Multivariate Calculus for Data Analysis

2

by **Fabrizio Donzelli**, with contributions from **Patrick Boily**

This chapter contains an essential introduction to multivariable calculus. The goal is to provide the readers interested in statistics and/or data science with some basic mathematical tools that are at the base of the algorithms and the mathematical models of statistical analysis. Theoretical details, such as rigorous proofs and definitions, will be kept at the minimal level.

A more detailed and complete introduction to multivariable calculus is found at the YouTube channel Calc with Fab and in [23, 24].

2.1 Points, Vectors, Coordinates, Dimensions

We denote by $\mathbb{R}^n$ the **n-dimensional (real) space**. A point P in $\mathbb{R}^n$ is located using the **orthogonal Cartesian coordinates** $(x_1, x_2, \cdots, x_n)$.*

This notation may be adapted according to the context. For instance, we will often denote a **specified point** in $\mathbf{R}^n$ by $\mathbf{a} = (a_1, a_2, \cdots, a_n)$, in contrast with the notation $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ which we reserve for a **generic point**. The number n of coordinates is the **dimension** of $\mathbb{R}^n$.

Given two sets A and B (for examples, two regions in $\mathbb{R}^n$) we write $A \subseteq B$ if A **is a subset of** B (that is, A is contained in B: every element of A is also in B, but the converse is not necessarily true). Let $P = (a_1, \cdots, a_n)$ be a point in $\mathbb{R}^n$, and $D \subseteq \mathbb{R}^n$. We write $P \in \mathbb{R}^n$ if the point **belongs** to the set D, **otherwise** we write $P \notin \mathbb{R}^n$.

The real line $\mathbb{R}$ contains **intervals**:

- **closed** $[a, b]$, the set of all x such that $a \leq x \leq b$;
- **open** (a, b), the set of all x such that $a < x < b$;
- **"clopens"** $(a, b]$ $(a < x \leq b)$ and $[a, b)$ $(a \leq x < b)$, and
- **unbounded** $(a, +\infty)$, $(-\infty, a)$, $(-\infty, +\infty)$.

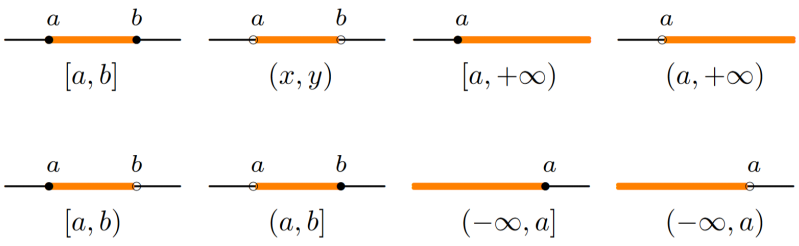


Figure 2.1: Intervals on the real line $\mathbb{R}$.

* We assume some familiarity with most of the following notions, but we suggest reading this short section before moving on to the rest of the chapter, as a refresher.

2.1.1 One Dimension

The (real) one-dimensional space is denoted by $\mathbb{R}$; it is represented by a **line**, oriented **from left to right** along the directon along which values increase. It is common to denote the position of the points along $\mathbb{R}$ by x, but one can choose another name for the variable.[1]

1: We often use t when the real line represents the passage of time, for instance.

The point with coordinate $x = 0$ is known as the **origin** of the line. Positive values of x are located to the **right** of the origin, negative values to the **left**, as in Figure 2.2.

Figure 2.2: The real line $\mathbb{R}$, with origin and direction.

2.1.2 Two and Three Dimensions

The (real) plane $\mathbb{R}^2$ is two-dimensional; we give it (Cartesian) coordinates (x, y), as shown in Figure 2.3.[2] The four **plane sectors** formed by the coordinate axes (red lines) are the plane's **quadrants**, labeled with Roman numerals in counterclockwise order.

2: As was the case in one-dimensional space, the notation of the coordinates may change according to the context: (x_1, x_2) is also used, for instance, but so are polar coordinates (r, θ).

For $\mathbb{R}^3$, we typically use the (Cartesian) coordinates (x, y, z) or (x_1, x_2, x_3).[3]

3: Other options: spherical coordinates, cylindrical coordinates.

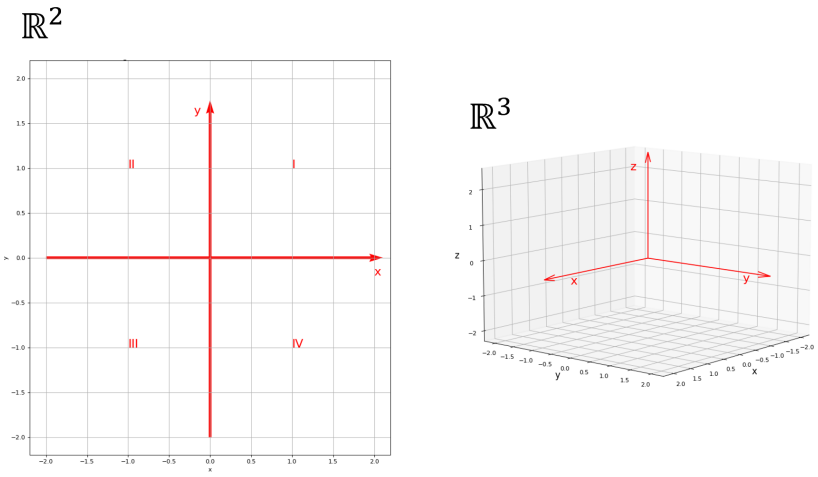


Figure 2.3: The real plane $\mathbb{R}^2$, with origin and quadrants (left); the real space $\mathbb{R}^3$ (right).

In general, we do not display the coordinate axes.[4]

4: Unless we do!

2.1.3 More Dimensions

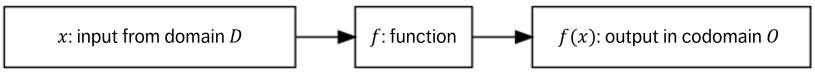
We define the n-dimensional (real) space $\mathbb{R}^n$ as the space described by Cartesian coordinates $\mathbf{x} = (x_1, x_2, \cdots, x_n)$. The point $\mathbb{0} = (0, 0, \cdots, 0)$ is the **origin** of $\mathbb{R}^n$, and it is the point of **common intersection** of the n coordinate axes.

In principle, $\mathbb{R}^n$ is not a vector space, but it can be treated as such and so we can perform vector algebra operation with elements of $\mathbb{R}^n$ (see Chapter 3, *Overview of Linear Algebra*).

2.2 Functions

Functions are the basic objects of calculus, and are the building blocks of mathematical modelling. Functions are in a general sense **input-output machines**, in the sense of the following general definition, which applies beyond calculus.

If D is a set of input values O is the set of output values, then a **function** $f : D \to O$ is a rule that assigns **to each input** element $x \in I$, a **unique output** value, which we denote by $f(x)$. The notation of the function, the input and output set can vary, as usual, according to the context. Once f has been specified, we refer to D as the **domain** of f and to O as its **codomain**.



If $f : D \to O$ is a function, the set $f(D) = \{f(x) \mid x \in D\} \subseteq O$ is called the **range** (or the **image**) of f.

Examples

1. Let P be the collection of patients in a COVID emergency hospital, and $O = \{p(\text{ositive}), n(\text{egative})\}$ be the set of possible test responses. We construct the "COVID-TEST" function $T : P \to O$ as follows: If $x \in P$,

$$T(x) = \begin{cases} p, & \text{if patient } x \text{ tests positive} \\ n, & \text{if patient } x \text{ tests negative} \end{cases}$$

   In this example the output values are **categorical**, since they classify the patients into a discrete set of (fixed) classes.[5]

2. Let S denote a sphere of arbitrary radius. A point on S can be located using two coordinates: its **longitude** and its **latitude**.[6] We can then define the temperature function $T : S \to \mathbb{R}$ by

$$T(\text{longitude}, \text{latitude}) = \text{temperature at the point.}$$

   The temperature function is usually assumed to be **continuous**.[7]

3. **Probability theory** is naturally expressed in the language of multivariate calculus (see Chapter 6). For instance, the **density function** of the **multivariate normal distribution** in 2 uncorrelated variables of expectation **0** is a function $f_{\sigma_1,\sigma_2} : \mathbb{R}^2 \to \mathbb{R}$ defined by:

$$f_{\sigma_1,\sigma_2}(x, y) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{x^2 + y^2}{2}\right).$$

   The probability that a randomly selected point $P = (x, y)$ from this distribution falls in $\Omega \subseteq \mathbf{R}^2$ is an integral:

$$\iint_\Omega f_{\sigma_1,\sigma_2}(x, y)\, dA.$$

   We will discuss such notions further in Section 2.6, 6.3, and 6.4.

5: In statistics, it is often convenient to represent categorical variables with **numeric** values. For example, we can assign $f(x) = 1$ if the patient x has a positive test, $f(x) = 0$ if their test is negative.

6: Assuming that a special point and great circle through that point have been identified.

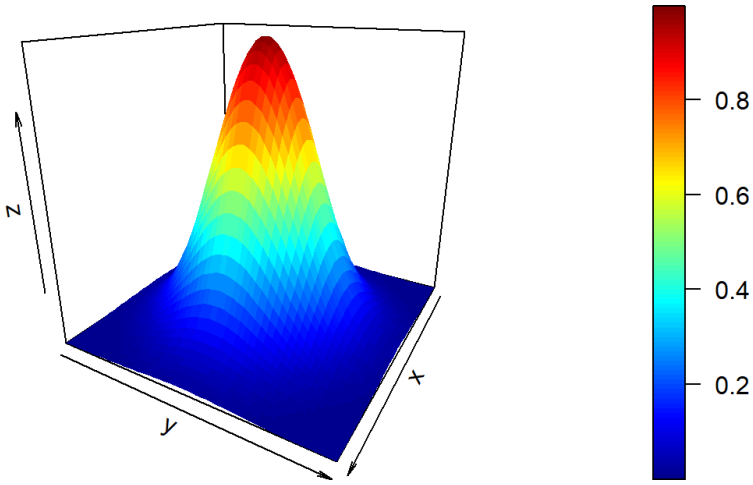
7: We will not be discussing this concept except in an intuitive manner: a continuous function is one in which there are no "jumps". An interesting corollary is that if we model the temperature on the Earth in that manner, we can show that at any given moment there are at least two antipodal points which have exactly the same temperature.

4. The following block of R code provides a display of the 3D surface $z = \exp(-x^2 - y^2)$ over $\{(x, y) \in \mathbb{R}^2 \mid -2 \leq x, y \leq 2\}$.

3D plotting in R

```
library(plot3D) # for 3D plotting

M <- mesh(seq(-2, 2, length.out = 50),
          seq(-2, 2, length.out = 50))
u <- M$x ; v <- M$y
x <- u
y <- v
z <- exp(-x^2-y^2)
surf3D(x, y, z, colvar = z, colkey = TRUE,
       box = TRUE, bty = "b", phi = 20, theta = 120)
```



Note: the domain of a function is part of the recipe, it is not automatically defined by the function itself. However, in calculus, when we use the word **domain**, we usually mean the **largest set** D_f to which the function could be applied. For any x in D_f, there is a **unique output** $f(x)$.[8]

8: That is not necessarily the case in the general framework of **multivalued functions**, which, while quite interesting from a geometrical perspective, are outside the scope of this document.

Examples

1. What is the (largest possible) domain D_f of the function defined by $f(x, y) = \frac{1}{x+y}$? We cannot divide by zero, so the denominator $x + y$ can never be zero when we apply the function $f(x, y)$; D_f therefore consists of all pairs (x, y) expect for those satisfying the equation $x + y = 0$, whose solution set is the line $y = -x$. Thus,

$$D_f = \{(x, y) \in \mathbb{R}^2 \mid x + y \neq 0\};$$

   in other words, the domain consists of the region above the line $y = -x$ and the region below the line $y = -x$.
2. What is the domain D_f of $f(x, y, z, w) = \ln(w) + x + y + z$? Recall that the (real) logarithm is defined only for positive input values. Hence the domain is $D_f = \{(x, y, z, w) \in \mathbb{R}^4 \mid w > 0\}$.

2.3 Graphical Representation of Functions

Human eyes (and brains) have a difficult time parsing large data files directly; we typically rely on **graphical representations** to make sense of data (see Chapter 16 and [25] for a *lot* more information on the topic).

Graphical representations are useful in calculus as well; we review a few standard ways of providing these for functions of several variables.

2.3.1 One Variable: Sketch the Graph

Let $f : (a, b) \to \mathbb{R}$ be a function of one variable x. The **graph** of f is the curve of equation $y = f(x)$; a point in the graph is given by coordinates $(x, f(x))$, for $x \in (a, b)$.

Example Sketch the graph of the function $f : [0, \infty) \to \mathbb{R}$ defined by

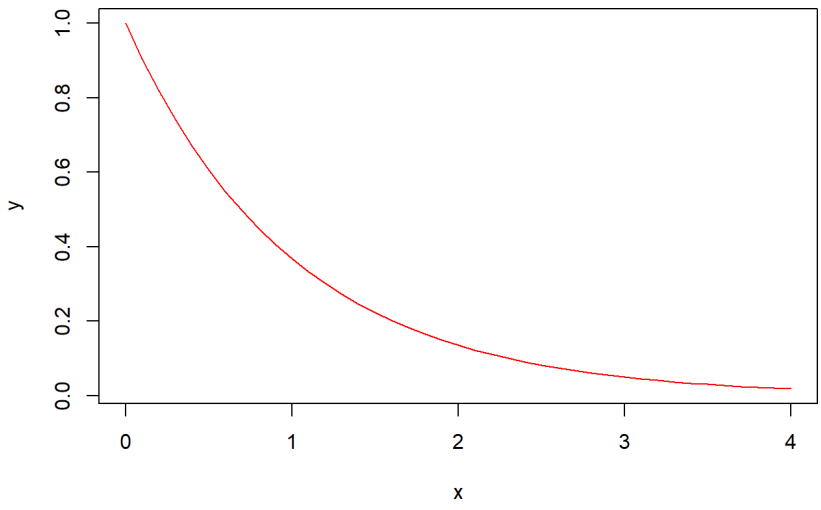
$$f(x) = e^{-x} \text{ for } x \geq 0.$$

Does the point $(1, 2)$ belong to the graph of f?[9]

Note that the domain is restricted to the half-real line $x \geq 0$; since the exponent is negative, e^{-x} decays to 0 as $x \to \infty$ (quite rapidly in fact).

