CHAPTER VIII
Calculus for Machine Learning

PHOK Ponna and NEANG Pheak

Institute of Technology of Cambodia
Department of Applied Mathematics and Statistics (AMS)

2022–2023

Contents

Contents

Calculus is the mathematical study of change. The effectiveness of calculus to solve a complicated but continuous problem lies in its ability to slice the problem into infinitely simpler parts, solve them separately, and subsequently rebuild them into the original whole.This strategy can be applied to study all continuous elements that can be sliced in this manner,be it the curvatures of geometric shapes, as well as the trajectory of an object in flight, or atime interval.

Contents

Calculus has been applied in many domains, from Newton's application in solving problems of mathematical physics, to the more recent application of Newton's ideas in the work done at NASA by mathematician, Katherine Johnson, and her colleagues.
In the 1860s, James Clerk Maxwell used calculus to recast the experimental laws of electricity and magnetism, eventually predicting not only the existence of electromagnetic waves, but also revealing the nature of light as an electromagnetic wave. Based on his work,Nikola Tesla created the first radio communication system, Guglielmo Marconi transmitted the first wireless messages, and eventually many modern-day devices, such as the televisionand the smartphone, came into existence.
Albert Einstein, in 1917, also applied calculus to a model of atomic transitions, in order to predict the effect of stimulated emission. His work later led to the first working lasers in the 1960s, which have since then been used in many different devices, such as compact-discplayers and bar code scanners.

“ Without calculus, we wouldn’t have cell phones, computers, or microwave ovens. We wouldn’t have radio. Or television. Or ultrasound for expectant mothers, or GPS for lost travelers. We wouldn’t have split the atom, unraveled the humangenome, or put astronauts on the moon. We might not even have the Declaration of Independence.”

More interestingly is the integral role of calculus in machine learning. It underlies important algorithms, such as **gradient descent**, which requires the computation of the gradient of a function and is often essential to train machine learning models. This makes calculus one of the fundamental mathematical tools in machine learning.

Contents

Definition 1

A **function** is a rule that defines the relationship between a dependent variable and an independent variable.

Examples are all around us: The average daily temperature for your city depends on, and is a function of, the time of year; the distance an object has fallen is a function of how much time has elapsed since you dropped it; the area of a circleis is a function of its radius; and the pressure of an enclosed gas is a function of its temperature.
In machine learning, a neural network learns a function by which it can represent the relationship between features in the input, the independent variable, and the expected output, the dependent variable. We canrepresent this mapping as follows:

$$\text{Output(s)} = \text{function(Input)}$$

More formally, however, a function is often represented by $y = f(x)$, which translates to y is a function of x. This notation specifies x as the independent input variable that we already know, whereas y is the dependent output variable that we wish to find.

Contents

Suppose $f(x)$ is defined when x is near the number a . (This means that is defined on some open interval that contains a , except possibly at a itself.)
Then we write

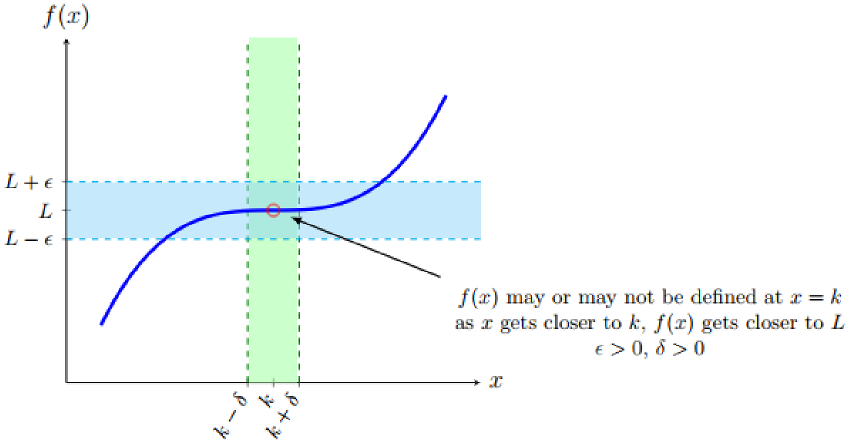
$$\lim_{x \to a} f(x) = L$$

and say "the limit of $f(x)$, as x approaches a, equals L "
if we can make the values of $f(x)$ arbitrarily close to L (as close to L as we like) by taking x to be sufficiently close to a (on either side of a) but not equal to a.

