u'_2)^2 (\Gamma_{22}^1 F + \Gamma_{22}^2 G)$$

Factoring out  $E$ 's,  $F$ 's, and  $G$ 's gives

$$\begin{aligned} &(u''_1 + (u'_1)^2 \Gamma_{11}^1 + 2u'_1 u'_2 \Gamma_{12}^1 + (u'_2)^2 \Gamma_{22}^1) E + (u''_2 + (u'_1)^2 \Gamma_{11}^2 + 2u'_1 u'_2 \Gamma_{12}^2 + (u'_2)^2 \Gamma_{22}^2) F \\ &(u''_1 + (u'_1)^2 \Gamma_{11}^1 + 2u'_1 u'_2 \Gamma_{12}^1 + (u'_2)^2 \Gamma_{22}^1) F + (u''_2 + (u'_1)^2 \Gamma_{11}^2 + 2u'_1 u'_2 \Gamma_{12}^2 + (u'_2)^2 \Gamma_{22}^2) G \end{aligned}$$

Defining

$$\begin{aligned} \alpha &= u''_1 + (u'_1)^2 \Gamma_{11}^1 + 2u'_1 u'_2 \Gamma_{12}^1 + (u'_2)^2 \Gamma_{22}^1 \\ \beta &= u''_2 + (u'_1)^2 \Gamma_{11}^2 + 2u'_1 u'_2 \Gamma_{12}^2 + (u'_2)^2 \Gamma_{22}^2, \end{aligned}$$

the condition that  $c'' \cdot f_x = c'' \cdot f_y = 0$  is equivalent to

$$\alpha E + \beta F = 0$$

$$\alpha F + \beta G = 0$$

Multiply the first equation by  $G$ , the second by  $-F$ , and add:

$$\alpha(EG - F^2) = 0.$$

Multiply the first equation by  $-F$  and the second by  $E$  and add:

$$\beta(EG - F^2) = 0.$$

Since  $I$  is positive definite,  $EG - F^2$  is never zero. Hence,  $c'' \cdot f_x = c'' \cdot f_y = 0$  is equivalent to  $\alpha = \beta = 0$ , as required.  $\square$

**Example 5.10.** (Hyperbolic plane.) Let  $H = \{(x, y) : y > 0\}$  be the upper half-plane endowed with the metric

$$\mathbb{I} = \begin{pmatrix} \frac{1}{y^2} & 0 \\ 0 & \frac{1}{y^2} \end{pmatrix}.$$

We will show that the geodesics are either vertical lines or parts of circles centered on the  $x$ -axis.

The equations (Definition 5.8) for a geodesic  $\gamma(t) = (x(t), y(t))$  are (cf. Exercise 6)

$$\ddot{x} - 2\frac{\dot{x}\dot{y}}{y} = 0, \quad \ddot{y} + \frac{\dot{x}^2 - \dot{y}^2}{y} = 0. \quad (9)$$

Using these equations, one may compute that

$$\frac{d}{dt}|\dot{\gamma}|^2 = \frac{d}{dt}\langle \dot{\gamma}, \dot{\gamma} \rangle = \frac{d}{dt} \left( \frac{\dot{x}^2 + \dot{y}^2}{y^2} \right) = 0,$$

(the first equalities are definitions, using our metric). Thus, any geodesic moves with constant speed. (In fact, this is a general result, not dependent on the particular metric we are considering in this example. This makes sense since one does not expect the speed to change if the acceleration is perpendicular to the velocity.) We will assume this speed is positive, ruling out the case of a trivial curve that just sits at a point.

Substituting  $z = \dot{x}$  in the first equation, separating variables, and integrating shows that

$$\dot{x} = ay^2 \quad (10)$$

for some constant  $a$ . If  $a = 0$ , the second of the geodesic equations (9) implies that  $(\dot{y}/y)' = 0$ , from which it easily follows that  $y = ce^t$  for some constant  $c > 0$ . Thus, the geodesic is  $t \mapsto (0, ce^t)$ , parametrizing a vertical line. We will now assume  $a > 0$ , the negative case being similar. From above, we have that  $\dot{x}^2 + \dot{y}^2 = by^2$  for some constant  $b > 0$ . Dividing this through by  $\dot{x}$  and using equation 10 yields

$$\left( \frac{dy}{dx} \right)^2 = \left( \frac{\dot{y}}{\dot{x}} \right)^2 = b/(a^2y^2) - 1.$$