9: This is essentially an example of the **exponential distribution**.

```
x <- seq(0,4,0.1)
y <- exp(-x)
plot(x, y, type='l', col = rainbow(25), lty=1)
```



To answer the last question, we evaluate $f(1)$; it is equal to $e^{-1} \neq 2$, and so the point is not on the graph.

2.3.2 Two Variables: Graphs or Level Curves

For function of two variables, there are two convenient ways to provide a graphical representation.

The Graph of a Function

Let $f : D \to \mathbb{R}$ be a function of two variables x, y, where $D \subseteq \mathbb{R}^2$. The **graph** of f is the **surface** of equation $z = f(x, y)$.

A point on the graph is given by coordinates $(x, y, f(x, y))$, where $(x, y) \in D$. We can interpret the graph as a **hilly region**, in which case (x, y) are the coordinates of the position with reference to $xy-$plane, and z is the altitude.

Example Sketch the graph of the function $f : D \to \mathbb{R}$ defined by

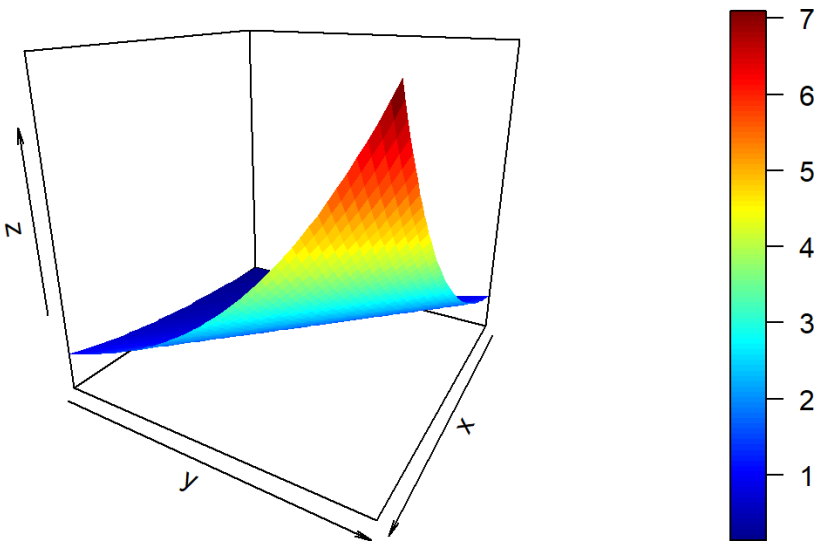
$$f(x, y) = e^{x+y}, \text{ for } -1 \leq x \leq 1, -1 \leq y \leq 1.$$

Interpret the graph.

We can recycle the code from one of the previous examples.

```
library(plot3D)

M <- mesh(seq(-1, 1, length.out = 50),
          seq(-1, 1, length.out = 50))
u <- M$x ; v <- M$y
x <- u
y <- v
z <- exp(x+y)
surf3D(x, y, z, colvar = z, colkey = TRUE,
       box = TRUE, bty = "b", phi = 20, theta = 120)
```



Level (Contour) Curves

Let $f : D \subseteq \mathbb{R}^2 \to \mathbb{R}$. Depending on the nature of f, the graph may be difficult to read (or to plot). An alternative may be to sketch the **level** (or contour) **curves**.

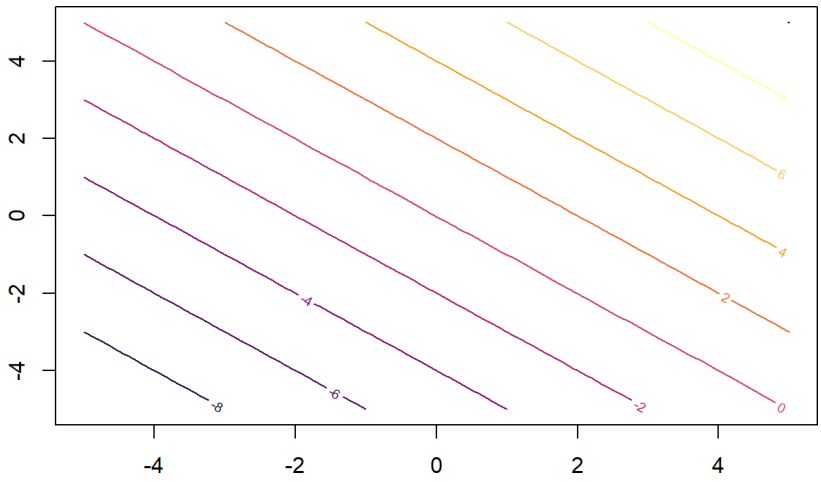
Let c be a value in the range of f, which is to say, a **possible output value** of f. Generically, the equation $f(x,y) = c$ is a **curve** in the $xy-$plane, a **level curve** (or **contour curve**) of f, which consists of **all (and only)** the points $(x,y) \in D$ where the function takes the value c.

Example Plot a few level curves of the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by $f(x,y) = x + y$.

For any fixed value c, the equation $x+y = c$ can be rewritten as $y = -x+c$. The level curves of f are thus all the lines in the $xy-$plane with slope -1. Along each line of equation $y = -x + c$, the value of f is given by the $y-$intercept.

Here is a sample code for plotting the level curves of f; the numbers displayed on top of the curves are the values c taken by the function along the curves displayed.

```
x <- seq(-5,5,length.out=50)
y <- seq(-5,5,length.out=50)
z <- outer(x,y,"+")
cols <- hcl.colors(10, "Inferno") #color palette
contour(x,y,z,col=cols)
```



We can use level curves to estimate the values of a function in a certain region of the domain.

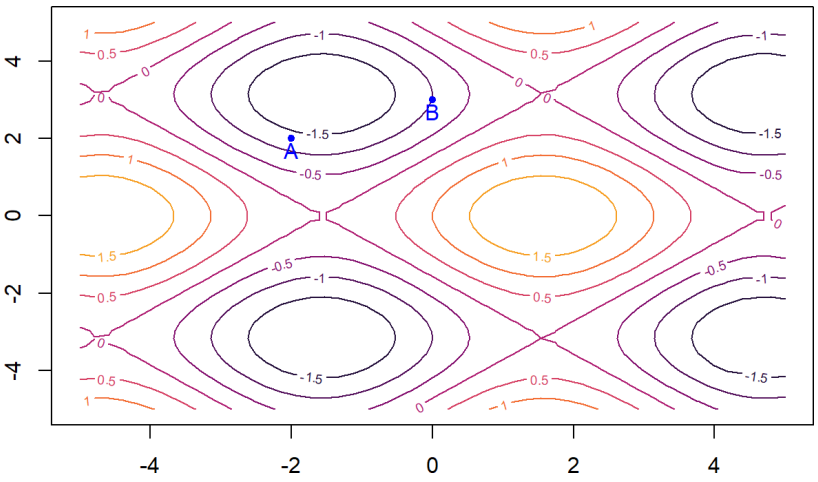
Example Given the following level curves of $f(x,y) = \sin(x) + \cos(y)$, estimate the value of f at A and B.

Level curves in R

```
x <- seq(-5,5,length.out=50)
y <- seq(-5,5,length.out=50)
z <- outer(sin(x),cos(y),"+")

cols <- hcl.colors(10, "Inferno") #color palette
```

```
contour(x,y,z, col=cols)
points(-2,2,col='blue',pch=20)
points(0,3,col='blue',pch=20)
points(-2,1.7,col='blue',pch="A")
points(0,2.7,col='blue',pch="B")
```



The point A is located between the level curves $f(x, y) = -1$ and $f(x, y) = -1.5$. Since it is slightly closer to the second curve, we can estimate $f(A) \approx -1.3$.

The point B seems to sit exactly along the level curve $f(x, y) = -1$, hence $f(B) \approx -1$.[10]

10: Of course, we can double check this estimate by finding the coordinates of A and B, and computing $f(A)$ and $f(B)$.

Example Level curves may **degenerate** to **lower dimensional regions**, or, even "worse", be empty when c is not in the range of f.

As an illustration, consider the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = x^2 + y^2$:

- for $c > 0$, the level curve $x^2 + y^2 = c$ is the circle of center $(0, 0)$ and radius $\sqrt{c}$;
- the level curve $x^2 + y^2 = 0$ degenerates to the point $(0, 0)$, the only point whose coordinates solve the equation $x^2 + y^2 = 0$;
- for $c < 0$, the level curve $x^2 + y^2 = c$ does not exist, since $x^2 + y^2 \geq 0$ for all real values of x and y.

2.3.3 Three or More Variables

The more variables we have, the more challenging it can be to provide graphical representations of a function.

However both graphs and level sets can be defined, in purely mathematical terms, over an arbitrary number of variables, without needing to be visualized.

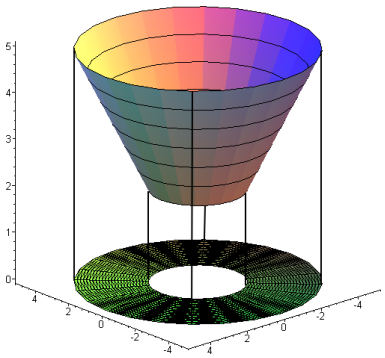
The Graph of a Function

Let $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ be a function of n variables $\mathbf{x} = (x_1, \cdots, x_n)$. The **graph** of f is the $n-$dimensional **hypersurface** in $\mathbb{R}^{n+1}$ defined by the equation $w = f(\mathbf{x}) = f(x_1, \cdots, x_n)$, for $\mathbf{x} \in D$. A point on the graph is therefore identified by coordinates

$$(\mathbf{x}, f(\mathbf{x})) = (x_1, x_2, \cdots, x_n, f(x_1, x_2, \cdots, x_n)),$$

with $\mathbf{x} = (x_1, x_2, \cdots, x_n) \in D$. We can interpret f as a way of bending and stretching the domain D into a new region embedded in $\mathbb{R}^{n+1}$.[11]

11: We illustrate this for $n = 2$ below:



The cone is a distortion in $\mathbb{R}^3$ of the ring in $\mathbf{R}^2$.

Level (Contour) Sets

Let $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ and let c be a value in the range of f. Generically, the equation $f(\mathbf{x}) = f(x_1, \ldots, x_n) = c$ is an $n-1$ dimensional region (**hypersurface**) in D, called a **level set** (or **contour set**) of f, which consists of **all (and only)** the points $\mathbf{x} = (x_1, \ldots, x_n) \in D$ where the function takes the value c.

Level sets may **degenerate** to **lower dimensional regions** $< n-1$, or be empty when c is not in the range of f.

Example Describe the level sets of the function $f : \mathbb{R}^3 \to \mathbb{R}$ defined by $f(x, y, z) = x^2 + y^2 + z^2$. Are there "degenerate" level sets?

In $\mathbb{R}^3$, the equation of the 2D sphere of radius $R > 0$ and centre at the origin $\mathbf{0} = (0, 0, 0)$ is $x^2 + y^2 + z^2 = R^2$. Thus, the level sets of the functions consists of spheres all centered at the origin.

If $R = 0$, the equation $x^2 + y^2 + z^2 = 0$ is satisfied only for the zero dimensional set $\{(x, y, z) \mid x = y = z = 0\}$; this level set is degenerate.

2.3.4 Scalar-Valued Functions and Vector Fields

Let $D \subseteq \mathbb{R}^n$ be a n-dimensional domain. A **real valued function** $f : D \to \mathbb{R}$ will be called a **function** (or a **scalar field**), in contrast with a **vector valued function** $\mathbf{F} : D \to \mathbb{R}^n$, which we call a **vector field**.

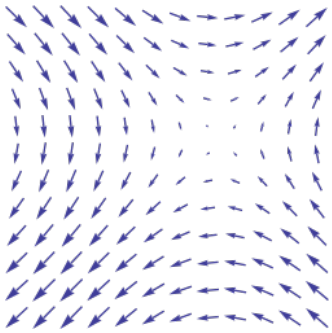


Figure 2.4: An illustration of the 2D vector field $\mathbf{F}(x, y) = (\sin y, \sin x)$ [author unknown].

Vector fields play a crucial role in vector calculus and its applications to physics and geometry, but this is out of scope for our purposes. We refer again the reader to [23].

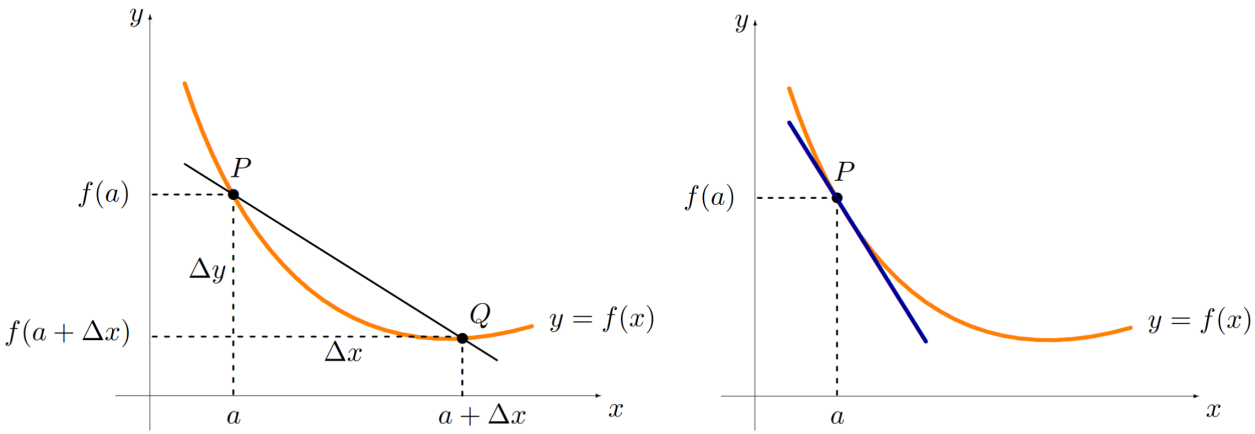


Figure 2.5: Difference quotient and slope of the tangent to $y = f(x)$ at $P(a, f(a))$.