Definition 2

The limit of $f(x)$ is L as x approaches k, if for every $\varepsilon > 0$, there is a positive number $\delta > 0$, such that:

$$\text{if } 0 < |x - k| < \delta \text{ then } |f(x) - L| < \varepsilon$$


$f(x)$
$L + \epsilon$
L
$L - \epsilon$
x
$k - \delta$
k
$k + \delta$
$f(x)$ may or may not be defined at $x = k$
as x gets closer to k, $f(x)$ gets closer to L
$\epsilon > 0$, $\delta > 0$

Definition 3

We write

$$\lim_{x \to a^-} f(x) = L$$

and say the left-hand limit of $f(x)$ as x approaches a [or the limit of $f(x)$ as x approaches a from the left] is equal to L if we can make the values of $f(x)$ arbitrarily close to L by taking x to be sufficiently close to a and x less than a.
Similarly, if we require that x be greater than a , we get "the right-hand limit of $f(x)$ as x approaches a is equal to L " and we write

$$\lim_{x \to a^+} f(x) = L$$

By comparing Definition of limit with the definitions of one-sided limits, we see that the following is true.

$$\lim_{x \to a} f(x) = L \text{ iff } \lim_{x \to a^-} f(x) = L \text{ and } \lim_{x \to a^+} f(x) = L$$

Example 1

Compute $\lim_{x \to -1} \sqrt{x+1}$

```python
from sympy import limit, sqrt, pprint
from sympy.abc import x
expression = sqrt(x+1)
result = limit(expression, x, -1)
print('Limit of')
pprint(expression)
print(' at x=-1 is', result)
```

```
Limit of

  _______
╲╱ x + 1
 at x=-1 is 0
```

Example 2

Compute $\lim_{x\to 0} x^2 \sin\frac{1}{x}$

```
from sympy import limit, sin, pprint
from sympy.abc import x
expression = x**2 * sin(1/x)
result = limit(expression, x, 0)
print('Limit of')
pprint(expression)
print('at x=0 is', result)
```

```
Limit of
 2    ⎛1⎞
x ⋅sin⎜─⎟
      ⎝x⎠
at x=0 is 0
```

Rules for limits

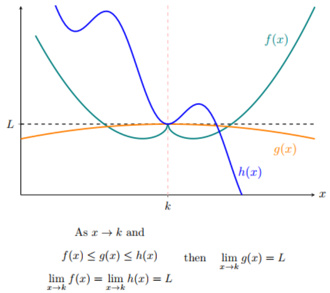
$\lim_{x\to k} f(x) = L \qquad \lim_{x\to k} g(x) = M$			
Constant multiple rule	$\lim_{x\to k}(af(x))$	$= aL$	
Sum rule	$\lim_{x\to k}(f(x)+g(x))$	$= L + M$	
Difference rule	$\lim_{x\to k}(f(x)-g(x))$	$= L - M$	
Product rule	$\lim_{x\to k}(f(x)\cdot g(x))$	$= L \cdot M$	
Quotient rule	$\lim_{x\to k}(\frac{f(x)}{g(x)})$	$= \frac{L}{M},$	$M \neq 0$
Root rule	$\lim_{x\to k}(\sqrt[n]{f(x)})$	$= \sqrt[n]{L},$	$n > 0,$ and if n is even then $L > 0$
Power rule	$\lim_{x\to k}((f(x))^n)$	$= L^n,$	$n > 0$

Theorem 1

The sandwich theorem This theorem is also called the squeeze theorem or the pinching theorem. It states that whenthe following are true:

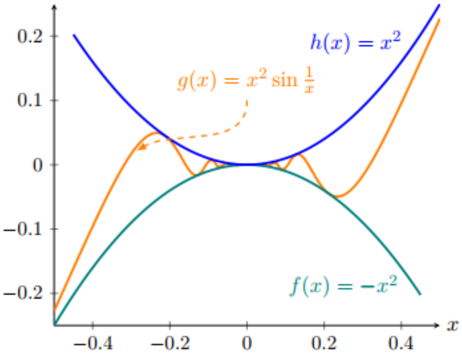
1. x is close to k
2. $f(x) \leq g(x) \leq h(x)$
3. $\lim_{x \to k} f(x) = \lim_{x \to k} h(x) = L$

then $\lim_{x \to k} g(x) = L$.



Example 3

Compute $\lim_{x\to 0} x^2 \sin \frac{1}{x}$.



As the following is true:

$$-1 \leq \sin \frac{1}{x} \leq +1$$

$$-x^2 \leq x^2 \sin \frac{1}{x} \leq +x^2$$

$$\lim_{x\to 0}(-x^2) = \lim_{x\to 0}(+x^2) = 0$$

From sandwich theorem:

$$\lim_{x\to 0} x^2 \sin \frac{1}{x} = 0$$

Definition 4

A function $f(x)$ is continuous at a point a, if the function's value approaches $f(a)$ when x approaches a. Hence to test the continuity of a function at a point $x = a$, check the following:

1. $f(a)$ should exist
2. $f(x)$ has a limit as x approaches a
3. The limit of $f(x)$ as $x \to a$ is equal to $f(a)$

If all of the above hold true, then the function is continuous at the point a.

In very simple words, the derivative of a function $f(x)$ represents its rate of change and is denoted by either $f'(x)$ or df/dx.

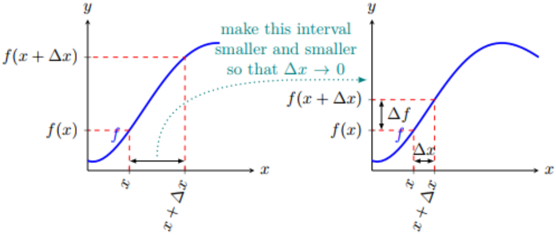
Definition 5

The **derivative of a function** f **at a number** a, denoted by $f'(a)$, is

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

if this limit exists.

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



Example 4

Calculate $f'(x)$ if $f(x) = x^2$.

```
from sympy import diff, sqrt, pprint
from sympy.abc import x
expression = x**2
result = diff(expression, x)
print('Derivative of')
pprint(expression)
print('with respect to x is')
pprint(result)
```

```
Derivative of
 2
x
with respect to x is
2·x
```

Example 5

Calculate $f'(x)$ if $f(x) = x^2, 3x^5, 4x^9$.

```
from sympy import diff, pprint
from sympy.abc import x
expressions = [x**2, 3*x**5, 4*x**9]
for expression in expressions:
    result = diff(expression, x)
    print('Derivative of')
    pprint(expression)
    print('with respect to x is')
    pprint(result)
    print()
```

```
Derivative of
 2
x
with respect to x is
2⋅x

Derivative of
   5
3⋅x
with respect to x is
    4
15⋅x

Derivative of
   9
4⋅x
with respect to x is
    8
36⋅x
```

Contents

Theorem 2

1. $\frac{d}{dx}c = 0$
2. $\frac{d}{dx}x^n = nx^{n-1}$
3. $\frac{d}{dx}\sin x = \cos x$
4. $\frac{d}{dx}\cos x = -\sin x$

Theorem 3

If f and g are differentiable at a and c is a real number, then $cf, f \pm g, f.g$ and $\frac{f}{g}, g(a) \neq 0$ are also differentiable at a and

1. $(cf)'(a) = cf'(a)$
2. $(f \pm g)'(a) = f'(a) \pm g'(a)$
3. $(f.g)'(a) = f'(a).g(a) + f(a).g'(a)$
4. $\left(\frac{f}{g}\right)'(a) = \frac{f'(a).g(a) - f(a).g'(a)}{(g(a))^2}, \qquad g(a) \neq 0$