The first step is justified by the chain rule and inverse function theorem from one-variable calculus. Taking the positive square root, the negative case being similar, gives

$$\frac{dy}{dx} = \frac{\sqrt{b - a^2y^2}}{ay}.$$

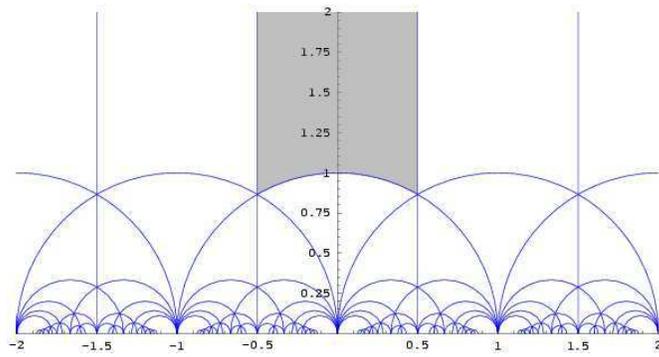
Separating variables and performing a simple integration yields

$$(x - c)^2 + y^2 = b/a^2$$

for some constant  $c$ . Thus, the geodesic is part of a circle centered at a point on the  $x$ -axis.

Given any initial point  $p$  and direction  $\vec{v}$  in  $H$ , we have shown that the geodesic starting at that point—which exists and is determined by a basic existence and uniqueness theorem from the theory of differential equations—is either a vertical line or a piece of a circular arc. If it is a circle, it is the unique circle with center on the  $x$ -axis, passing through  $p$ , and such that the line from the center out to  $p$  is perpendicular to  $\vec{v}$ .

A drawing with several geodesics grabbed from Wikipedia:



## EXERCISES

- (1) **Reparametrizations.** To a large extent, our definition of the length of a curve really describes the length of the image of the curve.
- (a) For each positive integer  $n$ , define the curve  $c_n(t) = t^n$  for  $t \in [-1, 1]$ . Two questions: (i) What is the image  $c_n$  for each  $n$ ? (ii) What is the length of  $c_n$  for each  $n$ ?
- (b) Now consider the three curves

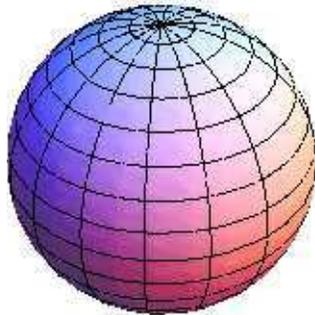
$$\begin{aligned} h: [0, \pi] &\rightarrow \mathbb{R} \\ j: [-\pi/2, \pi/2] &\rightarrow \mathbb{R} \\ k: [0, 2\pi] &\rightarrow \mathbb{R} \end{aligned}$$

all given by  $\cos t$ . In other words, consider the function  $\cos t$  restricted to three different domains. Same two questions: (i) What is the image of each of these functions? (ii) What is the length of each of these functions? (They are not all the same.)

- (c) An easier example of how the parametrization might not really reflect the length of its image, consider the curve  $\gamma(t) = (\cos t, \sin t)$ . What is the length of  $\gamma$  on the interval  $[0, 2\pi n]$  for each positive integer  $n$ ? Note that in each case, the image is just the unit circle.
- (d) Prove Proposition 4.4. The result should follow by a simple one-variable calculus  $u$ -substitution.
- (2) Consider the spherical coordinate parametrization of the unit sphere,

$$\phi(u, v) = (\sin u \cos v, \sin u \sin v, \cos u)$$

for  $u \in [0, \pi]$  and  $v \in [0, 2\pi]$ . The angle  $u$  controls the latitude, starting at the north pole, and  $v$  controls the longitude. The map  $\phi$  induces an inner product in the tangent space at each point  $(u, v)$ .

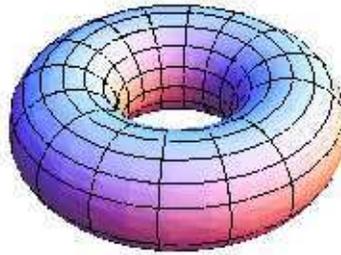


- (a) Compute the first fundamental form matrix for  $\phi$ . (The identity  $\cos^2 \theta + \sin^2 \theta = 1$  will be useful.)
- (b) Find the lengths of  $(1, 0)$  and  $(0, 1)$  as elements of the tangent space at an arbitrary point  $(u, v)$ . Explain the varying of these lengths by appealing to the picture above.
- (c) What is the angle between  $(1, 0)$  and  $(0, 1)$  as elements in the tangent space at an arbitrary point  $(u, v)$ ?
- (d) Show that the meridians  $t \mapsto (t, c)$ , where  $c$  is a constant are geodesics in  $[0, \pi] \times [0, 2\pi]$  with the metric induced by  $\phi$ .