2.4 Derivatives

After an introduction to functions, the next step is to define the **derivative**, which provides a unified way of measuring the rate of change of a function with respect to its variables.

2.4.1 Limit of Difference Quotients

Let $f : (c, d) \to \mathbb{R}$ be a function of one variable x and $x = a \in D_f = (c, d)$. The **derivative** of $f(x)$ at $x = a$ is denoted by $f'(a)$ and is defined as the limit (if it exists) of the difference quotients

$$f'(a) = \lim_{\Delta x \to 0} \frac{f(a + \Delta x) - f(a)}{\Delta x}.$$

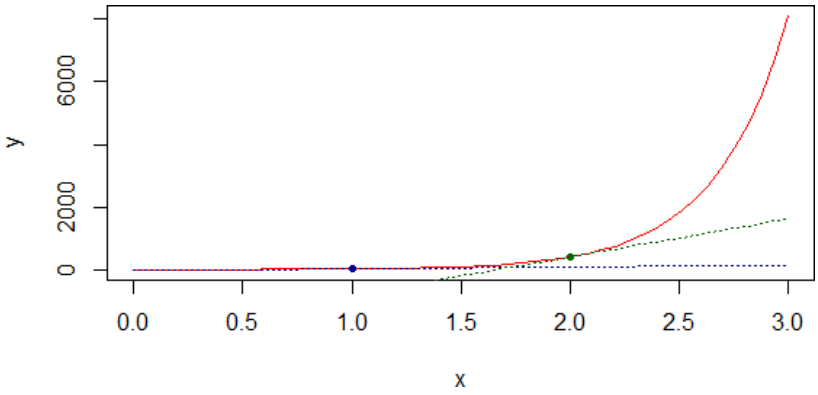
The **number** $f'(a)$ is a measure of the rate of change of f at $x = a$. Geometrically, the value $f'(a)$ is the slope of the tangent line to the graph of f at the point $(a, f(a))$.

In general, the value of the derivative of f depends on x; we therefore define the **derivative function** $f' : (c, d) \to \mathbf{R}$, which also carries the meaning of **slope function**.

Example Consider the exponential function f defined by $f(x) = e^{3x}$ on $\mathbb{R}$, whose graph is represented by the red curve below.

```
x <- seq(0, 3, length.out=50)
y <- exp(3*x)

plot(x, y, type='l', col=rainbow(25), lty=1)
lines(x, 3*exp(3)*x-3*exp(3)+exp(3),
      col='darkblue', lty=3)
points(1, exp(3), pch=20, col='darkblue')
lines(x, 3*exp(6)*x-3*2*exp(6)+exp(6),
      col='darkgreen', lty=3)
points(2, exp(6), pch=20, col='darkgreen')
```



The graph also shows two tangent lines. The slope of each tangent line is the **rate of change** of f at x. By comparing the slopes of the two tangent lines, we observe that the rate of change at $x = 2$ is much larger than the rate of change at $x = 1$, in accordance with the fact that the exponential function grows quite quickly.

The process of calculating the derivative of f is sometimes referred as **differentiation**. The derivative is denoted in two ways:

$$\frac{df(x)}{dx} \quad \text{or} \quad f'(x),$$

it is up to the reader which one (if not both) to use.

2.4.2 Rules of Differentiation

But there is no need to use the definition *via* the limit of differential quotients to compute the derivative of a function. The set of **differentiation rules** are recalled here for readers' convenience.[12]

12: A detailed discussion about differentiation can be found in [24, 26].

In the following list, x denotes the variable, while a and n are constants.

1. For a **constant function** f, $f'(x) = 0$
2. **Power rule:** $(x^n)' = nx^{n-1}$
3. **Exponentials:** $(e^{ax})' = ae^{ax}$
4. **Logarithms:** $(\ln(x))' = \frac{1}{x}$
5. **Product rule:** $(f(x)g(x))' = f'(x)g(x) + f(x)g'(x)$
6. **Quotient rule:** $\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}$
7. **Chain rule:** $f(g(x))' = f'(g(x))g'(x)$

The chain rule, for instance, is important for understanding the construction of the **backpropagation** algorithm of neural network models (see Chapter **??** and [27], say).

Example Using the rules, compute the derivative of $f(x) = e^{-x^2}$.[13] What is the value of the rate of change of $f(x)$ at $x = 2$?

13: We will stop using the convoluted phrasing "the function $f : A \to B$ defined by $f(x) = \ldots$" and substitute instead "the function $f(x) = \ldots$" when the context allows it.

From the exponentials derivative rule and the chain rule, we obtain:

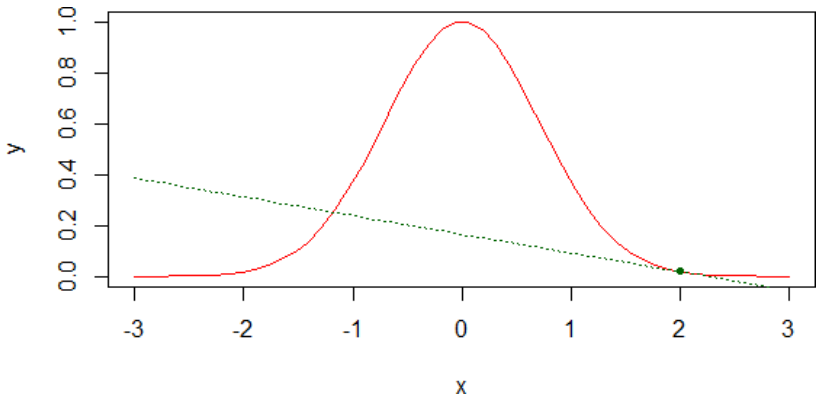
$$f'(x) = (e^{-x^2})' = e^{-x^2}(-x^2)' = -2xe^{-x^2}$$

At $x = 2$, the rate of change of $f(x)$ is

$$f'(2) = -2 \times 2 \times e^{-2^2} = -0.073$$

The **slope** (or **rate of change**) at $x = 2$ is negative, as expected by inspecting the shape of the bell curve representing the curve $y = e^{-x^2}$. Its value is "small", which is also expected since the function decays to zero quite rapidly.

```
x = seq(-3, 3, length.out=50)
y = exp(-x^2)
plot(x, y, type='l', col = rainbow(25), lty=1)
lines(x, -2*2*exp(-2**2)*(x-2)+exp(-2**2), col='darkgreen',
      lty=3)
points(2, exp(-2**2), pch=20, col='darkgreen')
```



2.4.3 Partial Derivatives

How do we expand this definition to functions of several variables? In this case, we are interested in defining and computing the rate of change with respect to any of the variables. This is done via **partial derivatives** which, computationally speaking, are a straightforward generalization of the notion of derivative of a function of one variable.

Partial Derivatives of Order 1

Let $f(x_1, \cdots, x_n)$, and pick any variable x_k, for some $k \in \{1, \cdots, n\}$, with respect to which we want to compute the rate of change of f. We can use the one-variable differentiation rules from Section 2.4.2 by treating the remaining variables as constant.

The **partial derivative of order one** of f with respect to the variable x_k, denoted in two alternative ways as follows:

$$\lim_{\Delta x \to 0} \frac{f(x_1, \ldots, x_k + \Delta x, \ldots, x_n) - f(x_1, \ldots, x_k, \ldots, x_n)}{\Delta x} = \frac{\partial f}{\partial x_k} = f_{x_k}.$$

Example Compute the 3 partial derivatives of $f(x, y, z) = x^2y + 3xz$.

We have 3 variables, and we compute the corresponding partial derivative for each of them:

$$f_x(x, y, z) = \frac{\partial(x^2y + 3xz)}{\partial x} = 2xy + 3z$$
$$f_y(x, y, z) = \frac{\partial(x^2y + 3xz)}{\partial y} = x^2$$
$$f_z(x, y, z) = \frac{\partial(x^2y + 3xz)}{\partial z} = 3x$$

Tangent Plane

If $f : \mathbb{R} \to \mathbb{R}$ is differentiable at $x = a$, the equation of the unique **tangent line to the graph** $y = f(x)$ at $P(a, f(a))$ is

$$y = f'(a)(x - a) + f(a).$$

More generally, if $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ is differentiable at $\mathbf{x} = \mathbf{a}$, there are infinitely many tangent lines to its graph $w = f(\mathbf{x})$ at $P(\mathbf{a}, f(\mathbf{a}))$. All of these lines lie in the same unique **tangent hyperplane**.

When $n = 2$, we have a **tangent plane** to $z = f(x, y)$ at $P(a, b, f(a, b))$; it is the plane that rests on the surface, touching it only at the point of tangency, as illustrated in the figure below.[14]

14: Near the point of tangency, the surface resembles the tangent plane: this is partly why that we've long believed the Earth to be flat!

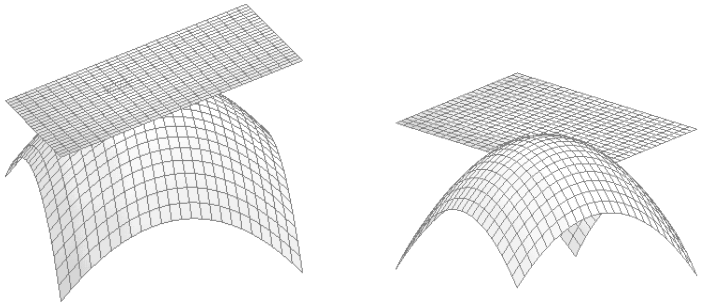


Figure 2.6: Tangent placn to $z = -x^2 + y^2$ at $(0, 1, 1)$, seen from two different angles.

When such a plane exists, as do the partial derivatives, the surface is said to be **differentiable** at the point in question.

If $z = f(x, y)$ is a differentiable surface $P(a, b, f(a, b))$, the equation of the tangent plane to the surface at point P is

$$z = f(a, b) + f_x(a, b)(x - a) + f_y(a, b)(y - b).$$

Example Find the tangent plane to $z = \sqrt{x - y}$ at $P(2, 1, 1)$

First, we verify that P is indeed on the surface. Since $a = 2$ and $b = 1$, we simply need to check that $\sqrt{a - b} = \sqrt{2 - 1} = 1$, which is indeed the case.

Next we compute the partial derivatives

$$f_x(x,y) = \frac{1}{2\sqrt{x-y}} \quad \text{and} \quad f_y(x,y) = -\frac{1}{2\sqrt{xx-y}}.$$

Thus

$$f_x(a,b) = f_x(2,1) = \frac{1}{2\sqrt{2-1}} = \frac{1}{2} \text{ and } f_y(a,b) = f_y(2,1) = -\frac{1}{2\sqrt{2-1}} = -\frac{1}{2},$$

so the equation of the tangent plane is

$$\begin{aligned} z &= f(2,1) + f_x(2,1)(x-2) + f_y(2,1)(y-1) \\ &= 1 + \frac{1}{2}\cdot(x-2) - \frac{1}{2}(y-1) = \frac{1}{2}(1+x-y). \end{aligned}$$

When the partial derivatives do not exist at a particular point on the surface, then either there is no tangent plane or it is **not unique**.

For example, the partial derivatives of $f(x,y) = 2 - \sqrt{x^2+y^2}$ are not defined when $(x,y) = (0,0)$ (which is in the domain of f); graphically, this translates into more than one tangent plane at the vertex of the cone $z = 2 - \sqrt{x^2+y^2}$, as shown below.

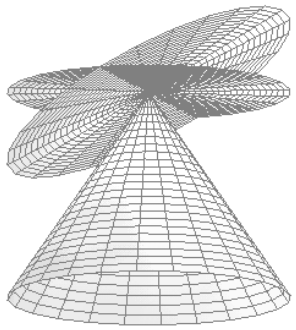


Figure 2.7: Two tangent planes at the vertex of the cone $z = 2 - \sqrt{x^2+y^2}$.

Partial Derivatives of Order 2

15: For example, in optimization.

In calculus problems,[15] it is convenient to have at hand the **partial derivatives of order two**. Let $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ and pick any two variables x_h, x_k, for $k, h \in \{1, 2, \cdots n\}$.

The **partial derivative of order two** with respect to x_h and x_k (in that order) is the function

$$f_{x_h x_k}(x_1, \cdots, x_n) = \frac{\partial^2 f(x_1, \cdots, x_n)}{\partial x_k \partial x_h}$$

obtained by first computing the partial derivative with respect to x_h, and then the partial derivative of that partial derivative with respect to x_k.

But what if, when computing a partial derivative of order two, we mistakenly change the order of differentiation with respect to the two chosen variables?

It turns out that for sufficiently regular functions the order does not matter, thanks to Clairaut's Theorem, which is explained in Figure 2.8; "higher order" means that we can keep differentiating f,[16] obtaining partial derivatives of order $3, 4, \ldots$ and so on.

16: When the function is differentiable, it needs to be added.