Theorem 4

1. $\frac{d}{dx}\tan x = \sec^2 x$
2. $\frac{d}{dx}\cot x = -\csc^2 x$
3. $\frac{d}{dx}\sec x = \sec x \tan x$
4. $\frac{d}{dx}\csc x = -\csc x \cot x$

Theorem 5 (Chain Rule)

Let $a \in I, f : I \longrightarrow \mathbb{R}$ and $g : J \longrightarrow \mathbb{R}$ such that $f(I) \subset J$. If f is differentiable at a and g is differentiable at $f(a)$, then $g \circ f$ is also differentiable at a and

$$(g \circ f)'(a) = g'(f(a)).f'(a).$$

Theorem 6

If u is a differentiable function of x and n is a rational number, then

1. $\frac{d}{dx}[u^n] = nu'u^{n-1}$
2. $\frac{d}{dx}\sin u = u'\cos u$
3. $\frac{d}{dx}\cos u = -u'\sin u$
4. $\frac{d}{dx}\tan u = u'\sec^2 u$
5. $\frac{d}{dx}\cot u = -u'\csc^2 u$
6. $\frac{d}{dx}\sec u = u'\sec u\tan u$
7. $\frac{d}{dx}\csc u = -u'\csc u\cot u$

Example 6

Compute the derivative of the following functions

1. $f(x) = \left(3x - 2x^2\right)^3$
2. $f(x) = \frac{-7}{(2x-3)^2}$
3. $f(x) = x^2\sqrt{1-x^2}$
4. $f(x) = \frac{x}{\sqrt[3]{x^2+4}}$
5. $f(x) = \left(\frac{3x-1}{x^2+3}\right)^2$
6. $f(x) = \cos(3x^2)$
7. $f(x) = \sin^3(4x)$
8. $f(x) = \sqrt{\cos x}$

Theorem 7 (Derivative of inverse function)

Let $a \in I, f : I \longrightarrow \mathbb{R}$ be continuous, bijective on I and differentiable at a with $f'(a) \neq 0$. Then, the inverse function of f, say f^{-1}, is differentiable at $f(a)$. Moreover,

$$\left(f^{-1}\right)'(f(a)) = \frac{1}{f'(a)}$$

Contents

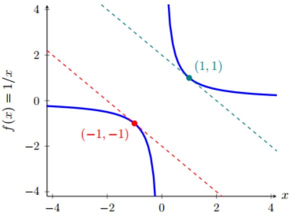
we'll define the tangent to a curve $f(x)$ at a point $A(x_0, f(x_0))$ as a line thatsatisfies two of the following:

1. The line passes through A
2. The slope of the line is equal to the slope of the curve at the point A.

Using the above two facts, we can easily determine the equation of the tangent line at a point $(x_0, f(x_0))$. A few examples are shown next.

Example 7

Let $f(x) = \frac{1}{x}$. Find the slopes and tangent lines to the curve $C: y = f(x)$ at $x = 1$ and $x = -1$.



```python
import numpy as np
def f(x):
    return 1/x
epsilon = np.finfo(np.float32).eps
for x in [1, -1]:
    slope = (f(x+epsilon) - f(x))/epsilon
    y = f(x)
    c = y - slope * x
    print('Slope at x={} is {}'.format(x, slope))
    print('Tangent line is y={:f}x{:+f}'.format(slope,c))
```

```
Slope at x=1 is -0.9999998807907104
Tangent line is y=-1.000000x+2.000000
Slope at x=-1 is -1.0000001192092896
Tangent line is y=-1.000000x-2.000000
```

Contents

Definition 6

A **function of two variables** is a rule that assigns to each ordered pair of real numbers (x, y) in a set D a unique real number denoted by $f(x, y)$. The set D is the **domain** of f and its **range** is the set of values that f takes on, that is, $\{f(x, y)|(x, y) \in D\}$.
We often write $z = f(x, y)$ to make explicit the value taken on by f at the general point (x, y). The variables x and y are **independent variables** and z is the **dependent variable**.