- (e) Show that the latitudes  $t \mapsto (c, t)$  where  $c$  is a constant are not generally geodesics. When are they?
- (3) Consider the parametrization of a torus

$$\tau(u, v) = ((2 + \cos v) \cos u, (2 + \cos v) \sin u, \sin v).$$

The distance from the origin to the center of the tube of the torus is 2 and radius of the tube is 1. Let the coordinates on  $\mathbb{R}^3$  be  $(x, y, z)$ . If  $p = \tau(u, v)$ , then  $u$  is the angle between the  $x$ -axis and the line connecting the origin to the projection of  $p$  onto the  $xy$ -plane. A slice of the torus through the origin and  $p$  by a plane perpendicular to the  $xy$ -plane cuts out two circles on the torus, one of which contains  $p$ . The angle  $v$  is the angle from the center of the circle containing  $p$  out to  $p$ .



- (a) Compute the first fundamental form matrix for  $\tau$ .
- (b) Find the lengths of  $(1, 0)$  and  $(0, 1)$  as elements of the tangent space at an arbitrary point  $(u, v)$ . Explain the varying of these lengths by appealing to the picture above.
- (c) What is the angle between  $(1, 0)$  and  $(0, 1)$  as elements in the tangent space at an arbitrary point  $(u, v)$ ?
- (d) Show that the circles  $t \mapsto (c, t)$ , where  $c$  is a constant are geodesics in  $[0, 2\pi] \times [0, 2\pi]$  with the metric induced by  $\tau$ .
- (e) Show that the circles  $t \mapsto (t, c)$  where  $c$  is a constant are not generally geodesics. When are they?
- (4) **Clairaut's relation.** (This exercise is based on an exercise from Do Carmo's book on Riemannian geometry.) Let  $\gamma(t) = (x(t), y(t))$  be a smooth curve in  $\mathbb{R}^2$  with domain the interval  $I$ . Assume that  $\gamma'(t) \neq 0$  and  $x(t) \neq 0$  for  $t \in I$ . Define

$$\begin{aligned} \varphi: [0, 2\pi] \times I &\rightarrow \mathbb{R}^3 \\ (\theta, t) &\mapsto (x(t) \cos \theta, x(t) \sin \theta, y(t)). \end{aligned}$$

The image of  $\varphi$  is the surface of revolution formed by rotating the curve  $t \mapsto (x(t), 0, y(t))$  about the  $z$ -axis. The curves  $\theta = \text{constant}$  and  $t = \text{constant}$  are called *meridians* and *parallels*, respectively.

- (a) Show that the first fundamental form matrix for  $\varphi$  is

$$\mathbb{I}(\theta, t) = \begin{pmatrix} x(t)^2 & 0 \\ 0 & \dot{x}^2(t) + \dot{y}^2(t) \end{pmatrix}$$

- (b) Verify  $\mathbb{I}$  defines a positive definite inner product (cf. Proposition 3.2).

- (c) Show that  $c(t) = (\theta(t), s(t))$  is a geodesic (with respect to the metric induced by  $\varphi$ ) if and only if the following two equations are satisfied:

$$\begin{aligned}\ddot{\theta} + 2\frac{\dot{x}}{x}\dot{\theta}\dot{s} &= 0, \\ \ddot{s} - \frac{x\dot{x}}{\dot{x}^2 + \dot{y}^2}\dot{\theta}^2 + \frac{\dot{x}\ddot{x} + \dot{y}\ddot{y}}{\dot{x}^2 + \dot{y}^2}s^2 &= 0.\end{aligned}$$

- (d) Let  $\phi \circ c$  be a geodesic on the surface of revolution; let  $\beta(t)$  be the angle between  $\phi \circ c$  and the parallel of the surface of revolution intersecting  $c$  at time  $t$ ; and let  $r(t)$  be the radius of this parallel, i.e., the distance from  $\phi(c(t))$  to the  $z$ -axis. Show that the first of the geodesic equations from above implies *Clairaut's relation*:

$$r(t) \cos \beta(t) = \text{constant}.$$

- (e) Continuing the notation from above, show that the second of the two geodesic equations means that the *energy*,  $|\dot{c}(t)|^2$ , of the geodesic is constant along  $c$ .  
 (f) Use Clairaut's relation to show that a geodesic of the paraboloid determined by  $\gamma(t) = (t, t^2)$  which is not a meridian intersects itself an infinite number of times.

- (5) Consider  $\mathbb{R}^2$  with metric induced by the matrix

$$\mathbb{I} = \begin{pmatrix} E & F \\ F & G \end{pmatrix}$$

where  $E$ ,  $F$ , and  $G$  are real-valued functions on  $\mathbb{R}^2$ . Give a necessary and sufficient condition on these functions in order for the standard basis vectors  $(1, 0)$  and  $(0, 1)$  in each tangent space to be perpendicular.