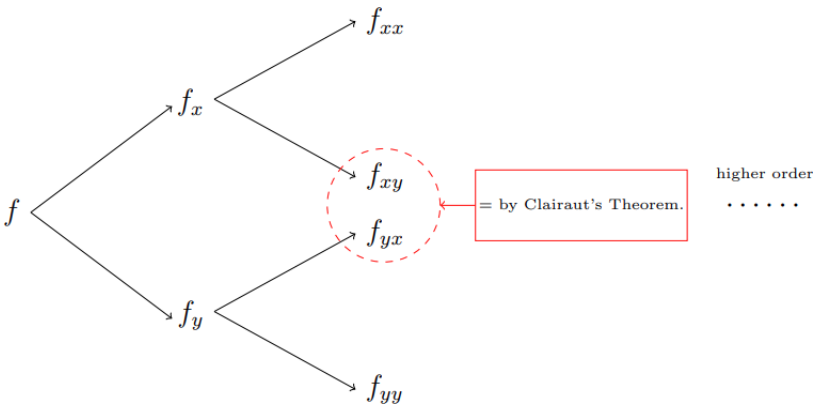


Figure 2.8: Illustration of Clairaut's theorem in 2 variables.

Clairaut's Theorem applies to the "standard functions" that we introduce in calculus courses, obtained by combining polynomials, rational funtions, trigonometric functions, exponentials and logarithmig functions, analytic functions (power series), etc.

Example Consider such a standard function of 3 variables (x, y, z). In theory, f has 9 partial derivatives of order 2:

$$f_{xx}, f_{xy}, f_{xz}, f_{yx}, f_{yy}, f_{yz}, f_{zx}, f_{zy}, f_{zz}$$

But thanks to Clairaut's Theorem, we have:

$$\begin{aligned} f_{xy} &= f_{yx} \\ f_{xz} &= f_{zx} \\ f_{yz} &= f_{zy} \end{aligned}$$

We only need to compute 6 partial derivatives of order 2 to obtain them all!

2.4.4 Gradients

From the point of view of data analysis, the most important vector fields are the gradients of multivariate functions $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$.

The **gradient** $\nabla f : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ is defined by:[17]

17: Pronounced "nabla".

$$\nabla f(x_1, \cdots, x_n) = \langle f_{x_1}(x_1, \cdots, x_n), \cdots, f_{x_n}(x_1, \cdots, x_n). \rangle$$

The $\langle \ldots \rangle$ notation is used to distinguish vector fields (and vectors) from points in $\mathbf{R}^n$, which are denoted using $(\ldots)$.[18]

18: The gradient is not only a way to collect the first order partial derivatives of a function into a vector, but it carries important geometrical information about the function, as we shall soon see.

Example We can easily ompute the gradient of $f(x, y, z) = x^2y + z$, and evaluate it at $(-1, 1, 2)$.

Indeed,

$$\nabla f(x, y, z) = \langle 2xy, x^2, 1\rangle.$$

At $(-1, 1, 2)$, the gradient becomes a 3-dimensional **vector**:

$$\nabla f(-1, 1, 2) = \langle 2 \cdot (-1) \cdot 1, (-1)^2, 1\rangle = \langle -2, 1, 1\rangle.$$

Gradient and Level Sets

There is a crucial property linking the gradient of a function $f : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ and its level sets: wherever $\nabla f(\mathbf{x}) \neq \mathbf{0}$, the gradient is **perpendicular** to the level sets of f.

More precisely, given a point $\mathbf{a} = (a_1, \cdots, a_n) \in D$, if $\nabla f(\mathbf{a}) \neq \mathbf{0} = (0, \ldots, 0)$, then $\nabla f(\mathbb{a}) \perp L_{\mathbf{a}}$, where $L_{\mathbf{a}}$ is the level set of f through $\mathbf{a}$. In $\mathbb{R}^2$, we can visualize this property quite easily.

Example Consider the function $f(x, y) = x^2 + y^2$, whose level curves are concentric circles. The gradient vector field is represented by the vectors in Figure 2.9. Since $\nabla f)(x, y) = \langle 2x, 2y\rangle$, the gradient is a radial vector field,[19] and the orthogonality is a simple consequence of Euclidean geometry.[20]

19: The vectors point along the radii of the level circles.

20: A radius meets its circle orthogonally [24].

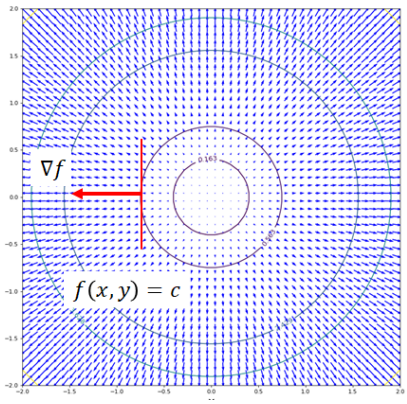


Figure 2.9: The gradient $\nabla f = \langle 2x, 2y\rangle$ is perpendicular to the level sets $x^2 + y^2 = c$, as is illustrated with $(x, y) = (-1, 0)$.

2.4.5 Directional Derivatives

In studying a function whose domain D is a region of $n-$dimensional space $\mathbb{R}^n$, we usually choose n **preferred** pairwise orthogonal directions, corresponding to the n cartesian coordinates $(x_1, \cdots, x_n)$. Those directions are given by the **canonical basis vectors**

$$\mathbf{e}_1 = \langle 1, 0, \cdots, 0\rangle$$
$$\vdots$$
$$\mathbf{e}_n = \langle 0, 0, \cdots, 1\rangle$$

Note that each canonical basis vector is of length 1. In $\mathbb{R}^3$ we also denote the canonical basis by $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\} = \{\mathbf{i}, \mathbf{j}, \mathbf{k}\}$.

The **rate of change** of f **along the direction** $\mathbf{e}_k$ is the partial derivative f_{x_k}. We can also use **any** direction $\mathbf{u}$ with unit length. We can find the appropriate formula using "minimally intuitive'' reasoning.[21]

21: To quote Dr. De Oliveira.

The vector $\mathbf{u}$ is a linear combination of the basis elements:

$$\mathbf{u} = c_1\mathbf{e}_1 + \cdots + c_n\mathbf{e}_n.$$

As we have discussed, the rate of change of f along $\mathbf{e}_k$ is f_{x_k}. If $\mathbf{u}$ is of length 1, we can interpret the linear combination above as a **signed weighted average** of the canonical basis vectors $\mathbf{e}_k$; consequently, it is reasonable to define the rate of change of f along $\mathbf{u}$ as the signed weighted average of the partial derivatives f_{x_k}, with the same coefficients c_k.[22]

22: The proof that this indeed the right approach to take is an easy consequence of the chain rule.

Link With the Gradient

Given a unit vector

$$\mathbf{u} = c_1\mathbf{e}_1 + \cdots + c_n\mathbf{e}_n,$$

the **directional derivative** of f along $\mathbf{u}$ is

$$D_{\mathbf{u}}f(x_1, \cdots, x_n) = c_1 f_{x_1}(x_1, \cdots, x_n) + \cdots + c_n f_{x_n}(x_1, \cdots, x_n).$$

Using the **dot product** of vectors, we can also write

$$D_{\mathbf{u}}f(x_1, \cdots, x_n) = \nabla f(x_1, \ldots, x_n) \cdot \mathbf{u}.$$

Example What is the directional derivative of $f(x, y) = \cos(xy) + y$ along the unit vector $\mathbf{u} = \frac{1}{\sqrt{2}}\langle 1, 1\rangle$ at the point $(1, 1)$?

We start computing the gradient of f:

$$\nabla f(x, y) = \langle -y\sin(xy), -x\sin(xy) + 1\rangle.$$

The directional derivative as a function (that is, for arbitrary x, y) is

$$\begin{aligned} D_{\mathbf{u}}f(x, y) &= \nabla f(x, y) \cdot \mathbf{u} = \langle -y\sin(xy), -x\sin(xy) + 1\rangle \cdot \frac{1}{\sqrt{2}}\langle 1, 1\rangle \\ &= -\frac{1}{\sqrt{2}}y\sin(xy) + \frac{1}{\sqrt{2}}(-x\sin(xy) + 1). \end{aligned}$$

At $x = 1, y = 1$ we obtain

$$D_{\mathbf{u}}f(1, 1) = -\frac{1}{\sqrt{2}}\sin(1) + \frac{1}{\sqrt{2}}(-1\sin(1) + 1) = -\sqrt{2}\sin(1) + \frac{1}{\sqrt{2}}$$

Minimum and Maximum Rate of Change

Let $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ and $\mathbf{a} = (a_1, \cdots, a_n) \in D$ with $\nabla f(\mathbf{a}) \neq \mathbf{0}$. The **maximum** rate of change of f at $\mathbf{a}$ occurs along the direction of the

gradient,

$$\frac{\nabla f(\mathbf{a})}{||\nabla f(\mathbf{a})||},$$

while the **minimum** rate of change of f at $\mathbf{a}$ occurs along the opposite direction.

To understand this last statement let us reason in the case of a function of two variables whose graph $z = f(x, y)$ is a surface. In order to climb or go down the hill along the steepest way, we move perpendicularly to the contour line of the hill located at a certain height. The orthogonal direction is given by the gradient.[23]

23: This property is crucial in understanding the **gradient descent** algorithm that searches for the minimum values of a function (the cost function). See Chapter **??**, *A Deep Learning Launchpad*.

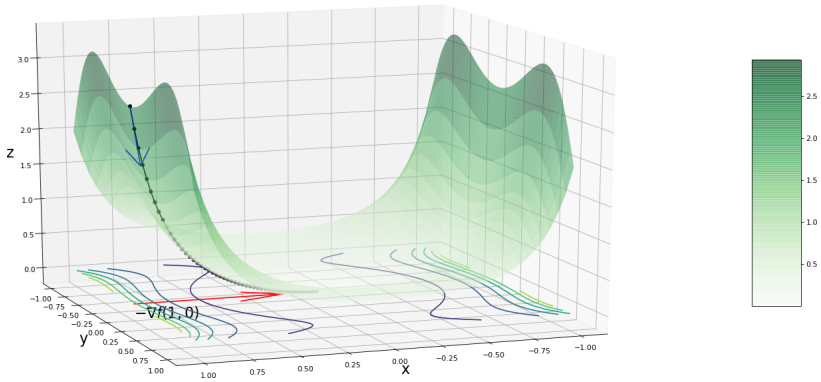


Figure 2.10: Gradient descent search for the minimum of $z = (x^2+y^2)\exp(x^4-y^4)$ [27].

Example What is the maximum rate of change of $f(x, y) = x^2 + y^2$ at $(1, 1)$?

We start with the calculation of the gradient

$$\nabla f(x, y) = \langle 2x, 2y \rangle.$$

At $(x, y) = (1, 1)$, the gradient is

$$\nabla f(1, 1) = \langle 2, 2 \rangle,$$

the unit vector corresponding to the direction of maximum rate of change is thus

$$\mathbf{u} = \nabla f(1, 1)||\nabla f(1, 1)||\frac{1}{\sqrt{2}}\langle 1, 1 \rangle.$$

The value of the maximum rate of change is thus given by:

$$D_{\mathbf{u}}f = \nabla f(1, 1) \cdot \mathbf{u} = \nabla f(1, 1) = \langle 2, 2 \rangle \cdot \frac{1}{\sqrt{2}}\langle 1, 1 \rangle = 2\sqrt{2}.$$

For a general $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ and $\mathbf{a} \in D$ such that $\nabla f(\mathbf{a}) \neq \mathbf{0}$, the value of the **maximum rate of change** of f at $\mathbf{a}$ is $||\nabla f(\mathbf{a})||$; conversely, the **minimum rate of change** of f at $\mathbf{a}$ is $-||\nabla f(\mathbf{a})||$.

2.5 Optimization

Optimization problems arise in many areas of sciences and mathematics.

1. In **regression analysis**, we minimize a "cost function" in order to find the parameters that best fit the available data (see Chapter 8);
2. in **machine learning**, we use algorithms to adjust the learning parameters, again by minimizing a cost function (see Chapters 19, 20, 21, and **??**);
3. in **general relativity**, objects move along **geodesics**, which are the trajectories of minimal length, and
4. in **geometry**, the shortest path joining two points on a sphere is the great circle passing through the points.[24]

Let $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$. The goal of **optimization** is to find where f reaches its **maximum** and **minimum** values, and to determine these values as well.[25]

24: These are crucial to navigation, especially when it comes to determining the fastest and cheapest air routes between two cities.

25: We provide a more in-depth look at optimization in Chapter 5.

Example In **linear regression**, we construct a linear model, in which a dependent variable (the **response**) is predicted by the independent variables (**predictors**) by means of a linear function.

Consider the case when we have only one independent variable, denoted by x. The goal is to find the linear relation that best determines the value of the response y as a function of x: $y = \beta_0 + \beta_1 x + \varepsilon$, where ε is the **error component** of the model.[26] The regression goal is to determine the **optimal** model parameters β_0 and β_1. But what does optimal mean in this context?

Let $(x_k, y_k), k = 1, \cdots, N$, be the observed/available data. In the **ordinary least squares** framework, the best estimate of the true parameters β_0, β_1 (assuming that the linear model was appropriate in the first place) are the values minimizing the **residual sum of squares**:

$$Q(\beta_0, \beta_1) = \sum_{k=1}^{N} (\beta_0 + \beta_1 x_k - y_k)^2.$$

26: In practice, the relation between x and y is unlikely to be exact, and the error component (which relies of distribution parameters) is part and parcel of the problem. We will discuss this in much more detail in Chapter 8.

In the rest of this section, we will review a few of the standard concepts and methods for solving optimization problems, which come in two flavours:

1. **analytical methods**, which are based on differential calculus – they yield exact solutions, but fail in practice when the underlying model is too complicated,[27] and
2. **numerical methods** which provide approximate solutions when that is the case.[28]

27: See Chapter 5 for more information.

28: See Chapter 4 for more information.

2.5.1 Critical Points

The properties of gradient mentioned above require that the gradient not be zero at the point of interest. But observations where the gradient is zero

are also important. These "equilibrium" points are location candidates for finding function's **extrema** (max/min).

Throughout, let $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ and $\mathbf{a} = (a_1, \ldots, a_n) \in D$. The latter is a **critical point** of f if

$$\nabla f(\mathbf{a}) = \mathbf{0} \quad \text{or} \quad \nabla f(\mathbf{a}) \text{ does not exist.}$$

The latter situation occurs at the cone's apex in Figure 2.7, for instance.