Example 8

For each of the following functions, evaluate and find and sketch the domain.

(a) $f(x, y) = \dfrac{\sqrt{x + y + 1}}{x - 1}$

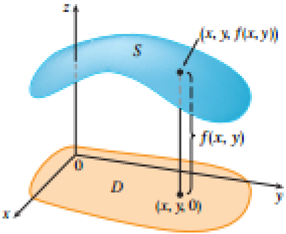
(b) $f(x, y) = x \ln\left(y^2 - x\right)$

Example 9

Find the domain and range of $g(x, y) = \sqrt{9 - x^2 - y^2}$.

Definition 7

If f is a function of two variables with domain D, then the **graph** of f is the set of all points (x, y, z) in $\mathbb{R}^3$ such that $z = f(x, y)$ and (x, y) is in D.



Example 10

Sketch the graph of the function $f(x, y) = 6 - 3x - 2y$.

Example 11

Sketch the graph of the function $g(x, y) = \sqrt{9 - x^2 - y^2}$.

Example 12

Sketch the graph of the function $f(x, y) = 4 - x^2 - y^2$.

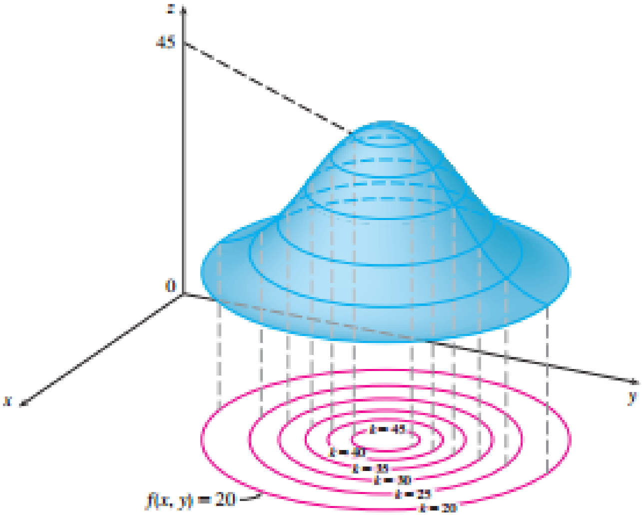
Example 13

Sketch the graph of the function $z = x^2 + y^2$.

Definition 8

The **level curves** of a function f of two variables are the curves with equations $f(x, y) = c$, where c is a constant (in the range of f).

A level curve $f(x, y) = c$ is the set of all points in the domain of f at which f takes on a given value c. In other words, it shows where the graph of f has height c.
You can see from Figure below the relation between level curves and horizontal traces. The level curves $f(x, y) = k$ are just the **traces** of the graph of f in the horizontal plane $z = k$ projected down to the xy-plane.


z
45
0
x
y
k = 45
k = 40
k = 35
k = 30
k = 25
k = 20
f(x, y) = 20

Example 14

Sketch the level curves of the function $f(x, y) = 6 - 3x - 2y$ for the values $c = -6, 0, 6, 12$.

Example 15

Sketch the level curves of the function $f(x, y) = \sqrt{9 - x^2 - y^2}$ for the values $c = 0, 1, 2, 3$.

Example 16

Sketch some level curves of the function $f(x, y) = 4x^2 + y^2 + 1$.

Functions of Three or More Variables

A **function of three variables**, f, is a rule that assigns to each ordered triple (x, y, z) in a domain $D \subset \mathbb{R}^3$ a unique real number denoted by $f(x, y, z)$. For instance, the temperature T at a point on the surface of the earth depends on the longitude x and latitude y of the point and on the time t, so we could write $T = f(x, y, t)$.

A **function of n variables** is a rule that assigns a number $z = f(x_1, x_2, \ldots, x_n)$ to an n-tuple of real numbers. We denote by $\mathbb{R}^n$ the set of all such n-tuples. The function f is a real-valued function whose domain is a subset of $\mathbb{R}^n$. Some times we will use vector notation to write such functions more compactly:

If $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, we often write $f(\mathbf{x})$ in place of $f(x_1, x_2, \ldots, x_n)$.