- (6) Let  $\gamma(t) = (x(t), y(t))$  be a curve in the upper half-plane with the hyperbolic metric (cf. Example 5.10).  
 (a) Show that  $\gamma(t)$  is a geodesic if and only if the following system of differential equations is satisfied:

$$\ddot{x} - 2\frac{\dot{x}\dot{y}}{y} = 0, \quad \ddot{y} + \frac{\dot{x}^2 - \dot{y}^2}{y} = 0.$$

- (b) Show that if  $\gamma$  is a geodesic then

$$\frac{d}{dt}|\dot{\gamma}|^2 = 0,$$

as claimed in the text.

- (7) Create your own metric on some region of  $\mathbb{R}^2$  and investigate some of its properties.  
 (a) Create a  $2 \times 2$  symmetric matrix,  $\mathbb{I}$ , with real-valued functions of two variables as entries.  
 (b) Use Proposition 3.2 to identify a region  $U \subset \mathbb{R}^2$  in which your metric is positive definite.  
 (c) Compute the lengths of  $(1, 0), (0, 1) \in T_p\mathbb{R}^2$  and the angles between these vectors for various points  $p$  in your region. Can you describe what is going on generally? (It might help to use a computer algebra system to plot some of these values.)  
 (d) Find the system of differential equations characterizing a geodesic with respect to  $\mathbb{I}$ .

- (e) Create a parametrized curve  $c : [0, 1] \rightarrow \mathbb{R}^2$  whose image is contained in your chosen region  $U$ . Compute the length of  $c$  with respect to the usual metric on  $\mathbb{R}^2$  and with respect to  $\mathbb{I}$ .
- (f) Use your favorite numerical integrator to compute and plot some geodesics.

## Set notation

A set is a collection of objects. If  $A$  is a set, we write  $x \in A$  if the object  $x$  is in  $A$ , or sometimes  $A \ni x$ ; if  $x$  is not in  $A$ , we write  $x \notin A$ . For example, if  $A = \{\clubsuit, \diamond, \heartsuit\}$ , then  $\diamond \in A$  and  $\spadesuit \notin A$ . The *empty set*, denoted  $\emptyset$ , is the set which contains no elements. We write  $\emptyset = \{\}$ .

If  $A$  and  $B$  are both sets, we write  $A \subseteq B$  or  $B \supseteq A$  if every element of  $A$  is an element of  $B$ . In this case, we say that  $A$  is a *subset* of  $B$  and  $B$  is a *superset* of  $A$ . The sets  $A$  and  $B$  are *equal*,  $A = B$ , if the objects in  $A$  are the same as the objects in  $B$ . In practice, it is often best to show  $A = B$  by separately demonstrating  $A \subseteq B$  and  $B \subseteq A$ . We will reserve the notation  $A \subset B$  and  $B \supset A$  to mean that  $A$  is a subset of  $B$  and  $A \neq B$ , mimicking the distinction between “less than or equal”,  $\leq$ , and “strictly less than”,  $<$ ; however, there is no agreement on this distinction in mathematical writing. The empty set is a subset of every set (otherwise it would have to contain at least one element: one not contained in another set).

An important way to describe a set, used throughout these notes, is the *set-builder notation*. For example, we can specify the set of all real numbers greater than 2 and less than 5 by

$$A = \{x \in \mathbb{R} \mid x > 2 \text{ and } x < 10\}.$$

The set-builder notation always takes the form  $\{\mathcal{O} \mid \mathcal{C}\}$ , meaning “the set of all objects of the form  $\mathcal{O}$  which satisfy the conditions specified by  $\mathcal{C}$ .” The central bar, “|”, can usually be translated as “such that.” Thus, the set  $A$  displayed above is described as “the set of all  $x \in \mathbb{R}$  such that  $x > 2$  and  $x < 10$ .”

Given two sets  $A$  and  $B$ , their *union* is the set of all elements that are in either  $A$  or  $B$ , denoted  $A \cup B$ . Their *intersection* is the set of all elements that are in both  $A$  and  $B$ , denoted  $A \cap B$ . Using set-builder notation,

$$\begin{aligned} A \cup B &:= \{x \mid x \in A \text{ or } x \in B\}, \\ A \cap B &:= \{x \mid x \in A \text{ and } x \in B\}. \end{aligned}$$

**[Attention:** Note the use of the “:=” symbol. We write  $E := F$  if  $E$  is *defined to be*  $F$ .] To demonstrate the notions of union and intersection, let  $A = \{1, 2, 3, 4, 5\}$  and  $B = \{4, 5, 6, 7\}$ ; then  $A \cup B = \{1, 2, 3, 4, 5, 6, 7\}$  and  $A \cap B = \{4, 5\}$ .