In term of equations, this means that $\mathbf{x} = (x_1, \ldots, x_n) = (x_1, \ldots, x_n) = \mathbf{a}$ is a solution of the system

$$\begin{aligned} f_{x_1}(a_1, \cdots, a_n) &= 0 \\ &\vdots \\ f_{x_n}(a_1, \cdots, a_n) &= 0. \end{aligned}$$

In general situations, it is typically somewhat difficult to find the critical points of a function, for two reasons:

1. the system of equations encoded in $\nabla f = \mathbf{0}$ is often **non-linear**, and so we can not use linear algebra methods to solve it;
2. but even when the system is linear, if the number of variables is large, it may be time consuming to use the Gauss-Jordan algorithm to obtain solution(s).[29]

29: See Chapter 3 for details.

We thus often have to rely on **numerical solvers**: the good news is that most programming languages come with libraries that do the work behind the scenes. But it remains important to have a basic understanding of the underlying mathematics, if we want to make conscientious use of such libraries.

Example Find the critical points of $f(x, y) = \sin(xy)$. Plot the graph and the contour curves of f as a solution.

We start by computing the gradient of f:

$$\nabla f(x, y) = \langle y \cos(xy), x \cos(xy) \rangle.$$

Next, we solve the system $\nabla f = \mathbf{0}$, which consists of the following equations:

$$y \cos(xy) = 0 \quad \text{and} \quad x \cos(xy) = 0.$$

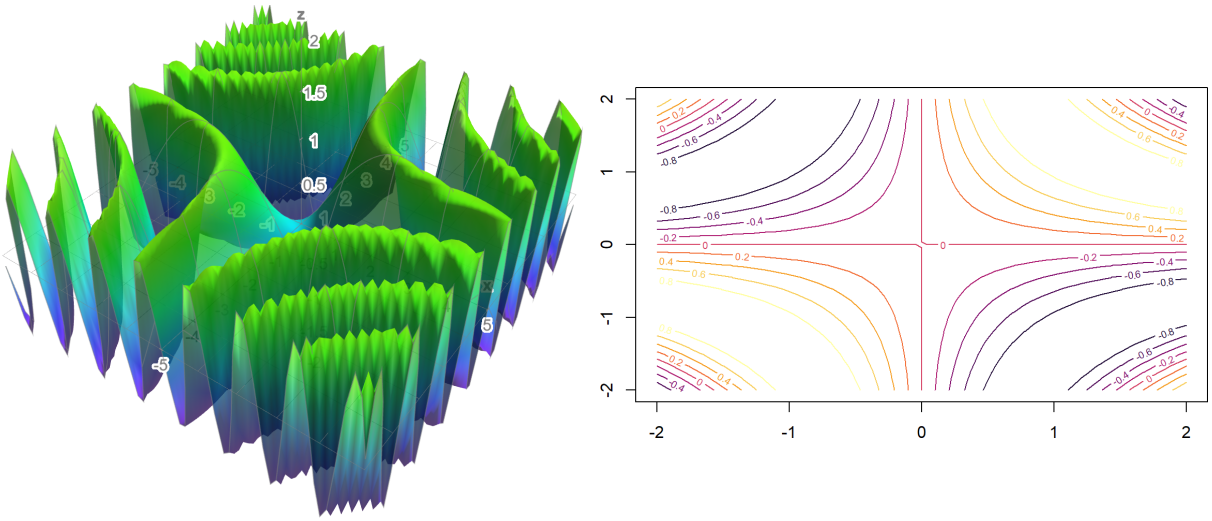
The first of these has two possible solutions: $y = 0$ or $\cos(xy) = 0$.

Substituting $y = 0$ in the second equation yields $x \cos(0) = x = 0$, which implies that $x = 0$ as well. Thus, $P = (0, 0)$ is a critical point of f.

If $\cos(xy) = 0$, then $xy = \frac{\pi}{2} + n\pi$, which automatically satisfies the second equation. We have thus found an infinite collection of critical points of f, namely all the points located along the the **hyperbolas** $xy = \frac{\pi}{2} + n\pi$. If we let $xy = t$, we see in fact that the graph of f looks like a "distorted cosine wave" drawn along each hyperbola.

```
# graph
library(plot3D)
M <- mesh(seq(-2, 2, length.out = 50),
          seq(-2, 2, length.out = 50))
u <- M$x ; v <- M$y
x <- u
y <- v
z <- sin(x*y)
surf3D(x, y, z, colvar = z, colkey = TRUE,
       box = TRUE, bty = "b", phi = 20, theta = 120)

# contour lines
x <- seq(-2,2,length.out=50)
y <- seq(-2,2,length.out=50)
z <- sin(outer(x,y,"*"))
cols <- hcl.colors(10, "Inferno") #color palette
contour(x,y,z, col=cols)
```



2.5.2 Local vs. Global

The extreme values of a function $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ fall into two main categories: **local** and **global**. In general, a **local** property is a property that is satisfied (detected) on a small subregion of the domain D; a **global** property is one that is satisfied everywhere in the domain.

Thus local extrema are extreme values in a sub-region of the domain D, global extrema are extreme values along the entire domain.

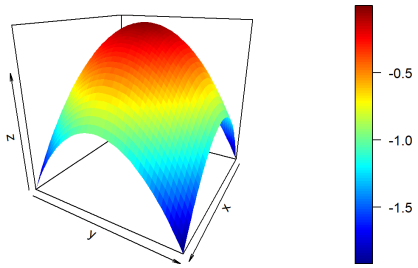
2.5.3 Local Extrema

We now discuss how to find the local extrema of multivariate functions using differential calculus.[30] Locally, the 3 standard shapes that we encounter at a critical point $\mathbf{x} = \mathbf{a} \in D$ where $\nabla f(\mathbf{a}) = \mathbf{0}$ resemble the following.

1. **Local maximum**

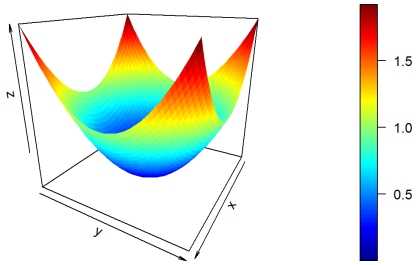
30: In order to keep things simple from a geometrical perspective, we will restrict our efforts to function f of two variables, but the concepts generalize to higher n. In this case, the graph is the surface $z = f(x, y)$, which can be interpreted as a hilly region over the domain D of f.

```
M <- mesh(seq(-1, 1, length.out = 50),
          seq(-1, 1, length.out = 50))
u <- M$x ; v <- M$y
x <- u
y <- v
z <- -x**2-y**2
surf3D(x, y, z, colvar = z, colkey = TRUE,
       box = TRUE, bty = "b", phi = 20, theta = 120)
```



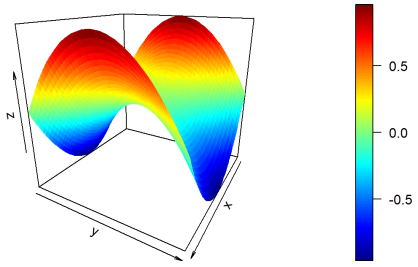
2. **Local mimimum**

```
z <- x**2+y**2
surf3D(x, y, z, colvar = z, colkey = TRUE,
       box = TRUE, bty = "b", phi = 20, theta = 120)
```



3. **Saddle point** ("hybrid": max on one direction, min on the other one)

```
z <- x**2-y**2
surf3D(x, y, z, colvar = z, colkey = TRUE,
       box = TRUE, bty = "b", phi = 20, theta = 120)
```



Definitions

We say that f has a **local minimum** at $\mathbf{a} = (a_1, \cdots, a_n)$ if $f(\mathbf{a}) \leq f(\mathbf{x})$ for all $\mathbf{x}$ in a small n-dimensional region of D centered at $\mathbf{a}$. In contrast, f has

a **local maximum** at $\mathbf{a}$) if $f(\mathbf{a}) \geq f(\mathbf{x})$ for all $\mathbf{x}$ in a small n-dimensional region of D centered at $\mathbf{a}$.

Critical Points and Local Extrema

It is the following result (presented without proof) that justifies the importance of critical points in the optimization context.

Theorem If f has a local extremum at $\mathbf{x} = \mathbf{a}$, then $\mathbf{x} = \mathbf{a}$ is a critical point of f.

The only candidates for **local** extrema are thus critical points.[31] The first step in the search of local extrema therefore consists in solving the system $\nabla f = \mathbf{0}$.

31: That is not necessarily the case for

Once that is done, we need to determine which critical points are local maxima and which are local minima. Thankfully, the **second derivative test** of introductory calculus can be generalized to any finite dimension n, as we shall see shortly.

The Hessian Matrix

We have already introduced the gradient of $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$, a vector field which provides **first-order** information about f. Second derivatives are collected into the **Hessian** matrix:

$$H(f)(\mathbf{x}) = \begin{bmatrix} f_{x_1x_1}(\mathbf{x}) & \cdots & f_{x_1x_n}(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ f_{x_nx_1}(\mathbf{x}) & \cdots & f_{x_nx_n}(\mathbf{x}). \end{bmatrix}$$

The Hessian matrix is symmetric (according to Clairaut's Theorem): a linear algebra result states that real symmetric matrix have real **eigenvalues**.[32]

32: We will discuss these notions in detail in Chapter 3.

Each eigenvalue λ of $H(f)(\mathbf{a})$ is associated to an **eigenvector** $\mathbf{v} \in \mathbb{R}^n$; the sign of the eigenvalue provides information about the local behaviour of f at $\mathbf{x} = \mathbf{a}$, along the direction determined by $\mathbf{v}$.

Second Derivative Test

Suppose $\mathbf{a} \in D$ is a critical point of f and let

$$H(f)(\mathbf{a}) = \begin{bmatrix} f_{x_1x_1}(\mathbf{a}) & \cdots & f_{x_1x_n}(\mathbf{a}) \\ \vdots & \ddots & \vdots \\ f_{x_nx_1}(\mathbf{a}) & \cdots & f_{x_nx_n}(\mathbf{a}). \end{bmatrix}$$

be the Hessian matrix of f at $\mathbf{a}$. If **all** eigenvalues of $H(f)(\mathbf{a})$ are **negative**, then f has a **local maximum** at $\mathbf{x} = \mathbf{a}$; if **all** eigenvalues of $H(f)(\mathbf{a})$ are **positive**, then f has a **local minimum** at $\mathbf{x} = \mathbf{a}$; if some are positive and some are negative, then f has a **saddle point** at $\mathbf{x} = \mathbf{a}$.[33]

33: What happens if some of the eigenvalues are 0?

If $f : D \subseteq \mathbb{R} \to \mathbb{R}$, this is simply the second derivative test in $\mathbb{R}$: let a be a critical point of f with $f'(a) = 0$:

- if $f''(a) < 0$, then f has a local maximum at $x = a$;
- if $f''(a) > 0$, then f has a local maximum at $x = a$, and
- if $f''(a) = 0$, we can not use the second derivative to determine the nature of the critical point.[34]

34: It may be a local maximum (such as $a = 0$ for $f(x) = -x^4$), a local minimum (such as $a = 0$ for $f(x) = x^4$, or an inflection point (such as $a = 0$ for $f(x) = x^3$. Which it is depends on the function in question.

Example Find and classify the critical points of the function $f : \mathbb{R}^3 \to \mathbb{R}$ defined by $f(x, y, z) = x^2 + y^2 + xz$.

We start by computing the gradient of f:

$$\nabla f(x, y, z) = \langle 2x + z, 2y, x \rangle.$$

The system $\nabla f = \mathbf{0}$ has a unique solution, $x = y = z = 0$; the only critical point of f is thus located at $\mathbf{0} = (0, 0, 0)$.

The Hessian matrix $H(f)(\mathbf{x})$ is constant since f was quadratic. In particular,

$$H(f)(\mathbf{0}) = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 2 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

We can compute the eigenvalues and the corresponding eigenvectors of $H(f)(\mathbf{0})$ **algebraically** (see Chapter 3), but we can also solve the eigenvalue/eigenvectors problem numerically with two lines of code in R:

```
H = matrix(c(2, 0, 1, 0, 2, 0, 1, 0, 0), 3, 3)
print(eigen(H))
```

$$\lambda_1 = 2.4 \quad \mathbf{v}_1 = \langle 0.9, 0, 0.4 \rangle$$
$$\lambda_2 = 2 \quad \mathbf{v}_2 = \langle 0, -1, 0 \rangle$$
$$\lambda_3 = -0.4 \quad \mathbf{v}_3 = \langle 0.4, 0, -0.9 \rangle$$

Two of the eigenvalues are positive, the other one is negative; the critical point $\mathbf{0} = (0, 0, 0)$ is a saddle point of f.

Geometrically, along the plane spanned by the vectors $\mathbf{v}_1$ and $\mathbf{v}_2$,[35] which corresponds to the positive eigenvalues λ_1 and λ_2 of $H(f)(\mathbf{0})$, f behaves like a function of two variables with a **local minimum**; along the line spanned by the vector $\mathbf{v}_3$ associated with the negative eigenvalue λ_3, f behaves like a function of one variable with a **local maximum**.

35: These concepts are discussed in Chapter 3.

2.5.4 Global Extrema

When we attempt of minimizing the cost function in a machine learning algorithm, we hope to find the **smallest possible cost**, which will correspond to the parameters associated with the "best learning". In mathematical terms we are looking for the **global minimum** of the cost function, which does not necessarily occur at a local minimum – indeed,

it is conceivable that the global minimum is reached on the boundary of the domain.

In other types of problems, it could be the **global maximum** that is of interest.

Definitions

Let $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$. We say that f reaches its **global minimum** at $\mathbf{a} \in D$ if $f(\mathbf{a}) \leq f(\mathbf{x})$ for all $\mathbf{x} \in D$; the value $f(\mathbf{a})$ is the global minimum value of f. For the **global maximum**, we replace "$\leq$" by "$\geq$".