For example, if a company uses n different ingredients in making a food product, c_i is the cost per unit of the ith ingredient, and x_i units of the ith ingredient are used, then the total cost C of the ingredients is a function of the n variables $x_1, x_2, \ldots, x_n$.

Contents

Definition 9

Let f be a function of two variables whose domain D includes points arbitrarily close to (a, b). Then we say that the **limit of $f(x, y)$ as (x, y) approaches (a, b)** is L and we write

$$\lim_{(x,y)\to(a,b)} f(x, y) = L$$

if for every number $\varepsilon > 0$ there is a corresponding number $\delta > 0$ such that

If $(x, y) \in D$ and $0 < \sqrt{(x-a)^2 + (y-b)^2} < \delta$ then $|f(x, y) - L| < \varepsilon$.

Remark 1

Other notations for the limit in the Definition are

$$\lim_{\substack{x\to a \\ y\to b}} f(x, y) = L \text{ and } f(x, y) \to L \text{ as } (x, y) \to (a, b)$$

Remark 2

If $f(x,y) \to L_1$ as $(x,y) \to (a,b)$ along a path C_1 and $f(x,y) \to L_2$ as $(x,y) \to (a,b)$ along a path C_2, where $L_1 \neq L_2$, then $\lim_{(x,y \to (a,b)} f(x,y)$ does not exist.

Example 17

Show that $\lim_{(x,y \to (0,0)} \dfrac{x^2 - y^2}{x^2 + y^2}$ does not exist.

Example 18

If $f(x,y) = \dfrac{xy}{(x^2 + y^2)}$, does $\lim_{(x,y) \to (0,0)} f(x,y)$ exist?

Example 19

If $f(x,y) = \dfrac{xy^2}{(x^2+y^4)}$, does $\displaystyle\lim_{(x,y)\to(0,0)} f(x,y)$ exist?

Remark 3

The Limit Laws of functions of one variable can be extended to functions of two variables: The limit of a sum is the sum of the limits, the limit of a product is the product of the limits, and so on. The Squeeze Theorem also holds.

Example 20

Find $\displaystyle\lim_{(x,y)\to(0,0)} \frac{3x^2y}{x^2+y^2}$ if it exists.

Definition 10

A function f of two variables is called **continuous** at (a, b) if

$$\lim_{(x,y \to (a,b)} f(x, y) = f(a, b)$$

We say f is continuous on D if f is **continuous** at every point (a, b) in D.

Definition 11

A **polynomial function of two variables** (or polynomial, for short) is a sum of terms of the form $cx^m y^n$, where c is a constant and m and n are nonnegative integers. A **rational function** is a ratio of polynomials.

Remark 4

All polynomials are continuous on $\mathbb{R}^2$. Likewise, any rational function is continuous on its domain because it is a quotient of continuous functions.

Example 21

Where is the function $f(x, y) = \frac{x^2-y^2}{x^2+y^2}$ continuous?

Example 22

Let

$$g(x, y) = \begin{cases} \frac{x^2-y^2}{x^2+y^2} & \text{if } (x, y) \neq (0, 0) \\ 0 & \text{if } (x, y) = (0, 0) \end{cases}$$

Here g is defined at $(0, 0)$ but g is still discontinuous there because $\lim_{(x,y)\to(0,0)} g(x, y)$ does not exist.

Example 23

Let

$$f(x,y) = \begin{cases} \frac{3x^2y}{x^2+y^2} & \text{if } (x,y) \neq (0,0) \\ 0 & \text{if } (x,y) = (0,0) \end{cases}$$

We know f is continuous for $(x,y) \neq (0,0)$ since it is equal to a rational function there. Also we have

$$\lim_{(x,y)\to(0,0)} f(x,y) = \lim_{(x,y)\to(0,0)} \frac{3x^2y}{x^2+y^2} = 0 = f(0,0)$$

Therefore f is continuous at $(0,0)$, and so it is continuous on $\mathbb{R}^2$.