Let  $I$  be some indexing set and suppose we are given sets  $A_i$  for each  $i \in I$ . Then

$$\cup_{i \in I} A_i := \{x \mid x \in A_i \text{ for at least one } i \in I\},$$

$$\cap_{i \in I} A_i := \{x \mid x \in A_i \text{ for all } i \in I\}.$$

If  $I = \{1, 2, 3, \dots\}$  we could alternatively write  $\cup_{i=1}^{\infty} A_i$  or  $\cap_{i=1}^{\infty} A_i$ ; or if  $I = \{1, 2, 3\}$ , we could write  $A_1 \cup A_2 \cup A_3$ , to denote the elements that are in at least one of  $A_1$ ,  $A_2$ , or  $A_3$ , etc. For example, if  $A_i = \{1, 2, \dots, i\}$  for  $i = 1, 2, \dots$ , then  $\cup_{i=1}^{\infty} A_i = \{1, 2, 3, \dots\}$  and  $\cap_{i=1}^{\infty} A_i = \{1\}$ .

If  $A$  and  $B$  are two sets, we define their *set difference* to be

$$A \setminus B := \{x \mid x \in A \text{ and } x \notin B\},$$

i.e., the set of all elements in  $A$  but not in  $B$ . This is also called the *complement of  $B$  in  $A$* . In a given context, all sets under consideration may be assumed to be subsets of some “universal set”, say  $U$ . In that case, we simply talk about the *complement of  $A$* , often denoted  $A^c$ :

$$A^c := U \setminus A.$$

For instance, suppose we are discussing subsets of the real numbers. Let  $A$  denote all real numbers greater than 10. Then  $A^c$  is the set of all real numbers less than or equal to 10.

## EXERCISES

Let  $A$  and  $B$  be sets. Also, let  $A_i$  be a set for each  $i$  in some indexing set  $I$ . Try proving various versions of De Morgan's laws:

$$(1) (A \cup B)^c = A^c \cap B^c.$$

$$(2) (A \cap B)^c = A^c \cup B^c.$$

$$(3) (\cup_{i \in I} A_i)^c = \cap_{i \in I} (A_i^c).$$

$$(4) (\cap_{i \in I} A_i)^c = \cup_{i \in I} (A_i^c).$$



# Real numbers

The real numbers,  $\mathbb{R}$ , form an ordered field satisfying the least upper bound property.

## 1. Field axioms

A *field* is a set  $F$  with two operations,  $+: F \times F \rightarrow F$  (addition) and  $\cdot: F \times F \rightarrow F$  (multiplication), satisfying the following axioms:

**F1.** Addition is commutative. For all  $x, y \in F$ ,

$$x + y = y + x.$$

**F2.** Addition is associative. For all  $x, y, z \in F$ ,

$$(x + y) + z = x + (y + z).$$

**F3.** There is an additive identity. There is an element of  $F$ , denoted  $0$ , such that for all  $x \in F$ ,

$$x + 0 = x.$$

**F4.** There are additive inverses. For all  $x \in F$ , there is an element  $y \in F$  such that

$$x + y = 0.$$

The element  $y$  is denoted  $-x$ . (Subtraction is then defined by  $x - y := x + (-y)$  for all  $x, y \in F$ .)

**F5.** Multiplication is commutative. For all  $x, y \in F$ ,

$$xy = yx.$$

**F6.** Multiplication is associative. For all  $x, y, z \in F$ ,

$$(xy)z = x(yz).$$

**F7.** There is a multiplicative identity. There is an element, denoted  $1$ , that is nonzero and satisfies

$$1x = x$$

for all  $x \in F$ .

**F8.** There are multiplicative inverses. For each nonzero  $x \in F$ , there is a  $y \in F$  such that

$$xy = 1.$$

The element  $y$  is denoted  $1/x$  or  $x^{-1}$ .

**F9.** Multiplication distributes over addition. For all  $x, y, z \in F$ ,

$$x(y + z) = xy + xz.$$

## 2. Order axioms

An *ordered field* is a pair  $(F, F^+)$  where  $F$  is a field and  $F^+$  is a subset of  $F$ , satisfying the following axioms

**O1.** The subset  $F^+$  is closed under addition:

$$x, y \in F^+ \implies x + y \in F.$$

**O2.** The subset  $F^+$  is closed under multiplication:

$$x, y \in F^+ \implies xy \in F.$$

**O3.** For all  $x \in F$ , exactly one of the following statements is true:

$$x \in F^+, \quad -x \in F^+, \quad x = 0.$$

The set  $F^+$  is called the set of *positive* elements of  $F$ . We can use it to define a notion of “greater than” or “less than.” For any two element  $x, y \in F$ , define

$$\begin{aligned} x > y & \text{ if } x - y \in F^+ \\ x < y & \text{ if } y - x \in F^+ \\ x \geq y & \text{ if } x > y \text{ or } x = y \\ x \leq y & \text{ if } x < y \text{ or } x = y. \end{aligned}$$

## 3. Least upper bound property

Let  $S$  be a subset of an ordered field  $F$ . An element  $M \in F$  is an *upper bound* for  $S$  if  $M \geq s$  for all  $s \in S$ . An element  $M \in F$  is the *least upper bound* or *supremum* for  $S$  if it is an upper bound and is less than any other upper bound. In this case, we write  $M = \text{lub } S$  or  $M = \text{sup } S$ . Similarly, an element  $m \in F$  is a *lower bound* for  $S$  if  $m \leq s$  for all  $s \in S$ . An element  $m \in F$  is the *greatest lower bound* or *infimum* for  $S$  if it is a lower bound and is greater than any other lower bound. In this case, we write  $M = \text{glb } S$  or  $M = \text{inf } S$ .