Note that global extrema do not necessarily exist: $f : (0, \infty) \to \mathbb{R}$, $x \mapsto \frac{1}{x}$ has neither a global maximum nor a global minimum.

Closed and Bounded Domains

A subset $D \subseteq \mathbb{R}^n$ is **bounded** if it can be contained in an $n-$ball of finite radius; formally, it there exists $M > 0$ such that

$$\|\mathbf{x}\|_2 = \sqrt{x_1^2 + \cdots + x_n^2} \leq M$$

for all $\mathbf{x} \in D$.

It is **closed** if it contains it boundary. This is perhaps more difficult to grasp than it looks. An alternative definition (in $\mathbb{R}^n$) is that D is closed if every $\mathbf{x} \notin D$ is contained in an $n-$ball centered at $\mathbf{x}$ which lies entirely outside of D.

Example The disk $D \subseteq \mathbb{R}^n$ defined by the inequality $x^2 + y^2 < 1$ is a bounded domain (use $M = 1$, but it not closed – its boundary, which consists of the circle $x^2 + y^2 = 1$, is not contained in D. The **closure** of D is $x^2 + y^2 \leq 1$.

Extreme Value Theorem

If $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ is **continuous** (roughly speaking, if it has no jump or break) over a closed and bounded domain, then f admits a global maximum and a global minimum on D.

The EVT is not useful from a computational point of view, but it gives some conditions that guarantee that the problems of searching for global extrema makes sense.

Example Let D be the open disk as in the previous example, and denote its closure by $\overline{D}$. Consider the function $f(x, y) = x^2 + y^2$ on D: the globabl minimum of f is 0, clearly attained at $x = y = 0$. However there is no global maximum, since the maximum value is "pushed" to the boundary circle, which is not part of the domain.

If we take the same function but extend it to the closed domain $\overline{D}$, then f does reach its maximum value of 1, at infinitely many points along the boundary circle.

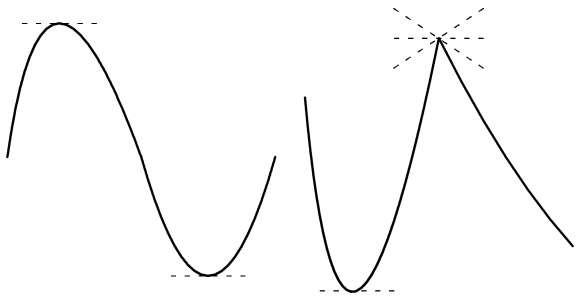


Figure 2.11: Critical points for continuous functions of a single real variable.

2.5.5 Lagrange Multipliers

We have already discussed the link between optimization and the derivative when it comes to finding local extrema. Is there a link for global optimization?

Recall that a differentiable function $f : [a, b] \to \mathbb{R}$ has a **critical point** at $x^* \in (a, b)$ if either $f'(x^*) = 0$ or $f'(x^*)$ is undefined (see Figure 2.11).

If additionally f is continuous, then the optimal solution of the problem

$$\begin{array}{rl} \max & f(x) \\ \text{s.t.} & x \leq b \\ & x \geq a \\ & x \in \mathbb{R} \end{array}$$

is found at one (or possibly, many) of the following **feasible solutions**: $x = a$, $x = b$, or $x = x^*$ where x^* is a critical point of f in (a, b).

This can be extended fairly easily to multi-dimensional domains, with the following result.

Theorem Let $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$ be a continuous function, where A is a closed subset of $\mathbb{R}^n$. Then f reaches its maximum (resp.minimum) value either at a critical point of f in A°, the **interior** of A, or somewhere on ∂A, the **boundary** of A.

Example Consider a company that sells gadgets and gizmos. If the company's monthly profits are expressed (in 1000$ dollars) according to

$$f(x, y) = 81 + 16xy - x^4 - y^4,$$

where x and y represent, respectively, the number of gadgets and gizmos sold monthly (in 10,000s of units), and if the company can produce up to 30,000 units of both gadgets and gizmos monthly, what is the optimal number of each items that the company must sell in order to maximize its monthly profits? The monthly profit function is shown in Figure 2.12.

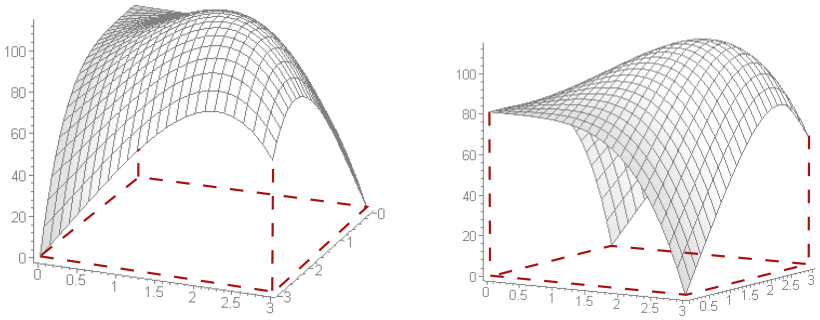


Figure 2.12: Monthly profit function for the gadgets and gizmos example.

Since f is continuous, the maximum value is reached at a critical value in

$$A^\circ = (0,3) \times (0,3)$$

or somewhere on the boundary

$$\partial A = \{(x,y) \in [0,3]^2 \mid x = 0 \text{ or } x = 3 \text{ or } y = 0 \text{ or } y = 3\}.$$

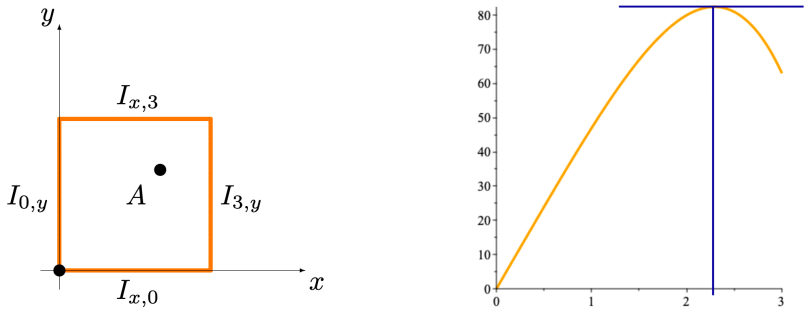


Figure 2.13: Boundary of the domain (left); profile for g_3 and h_3 (right) in the gadgets and gizmos example.

But f is smooth; the gradient $\nabla f(x,y)$ is thus always defined, and the only critical points are those for which $\nabla f(x,y) = (16y - 4x^3, 16x - 4y^3) = (0,0)$. At such a point, $4x = y^3$, which, upon substitution in f_x yields

$$0 = 16y - \frac{1}{16}y^9 = \frac{1}{16}y(256 - y^8) = \frac{1}{16}y(y-2)(y+2)(y^2+4)(y^4+16),$$

which is to say $y = -2, 0, 2$.

Only $y = 2$ can potentially yield a critical point in A°, however. When $y = 2$, we must have $x = \frac{1}{4}2^3 = 2$: the only critical point of f in A° is thus $(x^*, y^*) = (2,2)$, and the monthly profit function value at that point is

$$f(x^*, y^*) = 81 + 16(2)(2) - 2^4 - 2^4 = 113.$$

On the boundary ∂A, the objective function reduces to one of:

$$\begin{aligned}
f(0,y) &= g_0(y) = 81 - y^4, \quad \text{on } 0 \le y \le 3 \\
f(3,y) &= g_3(y) = 48y - y^4, \quad \text{on } 0 \le y \le 3 \\
f(x,0) &= h_0(x) = 81 - x^4, \quad \text{on } 0 \le x \le 3 \\
f(x,3) &= h_3(x) = 48x - x^4, \quad \text{on } 0 \le x \le 3
\end{aligned}$$

These are easy to optimize, being continuous functions of a single real variable; g_0 and h_0 are maximized at the origin, with the objective

function taking the value 81 there, while g_3 and h_3 are maximized at $12^{1/3}$, with the objective function taking the value ≈ 82.42 there (see Figure 2.13).

Combining all this information, we conclude that the company will maximize its monthly profits at 113,000$ if it sells 20,000 units of both gadgets and gizmos.

While the approach we just presented works in this case, there are many instances for which it can be substantially more difficult to find the optimal value on ∂A.

The method of **Lagrange multipliers** can simplify the computations, to some extent. Consider the problem

$$\left|\begin{array}{ll} \min/\max & f(\mathbf{x}) \\ \text{s.t.} & g_i(\mathbf{x}) \leq a_i \quad i = 1, \dots, m \\ & \mathbf{x} \in \mathcal{D}, \end{array}\right.$$

where f, g_i are continuous and differentiable on the (closed) region A described by the constraints $g_i \leq a_i$, $i = 1, \dots, m$.[36] If the problem is **feasible** and **bounded**,[37] then the optimal value is reached either at a critical point of f in A° or at a point $\mathbf{x} \in \partial A$ for which

$$\nabla f(\mathbf{x}) = \lambda_1 \nabla g_1(\mathbf{x}) + \cdots + \lambda_m \nabla g_m(\mathbf{x}),$$

where $\lambda_1, \dots, \lambda_m \in \mathbb{R}$ are the **Lagrange multipliers** of the problem.

36: Strictly speaking, differentiability is not required on the entirety of A.

37: See Chapter 5.

Example Consider a factory that produces various types of deluxe pickle jars. The monthly number of jars Q of a specific kind of pickled radish that can be produced at the factory is given by $Q(K, L) = 900K^{0.6}L^{0.4}$, where K is the number of dedicated canning machines, and L is the monthly number of employee-hours spent on the pickled radish.

The pay rate for the employees is 100$/hour (the pickles are extra deluxe, apparently); the monthly maintenance cost for each canning machine is 200$.

If the factory owners want to maintain monthly production at 36,000 jars of pickled radish, what combination of number of canning machines and employee-hour will minimize the total production costs? The optimization problem is

$$\left|\begin{array}{ll} \min & f(K, L) = 200K + 100L \\ \text{s.t.} & K^{0.6}L^{0.4} = 40; \qquad K, L \geq 0. \end{array}\right.$$

The **objective function** is linear and so has no critical point. The feasability region A can be described by the constraints $g_1(K, L) = K^{0.6}L^{0.4} \leq 40$ and $g_2(K, L) = -K^{0.6}L^{0.4} \leq -40$. Points of interest on the boundary ∂A are obtained by solving the Lagrange equation

$$(200, 100) = \lambda \left(0.6 \left(\tfrac{L}{K}\right)^{0.4}, 0.4 \left(\tfrac{K}{L}\right)^{0.6}\right),$$

since $\nabla g_1 = -\nabla g_2$, with $K^{0.6}L^{0.4} = 40$.

Numerically, there is only one solution, namely

$$(K_*, L_*, \lambda) \approx (35.65, 47.54, 297.10).$$

The objective function at that point takes on the value

$$f(K_*, L_*) \approx 200(35.65) + 100(47.54) \approx 11884.02,$$

and this value must either be the maximum or the minimum of the objective function subject to the constraints of the problem. But we know, that the point $(K_1, L_1) = (1, 40^{2.5})$ belongs to ∂A;[38] since

$$f(K_1, L_1) = 200(1) + 100(40^{2.5}) > f(K_*, L_*),$$

then (K_*, L_*) is indeed the minimal solution of the problem, and the minimal value of the objective function subject to the constraints is $\approx 11,884.02\$$.

38: As $1^{0.6}(40^{2.5})^{0.4} = 40$).

In practice, the value for K has to be an integer,[39] , so we might pick:

- a **sub-optimal** $K'_* = 36$ canning machines, which yields
- a **sub-optimal** $L'_* \approx 46.84$ employee-hours,
- which together yield a **sub-optimal** monthly operating cost of

$$f(K'_*, L'_*) \approx 200(36) + 100(46.84) \approx 11884.85.$$

This departure from optimality would nevertheless be quite likely to be acceptable to the factory owners.

39: Unless we consider using a different number of canning machines at various times during the month.

Given how straightforward the method is, it might seem that there is no real need to say anything else – why would anybody ever use something other than Lagrange multipliers to solve optimization problems?

One of the issues is that when the number of constraints is too high relative to the dimension n of A,[40] then **there may not be a finite number of candidate** solutions on ∂A, which makes this approach useless.

40: Which is usually the case in real-life situations.

Another difficulty that might arise is that the system of equations

$$\nabla f(\mathbf{x}) = \lambda_1 \nabla g_1(\mathbf{x}) + \cdots + \lambda_m \nabla g_m(\mathbf{x})$$

could be **ill-conditioned**, or **highly non-linear**, and numerical solutions could be hard to obtain. We will discuss this further in Chapters 4 and 5.

2.6 Riemann Integrals

Integration, as we will see, is the reverse process of differentiation. We start with a review of basic integration rules and methods, starting with one-variable methods which can then be generalized to multiple Riemann integrals in many variables.

2.6.1 Motivation: Local Densities vs. Total Quantities

The following argument, motivated by statistics, is one of many possible ways of introducing the concept of Riemann integrals.

In general, the (multi-variable) **Riemann integral**

$$\int_D f(x_1, \cdots, x_n) dV$$

is the continuous version of the infinite series

$$\sum_{k_1, \cdots, k_n = 1}^{\infty} f_{k_1, \cdots, k_n} \Delta V.$$

This realization is at the centre of all approaches to Riemann integration.

Consider a real random variable x with **probability density function** $f(x)$.[41] Let x_0 be an arbitrary value of x. The **probability** that x takes a value in the interval $[x_0, x_0 + \Delta x]$ of length (size) Δx (which is usually quite small) is approximately

41: See Chapter 6 for details.

$$f(x_0)\Delta x.$$

Assume that $[a, b]$ is a finite interval. We compute the probability that x belongs to the (large) interval $[a, b]$ by using **Riemann sums approximations**.