Remark 5

Just as for functions of one variable, composition is another way of combining two continuous functions to get a third. In fact, it can be shown that if f is a continuous function of two variables and g is a continuous function of a single variable that is defined on the range of f, then the composite function $h = g \circ f$ defined by $h(x, y) = g(f(x, y))$ is also a continuous function.

Example 24

Where is the function $h(x, y) = \arctan(y/x)$ continuous?

Solution

The function $f(x, y) = y/x$ is a rational function and therefore continuous except on the line $x = 0$. The function $g(t) = \arctan(t)$ is continuous everywhere. So the composite function $g(f(x, y)) = \arctan(y/x) = h(x, y)$ is continuous except where $x = 0$.

Definition 12 (Functions of More Than Two Variables)

Let f be a function of n variables. If f is defined on a subset D of $\mathbb{R}^n$, then $\lim_{\mathbf{x}\to\mathbf{a}} f(\mathbf{x}) = L$ means that for every number $\varepsilon > 0$ there is a corresponding number $\delta > 0$ such that

$$\text{if } \mathbf{x} \in D \text{ and } 0 < |\mathbf{x} - \mathbf{a}| < \delta \text{ then } |f(\mathbf{x}) - L| < \varepsilon.$$

For the case $n = 2$, we have $\mathbf{x} = (x, y), \mathbf{a} = (a, b)$, and $|\mathbf{x} - \mathbf{a}| = \sqrt{(x-a)^2 + (y-b)^2}$.
In each case the definition of continuity can be written as

$$\lim_{\mathbf{x}\to\mathbf{a}} f(\mathbf{x}) = f(\mathbf{a}).$$

Contents

Definition 13

If f is a function of two variables, its partial derivatives are the functions f_x and f_y defined by

$$f_x(x, y) = \lim_{h \to 0} \frac{f(x + h, y) - f(x, y)}{h}$$
$$f_y(x, y) = \lim_{h \to 0} \frac{f(x, y + h) - f(x, y)}{h}$$

provided the limits exist.

Remark 6 (Notations for Partial Derivatives)

If $z = f(x, y)$, we write

$$f_x(x, y) = f_x = \frac{\partial f}{\partial x} = \frac{\partial}{\partial x} f(x, y) = \frac{\partial z}{\partial x} = f_1 = D_1 f = D_x f$$
$$f_y(x, y) = f_y = \frac{\partial f}{\partial y} = \frac{\partial}{\partial y} f(x, y) = \frac{\partial z}{\partial y} = f_2 = D_2 f = D_y f$$

Rule for Finding Partial Derivatives of $z = f(x, y)$

1. To find f_x, regard y as a constant and differentiate $f(x, y)$ with respect to x.
2. To find f_y, regard x as a constant and differentiate $f(x, y)$ with respect to y.

Example 25

If $f(x, y) = x^2 + y^2$, find f_x and f_y.

```
from sympy.abc import x, y
from sympy import diff, pprint
f = x**2 + 2 * y**2
dx = diff(f, x)
dy = diff(f, y)
print('Derivative of')
pprint(f)
print('with respect to x is')
pprint(dx)
print('and with respect to y is')
pprint(dy)
```

```
Derivative of
 2      2
x  + 2⋅y
with respect to x is
2⋅x
and with respect to y is
4⋅y
```

Definition 14 (Functions of More Than Two Variables)

If u is a function of n variables, $u = f(x_1, x_2, \ldots, x_n)$, its partial derivative with respect to the ith variable x_i is

$$\frac{\partial u}{\partial x_i} = \lim_{h \to 0} \frac{f(x_1, \ldots, x_{i-1}, x_i + h, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_1, \ldots, x_n)}{h}$$

provided the limit exists and we also write

$$\frac{\partial u}{\partial x_i} = \frac{\partial f}{\partial x_i} = f_{x_i} = f_i = D_i f$$

Example 26

Find f_x, f_y and f_z if $f(x, y, z) = e^{xy} \ln z$.