An ordered field  $F$  has the *least upper bound property* if every nonempty subset  $S \subseteq F$  which has an upper bound has a least upper bound in  $F$ .

#### 4. Interval notation

You are probably familiar with the standard notation for intervals of real numbers. If  $x, y \in \mathbb{R}$ , we write

$$\begin{aligned} [x, y] &:= \{z \in \mathbb{R} \mid x \leq z \leq y\}, & (x, y) &:= \{z \in \mathbb{R} \mid x < z < y\} \\ [x, y) &:= \{z \in \mathbb{R} \mid x \leq z < y\}, & (x, y] &:= \{z \in \mathbb{R} \mid x < z \leq y\} \\ (-\infty, y] &:= \{z \in \mathbb{R} \mid z \leq y\}, & [x, \infty) &:= \{z \in \mathbb{R} \mid x \leq z\} \\ (-\infty, y) &:= \{z \in \mathbb{R} \mid z < y\}, & (x, \infty) &:= \{z \in \mathbb{R} \mid x < z\}, \\ (-\infty, \infty) &:= \mathbb{R}. \end{aligned}$$



# Maps of Constant Rank

In the following, let  $U \subseteq \mathbb{R}^n$  and  $V \subseteq \mathbb{R}^m$  be open subsets.

**Definition 0.1.** A mapping  $f: U \rightarrow V$  is smooth if all partial derivatives of all orders for  $f$  exist.

**Definition 0.2.** The mapping  $f: U \rightarrow V$  is a diffeomorphism if it is smooth and bijective (1-1 and onto) and its inverse  $f^{-1}: V \rightarrow U$  is smooth.

**Proposition 0.3.** If  $f: U \rightarrow V$  is a diffeomorphism and  $p \in U$ , then  $Df_p$  is a linear isomorphism (i.e. it is invertible as a linear function). In particular,  $n = m$ .

There is a sort of converse to the preceding proposition. It says that if the derivative is an isomorphism, then  $f$  is locally invertible. This is a striking and important result: linear information at a single point determines the behavior in a whole open neighborhood about that point.

**Theorem 0.4.** (INVERSE FUNCTION THEOREM). For arbitrary smooth  $f: U \rightarrow V$  and  $p \in U$ , if  $Df_p$  is a linear isomorphism, then there exist open sets  $\tilde{U} \subseteq U$  containing  $p$  and  $\tilde{V} \subseteq V$  containing  $f(p)$  such that the restriction of  $f$  to  $\tilde{U}$  is a diffeomorphism onto  $\tilde{V}$ .

**Definition 0.5.** A smooth mapping  $f: U \rightarrow V$  has constant rank  $k$  if the dimension of the image of  $D_p f: \mathbb{R}^n \rightarrow \mathbb{R}^m$  is  $k$  for all  $p \in U$ .

The following generalization of the inverse function theorem is actually equivalent to it:

**Theorem 0.6.** (CONSTANT RANK THEOREM). Suppose  $f: U \rightarrow V$  is a smooth mapping and that  $f$  has constant rank  $k$  in some open neighborhood of  $p \in U$ . Then there exist open sets  $\tilde{U} \subseteq U$  containing  $p$  and  $\tilde{V} \subseteq V$  containing  $f(p)$  along with diffeomorphisms

$$\begin{aligned}\phi: \tilde{U} &\rightarrow U' \\ \psi: \tilde{V} &\rightarrow V'\end{aligned}$$

onto open subsets  $U' \subseteq \mathbb{R}^n$  and  $V' \subseteq \mathbb{R}^m$  such that

$$\psi \circ f \circ \phi^{-1}(x_1, \dots, x_n) = (x_1, \dots, x_k, 0, \dots, 0).$$

In the constant rank theorem says that by changing coordinates (via  $\phi$  and  $\psi$ ), the mapping  $f$  becomes the simple mapping  $\pi(x_1, \dots, x_n) = (x_1, \dots, x_k, 0, \dots, 0)$ . The following commutative diagram is helpful:

$$\begin{array}{ccc} \tilde{U} & \xrightarrow{f} & \tilde{V} \\ \phi \downarrow & & \downarrow \psi \\ U' & \xrightarrow{\pi} & V' \end{array}$$

(To say the diagram *commutes* means that  $\psi \circ f = \pi \circ \phi$ .)