First, we sub-divide the interval $[a, b]$ into N **sub-intervals** of equal length $\Delta x = \frac{b-a}{N}$: if we label the endpoints of each sub-interval as

$$x_0 = a, x_1 = x_0 + \Delta x, \cdots, x_{N-1} = x_0 + (N-1)\Delta x, x_N = b,$$

then the sub-interval I_k can be written as

$$I_k = [x_{k-1}, x_k].$$

If Δx is sufficiently small, then we can say that, since the probability of finding x within I_k is approximately $f(x_{k-1})\Delta x$, then the probability of finding x in $[a, b]$ is approximated by the sum of those "local" (**infinitesimal**) probabilities:

$$P(x \in [a, b]) \approx \sum_{k=1}^{N} f(x_{k-1})\Delta x.$$

At this point, we may be nonplussed to realize that this formula is only going to yield an **estimate** (or an **approximation**) of the exact value of the probability.

But the theory of Riemann integrals shows that as we increase the number N of sub-intervals I_k,[42] the estimated value **converges** (gets closer and closer) to the exact value, and in the limiting case $N \to \infty$, we obtain

42: And therefore sending $\Delta x \to 0$.

$$P(x \in [a, b]) = \lim_{N \to \infty} \sum_{k=1}^{N} f(x_{k-1})\Delta x.$$

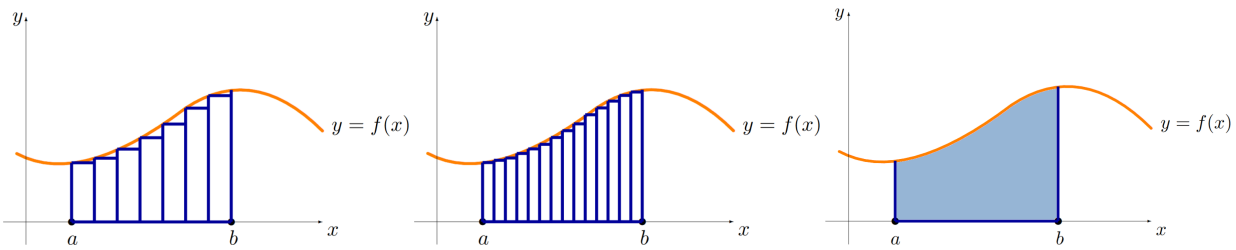


Figure 2.14: Graphical illustration of the Riemann integral $\int_a^b f(x)dx$: approximations with left-most sample points and $N = 7$ (left); $N = 14$ (middle); Riemann integral (right).

2.6.2 One Variable

Using the same reasoning, we define the Riemann integral for any continuous function $f : [a, b] \to \mathbb{R}$ by

$$\int_a^b f(x)\, dx = \lim_{N\to\infty} \sum_{k=1}^{N} f(x_{k-1})\Delta x,$$

where N is the number of sub-interval I_k of length $\Delta x = (b - a)/N$ and x_k is a sample point in I_k (the centre of the interval, say).

Different choices of sample points lead to **different versions** of the Riemann sum approximation. In the limiting case $N \to \infty$, however, all approximations converge to the same value, which is the {**Riemann integral of** f **over** $[a, b]$; the process is illustrated in Figure 2.14.

2.6.3 Fundamental Theorem of Calculus

As is the case with derivatives, the calculation of Riemann integrals can (in principle) be performed without going through the process of Riemann sum approximations.

For a continuous function $f : [a, b] \to \mathbb{R}$, there is a function $F : [a, b] \to \mathbb{R}$ (the **antiderivative** or **indefinite integral** of f), which satisfies $F'(x) = f(x)$ and which we denote by

$$F(x) = \int f(x)dx,$$

The antiderivative is **unique** up to an additive constant c:

$$(F(x) + c)' = f(x).$$

The **Fundamental Theorem of Calculus** states that, for any antiderivative F of f, then

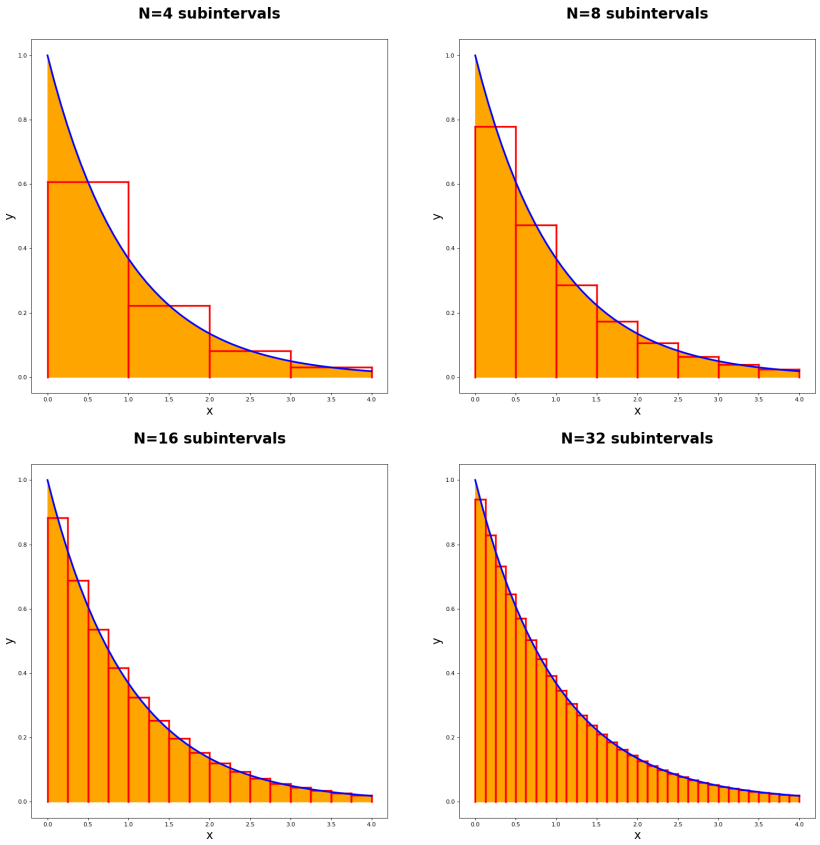
$$\int_a^b f(x)dx = F(b) - F(a) = \Big[F(x)\Big]_a^b.$$

Note that we also denote the difference $F(b) - F(a)$ by $\Big[F(x)\Big]_a^b$.

Example Here are the Riemann sum approximations with 4 different sub-interval sub-divisions, for the integral

$$\int_0^4 e^{-x}dx = \Big[-e^{-x}\Big]_0^4 = 1 - e^{-4}.$$

For any of the approximations, the area of each vertical rectangle is $f(\overline{x})\Delta x$, where $\overline{x}$ is the midpoint of the small interval.



The antiderivative F of a continuous function f always exists. However, if the analytic expression of the function is too complicated, it may not be possible to find the antiderivative F of f.[43] What to do, then? We have no choice but to proceed with numerical integration.[44]

2.6.4 Finding Antiderivatives

Computing derivatives is usually easy, since it is (almost) a one-directional, no-choice algorithm: **follow the rules** and all is good to go.

When we find an antiderivative, we are "climbing back" to the source, and that can actually be much harder.[45]

Here are some basic rules for finding antiderivatives. For more advanced techniques, we let the reader look into the literature [24, 26].

43: It still exists, however.

44: There are several approaches used to compute a Riemann integrals numerically. In the previous example, we used the midpoint approximation; there are other ways of approximating the integral (left-most point, right-most point, Simpson rule, Gaussian quadratures, Monte Carlo, etc.). We will discuss these in Chapter 4.

45: There are methods, but typically harder to use or understand: how do we select the right u-substitution? Or the $u\,dv$ term in integration by parts?

1. **Linearity:**

$$\int (af(x) + bg(x))\,dx = a\int f(x)\,dx + b\int g(x)\,dx$$

2. **Power rule:**

$$\int x^n\,dx = \frac{x^{n+1}}{n+1} + C, \text{ for } n \neq -1$$

3. **Power rule special case:**

$$\int \frac{dx}{x} = \ln|x| + C$$

4. **Exponentials:**

$$\int e^{ax}\,dx = \frac{e^{ax}}{a} + C$$

5. **Integration by parts:**

$$\int f'(x)g(x)\,dx = f(x)g(x) - \int f(x)g'(x)\,dx$$

6. **Integration by substitution:**

$$\int f(x)\,dx = \int f(x(u))\frac{dx}{du}\,du$$

Note that integration by substitution is a sort of inverse of the chain rule, and integration by parts the same for the product rule.

2.6.5 Several Variables

We are now ready to introduce **multiple integrals**, that is Riemann integrals of a function defined over a domain of arbitrary dimension.

Let $D \subset \mathbb{R}^n$ and f be a **density function** on D, such as a **probability density function** for the configuration $(x_1, \cdots, x_n)$ of n random variables. Let $\mathbf{a} \in D$. If we pick a point $\mathbf{x}$ at random the probability that we find it in a region centered at $\mathbf{x} = \mathbf{a}$ of $n-$volume ΔV is approximated by

$$f(\mathbf{a})\Delta x_1 \cdots \Delta x_n.$$

Let $S \subset D$ be a subregion of the whole **sample space domain** D. The probability $p(S)$ of finding $\mathbf{x} \in S$ is approximated as follows. Subdivide S into N small sample regions S_k ($k = 1, \cdots, N$), each of volume ΔV. Pick, for each k, a sample point P_k in S_k. According to the formula above,

we have

$$p(\mathbf{x} \in S) \approx \sum_{k=1}^{N} f(P_k)\Delta V$$

The exact value is obtained in the limiting case $N \to \infty$. This is the multivariate **Riemann integral** construction. If we use **Cartesian coordinates** $(x_1, \cdots, x_n)$, the volume is

$$\Delta V = \Delta x_1 \cdots \Delta x_n,$$

and so

$$p(\mathbf{x} \in S) = \int_S f(S)\, dV = \lim_{N\to\infty} \left(\sum_{k=1}^{N} f(P_k)\Delta x_1 \cdots \Delta x_n \right).$$

The Riemann sum approximation is used to define the Riemann integral for an arbitrary **continuous** function, not necessarily one carrying the meaning of probability.

The **double integral** ($n = 2$ variables) is often denoted by $\iint$, the **triple integral** ($n = 3$ variables) by $\iiint$. If the dimension of the integral is not important (for example, if we are interested in general properties of Riemann integrals) we simply use the symbol $\int$.

2.6.6 Applications to Statistics

Let f be a probability density function of n independent continuous random variables, on a domain $D \subset \mathbb{R}^n$. Let $g(x_1, \cdots, x_n)$ be an arbitrary random variable.[46]

46: We can assume that is a continuous function.

The **average value** of g is the integral

$$E\{g\} = \int_D g(x_1, \cdots, x_n) f(x_1, \cdots, x_n)\, dx_1 \cdots dx_n.$$

The **variance** of g is the integral

$$\sigma^2 = E\left\{(g - E\{g\})^2\right\} = \int_D (g(x_1, \cdots, x_n) - E\{g\})^2 f(x_1, \cdots, x_n)\, dx_1 \cdots dx_n.$$

The **standard deviation** of g is the integral

$$\sigma = \sqrt{\int_D (g(x_1, \cdots, x_n) - E\{g\})^2 f(x_1, \cdots, x_n)\, dx_1 \cdots dx_n}.$$

The **covariance** between two random variables g and h is

$$\sigma\{g, h\} = \int_D (g(x_1, \cdots, x_n) - E\{g\})(h(x_1, \cdots, x_n) - E\{h\}) f(x_1, \cdots, x_n)\, dx_1 \cdots dx_n.$$

Computing Riemann Integrals in Several Variables

Several methods can be used to calculate the Riemann integral of a function of several variables. In Cartesian coordinates, we can deduce a formula starting, once again, from the "infinitesimal" point of view.

For simplicity, we can consider a 2D domain $D \subset \mathbb{R}^2$ defined by the inequalities

$$D : a \leq x \leq b, \quad c(x) \leq y \leq d(x).$$

Let $f : D \subseteq \mathbb{R}^2 \to \mathbb{R}$ be continuous. In order to compute the integral

$$\iint_D f(x, y)\, dydx,$$

we can proceed by iterating the integration process, one iteration per variable, as follows.

First, for each value of $x \in [a, b]$, we can integrate $\int f(x, y)\, dy$ along the vertical direction. Since y satisfies the bounds $c(x) \leq y \leq d(x)$ for each $x \in [a, b]$, we start by computing the integral along the vertical strips of width dx:

$$\int_{c(x)}^{d(x)} f(x, y)dy.$$

Next, we integrate the contributions of each individual strip, by integrating over the remaining variable x. We therefore obtain a formula for computing a double integral in Cartesian coordinates, integrating first by vertical strips:

$$\int_D f dA = \int_a^b \left(\int_{c(x)}^{d(x)} f(x, y)dy \right).$$

Note that the role of the variables can be interchanged; refer to [23] for more details.

In general, if a domain $D \subset \mathbb{R}^n$ is described by Cartesian coordinate inequalities $(x_1, \cdots, x_n)$, such as:

$$\begin{aligned}
&a_1 \leq x_1 \leq b_1 \\
&a_2(x_1) \leq x_2 \leq b_2(x_1) \\
&\cdots \\
&a_n(x_1, x_2, \cdots, x_{n-1}) \leq x_n \leq b_n(x_1, x_2, \cdots, x_{n-1})
\end{aligned}$$

then the $n-$ integral of f over D can be computed by the **iterated integral**

$$\int_D f dV = \int_{a_1}^{b_1} \int_{a_2(x_1)}^{b_1(x_1)} \cdots \int_{a_n(x_1,x_2,\cdots,x_{n-1})}^{b_n(x_1,x_2,\cdots,x_{n-1})} f(x_1, x_2, \cdots, x_n)dx_n dx_{n-1} \cdots dx_1.$$

The idea is to integrate one variable per time, using the one-variable rules of integration. As is the case for integration in $\mathbb{R}$, there is a **change of variables** (substitution) formula for integrals in several variables.

We can then derive formulas for double integrals in **polar** coordinates, or triple integrals in **cylindrical** or **spherical** coordinates.[47]

47: Again, refer to [23] for more details.

Let $D \subset \mathbb{R}^n$ and $f : D \to \mathbb{R}$. The Riemann integral of f over D, defined as the limit of Riemann sums, is denoted by

$$\int_D f \, dV.$$

The symbol dV denotes the **infinitesimal $n-$dimensional volume element**, and the **infinitesimal quantity** $f \, dV$ represents the infinitesimal portion of f contained in the infinitesimal region of measure dV. The **total** ("grand sum") is obtained by integrating $f \, dV$ over the full domain.

The expression of the volume element depends of the choice of coordinates. In Cartesian coordinates, the volume is as expressed above:

$$dV = dx_1 \cdots dx_n.$$

Thus, if $f \equiv 1$, $\int_D f \, dV$ represents the $n-$**volume** of D. For other types of coordinate systems, and the corresponding integration formulas, we once again refer to [23].

Example Let E be the **solid** region located above the triangle of the $xy-$plane defined by the inequalities $|x| \leq 1, 0 \leq y \leq 1-x$, and below the surface $z = x^2+y^2$. Compute the triple integral of $f(x, y, z) = x$ over E.

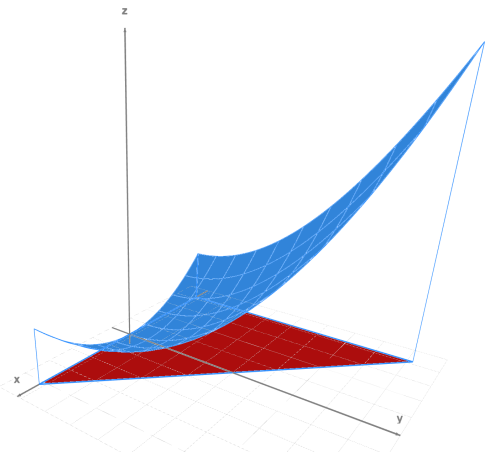
The bounds of the triangle define the region of the $xy-$plane:

$$-1 \leq x \leq 1, \quad 0 \leq 1 \leq 1 - x.$$

The solid is therefore described by the inequalities

$$-1 \leq x \leq 1, \quad 0 \leq y \leq 1 - x, \quad 0 \leq z \leq x^2 + y^2,$$

as shown below.



Therefore, the triple integral is:

$$\begin{aligned}
\iiint_E f\,dV &= \int_{-1}^{1}\int_{0}^{1-x}\int_{0}^{x^2+y^2} x\,dz\,dy\,dx \\
&= \int_{-1}^{1}\int_{0}^{1-x}\left[xz\right]_{z=0}^{z=x^2+y^2} = \int_{-1}^{1}\int_{0}^{1-x}(x^3+xy^2)\,dy\,dx \\
&= \int_{-1}^{1}\left[x^3y + x\frac{y^3}{3}\right]_{y=0}^{y=1-x} dx = \int_{-1}^{1}\left(x^3(1-x) + x\frac{(1-x)^3}{3}\right) dx \\
&= \int_{-1}^{1}\left(\frac{-4x^4}{3} + 2x^3 - x^2 + \frac{x}{3}\right) dx \\
&= \left[-\frac{4x^5}{15} + \frac{2x^4}{4} - \frac{x^3}{3} + \frac{x^2}{6}\right]_{-1}^{1} = -\frac{6}{5}.
\end{aligned}$$

2.7 Exercises

1. The price at which an item sells is given by $P(d,s) = k\frac{d^2}{s+10}$, where k is a constant, and s and d are the product supply and demand, respectively.
   a) For what value(s) of d is $P(d,90) = 100k$?
   b) For what value(s) of s is $P(10,s) = 10k$?
   c) If $d = 9$ and $s = 10$, how does P change when d goes from 9 to 11?
   d) If $d = 9$ and $s = 10$, how does P change when s goes from 10 to 8?
   e) Compute and interpret $P(6,3)$.
   f) Compute and interpret $P_d(6,3)$.
   g) Compute and interpret $P_s(6,3)$.
2. Find the largest possible domain (in $\mathbb{R}^2$) and the range (in $\mathbb{R}$) of the following functions.
   a) $f(x,y) = x^2 + 2xy + y^2$.
   b) $f(x,y) = \ln(x-y)$.
   c) $f(x,y) = \frac{1}{(y-2)\ln x}$.
   d) $f(x,y,z) = \frac{xy}{1-z}$.
   e) $f(x,y,z) = \sqrt{36 - x^2 - 4y^2}$.
   f) $f(x,y,z) = \frac{x^2z^2}{(y-2)^2}$.
   g) $f(x,y) = \sqrt{x+y}$.
   h) $f(x,y) = \sqrt{4 - x^2 - y^2}$.
   i) $f(x,y) = \frac{1}{4-x^2-y^2}$.
   j) $f(x,y) = \frac{1}{e^{x^2+y^2}}$.
3. Find the equation of the tangent plane to the surface $z = f(x,y)$ at the given point.
   a) $f(x,y) = x^4 + y^4 - 4xy + 1, (0,0)$.
   b) $f(x,y) = x^2 + y^2 + 4x - 6y, (1,0)$.
   c) $f(x,y) = 2x^3 + xy^2 + 5x^2 + y^2, (0,1)$.

d) $f(x,y) = x^2 + y^2 + x^2y + 4$, $(1,2)$.
e) $f(x,y) = y\sqrt{x} - y^2 - x + 6y$, $(1,-1)$.
f) $f(x,y) = xy - 2x - y$, $(2,3)$.
g) $f(x,y) = xy(1-x-y)$, $(-3,2)$.
h) $f(x,y) = x^2 + y^2 + \frac{1}{x^2y^2}$, $(-1,0)$.
i) $f(x,y) = x^3 + y^3 + 4xy$, $(0,-2)$.
j) $f(x,y) = \frac{1}{xy}$, $(1,-1)$.
k) $f(x,y) = \ln(x^2+y^2)$, $(1,0)$.
l) $f(x,y) = x^y$, $(2,2)$.
m) $f(x,y) = (x+y)e^x$, $(0,2)$.
n) $f(x,y) = \frac{x+y}{x-y}$, $(2,-1)$.
o) $f(x,y) = y\ln(x+2)e^{\sqrt{y}}$, $(-1,4)$.
p) $f(x,y) = xye^{1/y}$, $(-1,1)$.

4. Classify the critical points of the following functions.

a) $f(x,y) = x^4 + y^4 - 4xy + 1$.
b) $f(x,y) = x^2 + y^2 + 4x - 6y$.
c) $f(x,y) = 2x^3 + xy^2 + 5x^2 + y^2$.
d) $f(x,y) = x^2 + y^2 + x^2y + 4$.
e) $f(x,y) = y\sqrt{x} - y^2 - x + 6y$.
f) $f(x,y) = xy - 2x - y$.
g) $f(x,y) = xy(1-x-y)$.
h) $f(x,y) = x^2 + y^2 + \frac{1}{x^2y^2}$
i) $f(x,y) = x^3 + y^3 + 4xy$.

5. Compute the 2nd order partial derivatives of the following functions.

a) $f(x,y) = \frac{1}{\sqrt{x^2+y^2}}$.
b) $f(x,y,z) = xyz$.
c) $f(x,y,z) = \ln\left(\frac{x+y}{x+z}\right)$.
d) $f(x,y) = \frac{x^2+y^2}{1+x}$.
e) $f(x,y,z) = \sqrt{1+x+y-2z}$.
f) $f(x,y,z) = x^2yz^3 + xy^2\sqrt{z}$.
g) $f(x,y) = \frac{xy^2}{\sqrt{x+3}}$.
h) $f(x,y,z) = xz\sqrt{y}$.
i) $f(x,y,z) = x^3\ln(zx)yz^2e^{yx}$.
j) $f(x,y) = xy\sqrt{x^2+7}$.
k) $f(x,y,z) = \frac{1}{xyz}$.
l) $f(x,y,z) = \frac{x^3y-z^2}{3x+y+2z}$.

6. Compute $\int_0^2 \int_0^x e^{x^2}\,dy\,dx$ by first sketching the area of integration.
7. Compute $\int_0^3 \int_{y^2}^9 y\sin(x^2)\,dx\,dy$.
8. What is the volume of the solid bounded by the planes $z = x+2y+4$ and $z = 2x+y$, above the triangle in the xy plane with vertices $A(1,0,0)$, $B(2,1,0)$ and $C(0,1,0)$?
9. Compute $\int_W h\,dV$, where $h(x,y,z) = ax+by+cz$, $W = \{(x,y,z) : 0 \le x \le 1, 0 \le y \le 1, 0 \le z \le 2\}$.
10. Sketch the region of integration W of the triple integral $\int_0^1 \int_0^{2-x} \int_0^3 f(x,y,z)\,dz\,dy\,dx$
11. What is the volume of the solid defined by the intersection of the two cylinders $x^2 + z^2 = 1$ and $y^2 + z^2 = 1$?

12. Compute $\int_0^{\sqrt{2}} \int_0^{\sqrt{4-y^2}} xy \, dx \, dy$.
13. Compute $\int_W \sin(x^2 + y^2) \, dV$, where W is the cylinder centered about the z axis from $z = -1$ to $z = 3$ with radius 1.
14. Compute

$$\int_0^1 \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \int_{-\sqrt{1-x^2-z^2}}^{\sqrt{1-x^2-z^2}} (x^2 + y^2 + z^2)^{-1/2} \, dy \, dz \, dx.$$

15. Compute

$$\int_0^1 \int_{-1}^1 \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} (x^2 + y^2)^{-1/2} \, dy \, dx \, dz.$$

16. What is the volume of the solid Q directly above the region bounded by $0 \le x \le 1, 1 \le y \le 2$ in the $xy-$plane and below the plane $z = 4 - x - y$?
17. Compute $\int_0^1 \int_{\sqrt{x}}^1 e^{y^3} \, dy \, dx$.
18. Sketch the solid bounded by the the surfaces $z = 0$, $y = 0$, $z = a - x + y$ and $y = a - \frac{1}{a}x^2$, where a is a positive constant. What is the volume of that solid?
19. Evaluate $\int_0^{\ln 2} \int_0^{\ln 5} e^{2x-y} \, dx \, dy$.
20. Evaluate $\int_0^1 \int_0^1 \frac{xy}{\sqrt{x^2+y^2+1}} \, dx \, dy$.
21. Let $D = \{(x, y) : 1 \le y \le e, y^2 \le x \le y^4\}$. Compute $\iint_D \frac{1}{x} \, dA$.
22. What is the volume of the solid lying under the paraboloid $z = x^2 + y^2$ and above the domain bounded by $y = x^2$ and $x = y^2$?
23. Let R be the disk of radius 5, centered at the origin. Evaluate $\iint_R x \, dA$.
24. What is the volume of the solid lying under the cone $z = \sqrt{x^2 + y^2}$ and above the ring $4 \le x^2 + y^2 \le 25$ located in the $xy-$plane?
25. Evaluate $\int_0^3 \int_0^{\sqrt{9-x^2}} \int_0^x yz \, dy \, dz \, dx$.
26. Compute $\iiint_E e^x \, dV$, where

$$E = \{(x, y, z) : 0 \le y \le 1, 0 \le x \le y, 0 \le z \le x + y\}.$$

27. Compute $\iiint_E xz \, dV$, where E is the pyramid with vertices $(0, 0, 0)$, $(0, 1, 0)$, $(1, 1, 0)$ and $(0, 1, 1)$.
28. Let W be a three-dimensional solid. Its volume can be computed by the following iterated integral:

$$V(W) = \int_0^{2\pi} \int_0^2 \int_0^{4-r^2} r \, dz \, dr \, d\theta.$$

Find W and $V(W)$.
29. Compute $\iiint_B (x^2+y^2+z^2) \, dV$, where B is the unit ball $x^2+y^2+z^2 \le 1$.
30. Evaluate

$$\int_0^3 \int_0^{\sqrt{9-y^2}} \int_{\sqrt{x^2+y^2}}^{\sqrt{18-x^2-y^2}} (x^2 + y^2 + z^2) \, dz \, dx \, dy.$$

31. Evaluate the integral $\iint_D x^2 y \, dx \, dy$ where D is the region bounded by the curves $y = x^2$ and $x = y^2$ in the first quadrant.

32. Compute the volume of the solid bounded by the cone $z = \sqrt{x^2 + y^2}$ and the sphere of radius $a > 0$ whose center is located at the origin.
33. Compute the volume of the solid bounded by the paraboloïds $z = 10 - x^2 - y^2$ and $z = 2(x^2 + y^2 - 1)$.
34. Compute the area of the planar region bounded by $y = x^2$, $y = 2x^2$, $x = y^2$, and $x = 3y^2$.
35. Find the volume of the solid bounded by the interior of the sphere $x^2 + y^2 + z^2 = a^2$ and the interior of the cylinder $x^2 + y^2 = a^2$, $a > 0$.
36. Find the volume of the solid bounded by the interior of each of the cylinders $x^2 + y^2 = a^2$, $x^2 + z^2 = a^2$ and $y^2 + z^2 = a^2$, $a > 0$.
37. Find the volume of the solid bounded by the interior of the cone $z^2 = x^2 + y^2$ lying above the paraboloïd $z = 6 - x^2 - y^2$.
38. Find the volume of the solid bounded by the plane $z = 3x + 4y$ lying below the paraboloïd $z = x^2 + y^2$.
39. Let S be the sphere of radius $a > 0$ centered at $(0, 0, a)$. Show that $\iiint_S z^2 \, dx \, dy \, dz = \frac{8}{5}\pi a^5$.
40. Compute $\iiint_{\mathbb{R}^3} e^{-(x^2+y^2+z^2)} \, dx \, dy \, dz$.