**Corollary 0.7.** (IMPLICIT FUNCTION THEOREM). *Suppose  $f: U \rightarrow \mathbb{R}$  is smooth and  $p \in U$  with  $f(p) = q$ . If, for some  $i$ ,*

$$\frac{\partial f}{\partial x_i}(p) \neq 0,$$

*then there is an open neighborhood  $\tilde{U}$  about  $p$  such that  $f^{-1}(q) \cap \tilde{U}$ , the portion of the level set  $f^{-1}(q)$  in  $\tilde{U}$ , can be parametrized by*

$$(x_1, \dots, x_{i-1}, g(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n), x_{i+1}, \dots, x_n).$$

*for some smooth function  $g$ .*

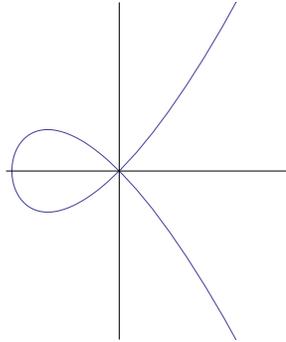
## EXERCISES

- (1) Let  $f(x, y) = (x^2, y)$ . This function takes  $\mathbb{R}^2$  and folds it in half. The Jacobian of  $f$  is

$$Jf(x, y) = \begin{pmatrix} 2x & 0 \\ 0 & 1 \end{pmatrix}.$$

The image of the derivative mapping  $D_{(x,y)}f: \mathbb{R}^2 \rightarrow \mathbb{R}^2$  is the span of  $(2x, 0)$  and  $(0, 1)$ , the columns of  $Jf(x, y)$  (notice that  $x$  is fixed). Thus, the derivative mapping is a linear isomorphism exactly when  $x \neq 0$ . Verify (by finding explicit local inverses) that the function  $f$  is locally invertible exactly at such points, in accordance with the inverse function theorem.

- (2) Let  $c(t) = (t^2 - 1, t(t^2 - 1))$ . The image of  $c$  is shown below.



Check that the dimension of the image of the derivative of  $c$  is 1 at each point, and explain how the constant rank theorem applies, especially at the points  $t = \pm 1$ .

- (3) Let  $f(x, y) = x^2 + y^2$ . The level set  $f^{-1}(1)$  is a circle of radius 1. What does the implicit function say? Especially consider the points  $(1, 0)$  and  $(0, 1)$ .
- (4) Prove that the inverse function theorem is a consequence of the constant rank theorem. You may use the fact that  $Df_p$  is 1-1 if and only if the Jacobian matrix,  $Jf(p)$ , has non-zero determinant. Also, since  $Df_p$  is a linear map from  $\mathbb{R}^n$  to  $\mathbb{R}^n$ , it is 1-1 if and only if it is onto. We make the assumption that the partials of  $f$  are continuous so that the determinant of  $Jf$  is a continuous function. Hence, if it is non-zero at  $p$ , it is non-zero in an open neighborhood about  $p$ .
- (5) Let  $f(x, y) = xy$  and  $p = (1, 0)$ . Show that the implicit function theorem applies at  $p$  and display a corresponding neighborhood  $\tilde{U}$  and the function  $g$  discussed in our statement of the implicit function theorem.



---

# Index

- $\mathbb{R}^n$ , 6, 35
- $\nabla$ , *see also* gradient
- absolute value, *see also* length
- addition, 12, 35
  - of linear functions, 54
  - of matrices, 51
  - of vectors, *see also* vector, addition
- affine
  - function, 56
  - pronunciation, 11
  - subspace, 46
- angle, 42
  - why our definition is reasonable, 42
- approximation, *see also* best affine approximation
- basis, 45
  - standard, 43
- best affine approximation, 11–13, 72–73
  - definition, 12, 72
- bijection, 18
- boundary of a set, 97
- bounded set, 102
- Cartesian product, 17
- Cauchy-Schwarz, 40
- chain rule, 73–76
  - proof of, 73
- closed ball, 65
- closed set, 67
  - and limit points, 68
- codomain, 6, 18
- component
  - along a vector, 40
  - of a vector, 35
- composition of functions, 18
- conservation of energy, 86
- constant rank theorem, 141
- constant rank, smooth function of, 141
- continuity, 69
  - basic properties of, 69
  - implied by differentiability, 85
  - using sequences, 85
- contour, 27
- critical point, 89, 97
  - analysis via Taylor’s theorem, 98–102
  - relation to extrema, 97
- curve on a surface, 25
- derivative, 63–87
  - crash course in calculating, 6–13
  - definition, 70
  - directional, 89–91
    - definition, 89
    - relation to the gradient, 90
  - exists if partials are continuous, 80
    - continuity is necessary, 86
  - geometric interpretations, 15–34
  - given by the Jacobian matrix, 79
  - higher order, 92–97
  - of a composition, 73
  - partial, 76–78
    - definition, 77
    - higher order, 78
    - interpretation, 77
    - introduction, 7
    - order matters, sometimes, 78, 86
  - the case of one variable, 11, 71
  - uniqueness of, 73
- diffeomorphism, 141
- dimension, 45
- directional derivative, *see also* derivative, directional
- discriminant, *see also* quadratic form
- distance, 41
  - basic properties of, 41
- dot product, 37
  - basic properties of, 37
- Euclidean  $n$ -space, 6, 35
- extended binomial theorem, 111

- extreme value theorem, 102
- extremum, 97
  - finding via Taylor's theorem, 98–102
  - guaranteed on a closed bounded set, 102
  - strategy for finding, 102
- function
  - continuous, *see also* continuity
  - definition, 18
  - differentiable, *see also* derivative, definition
  - graph of, 18
  - interpreting, 19–31
- gradient
  - basic properties of, 91
  - definition, 28, 90
  - introduction, 26–29
  - relation to extrema, 97
  - relation to the directional derivative, 90
- helix, 22
- Hessian, 106–109
  - matrix, 108
  - the second derivative test, 109
- hyperplane, 46–49
  - definition, 46
  - equation for, 47
- identity matrix, 53
- image, 18
- implicit function theorem, 142
- injective, *see also* one-to-one
- inner product, *see also* dot product
- interior of a set, 97
- inverse function theorem, 141
- inverse image, 18
- Lagrange multipliers, 104–106
- law of cosines, 59
- length, 38
  - basic properties of, 40
- level set, 27
- limit of a function, 68–70
  - basic properties of, 68
  - definition, 68
- limit of a sequence, 84
- limit point, 67
- linear combination, 45
- linear function, 49–57
  - addition, 54
  - corresponding matrix, 54
  - definition, 49
- linear mapping, *see also* linear function
- linear subspace, 43
  - basis for, 45
  - dimension of, 45
- linear transformation, *see also* linear function
- linearly dependent, 46
- linearly independent, 46
- matrix, 51–56
  - addition, 51
  - column matrix, 54
  - corresponding linear function, 54
  - definition, 51
  - identity matrix, 53
  - multiplication, 52
    - our definition is reasonable, 55
  - scalar multiplication, 51
- maximum
  - finding via Taylor's theorem, 98–102
  - global, 97
    - guaranteed on a closed bounded set, 102
  - local, 97
    - strategy for finding, 102
- metric, 42
- minimum
  - finding via Taylor's theorem, 98–102
  - global, 97
    - guaranteed on a closed bounded set, 102
  - local, 97
    - strategy for finding, 102
- norm, *see also* length
- one-to-one, 18
- onto, 18
- open ball, 65
- open set, 66, 67
- optimization, 89–113
- orthogonal, 38
- orthogonal matrix, 107
- parametrization, 19–26
  - definition, 56
- parametrized curve, 20–22
  - speed, 21
  - tangent vector, 21
  - velocity, 21
- parametrized solid, 26
- parametrized surface, 22
  - tangent plane, 25
  - tangent space, 25
- partial derivative, *see also* derivative, partial
- perpendicular, 38
- plane, 46
- point, 35
- potential, 30
- product rule
  - for dot products, 60
- projection, 40
- Pythagorean theorem, 39
- quadratic form
  - positive definite, negative definite, indefinite, 108
    - associated matrix, 107
  - at a critical point, 99
  - discriminant, 109
- range, 18

- saddle point, 97
- scalar product, *see also* dot product
- second derivative test, *see also* Hessian
- singularity, 26
- smooth function, 141
- span, 44
  - definition, 45
- spectral theorem, 107
- speed, 21
- stationary point, *see also* critical point
- surjective, *see also* onto
- swimming, 104
- symmetric matrix, 107
  
- tangent
  - to a curve, 21
  - to a surface, 25, 33
- Taylor polynomial, 95
- Taylor series, 95
- Taylor's theorem, 92–97
  - proof for one variable, 95, 110
  - proof for several variables, 95
- topographical map, 27
- topological space, 67
- topology, 65–68
  - definition, 67
- translation, 37
- transpose, 106
- triangle inequality, 40
  - reverse, 59
  
- unit vector, 43
  
- vector, 35
  - addition, 35
    - geometric interpretation, 36
    - parallelogram rule, 36
    - properties of, 36
  - component, 35
  - coordinate, 35
  - scalar multiplication, 35
  - standard basis, 43
  - unit, 43
- vector field, 19, 30
- velocity, 21