Higher Derivatives

If f is a function of two variables, then its partial derivatives f_x and f_y are also functions of two variables, so we can consider their partial derivatives $(f_x)_x\,(f_x)_y\,,(f_y)_n$ and $(f_y)_y$, which are called the **second partial derivatives** of f. If $z = f(x, y)$, we use the following notation:

$$(f_x)_x = f_{xx} = f_{11} = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right) = \frac{\partial^2 f}{\partial x^2} = \frac{\partial^2 z}{\partial x^2}$$
$$(f_x)_y = f_{xy} = f_{12} = \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right) = \frac{\partial^2 f}{\partial y \partial x} = \frac{\partial^2 z}{\partial y \partial x}$$
$$(f_y)_x = f_{yx} = f_{21} = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right) = \frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 z}{\partial x \partial y}$$
$$(f_y)_y = f_{yy} = f_{22} = \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial y}\right) = \frac{\partial^2 f}{\partial y^2} = \frac{\partial^2 z}{\partial y^2}$$

Thus the notation f_{xy} (or $\partial^2 f/\partial y \partial x$) means that we first differentiate with respect to x and then with respect to y, whereas in computing f_{yx} the order is reversed.

Theorem 8 (Clairaut's Theorem)

Suppose f is defined on a disk D that contains the point (a, b) If the functions f_{xy} and f_{yx} are both continuous on D, then

$$f_{xy}(a, b) = f_{yx}(a, b)$$

Example 27

Find the second partial derivatives of

$$f(x, y) = x^3 + x^2y^3 - 2y^2$$

Example 28

Calculate f_{xxyz} if $f(x, y, z) = \sin(3x + yz)$.

Example 29

Partial derivatives occur in partial differential equations that express certain physical laws. For instance, the partial differential equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

is called **Laplace's equation** after Pierre Laplace (1749-1827). Solutions of this equation are called **harmonic functions**; they play a role in problems of heat conduction, fluid flow, and electric potential. Show that the function $u(x, y) = e^x \sin y$ is a solution of Laplace's equation.

Example 30

The **wave equation**

$$\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2}$$

describes the motion of a waveform, which could be an ocean wave, a sound wave, a light wave, or a wave traveling along a vibrating string. For instance, if $u(x,t)$ represents the displacement of a vibrating violin string at time t and at a distance x from one end of the string (as in Figure 8), then $u(x,t)$ satisfies the wave equation. Here the constant a depends on the density of the string and on the tension in the string. Verify that the function $u(x,t) = \sin(x - at)$ satisfies the wave equation.

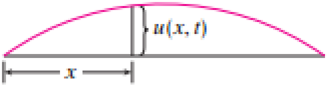


FIGURE 8

Contents

Theorem 9

Suppose f has continuous partial derivatives. An equation of the tangent plane to the surface $z = f(x, y)$ at the point $P(x_0, y_0, z_0)$ is

$$z - z_0 = f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0)$$

Example 31

Find the tangent plane to the elliptic paraboloid $z = 2x^2 + y^2$ at the point $(1, 1, 3)$.

Definition 15

The linear function whose graph is the tangent plane to the graph of a function f of two variables at the point $(a, b, f(a, b))$, namely

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

is called the **linearization** of f at (a, b) and the approximation

$$f(x, y) \approx f(a, b) + f_x(a, b)(x - a) + f_y(a, b)(y - b)$$

is called the **linear approximation** or the **tangent plane approximation** of f at (a, b).

Definition 16

If $z = f(x, y)$, then f is differentiable at (a, b) if the increment of z, Δz, can be expressed in the form

$$\Delta z = f_x(a,b)\Delta x + f_y(a,b)\Delta y + \varepsilon_1 \Delta x + \varepsilon_2 \Delta y$$

where ε_1 and $\varepsilon_2 \longrightarrow 0$ as $(\Delta x, \Delta y) \longrightarrow (0, 0)$.

Theorem 10

If the partial derivatives f_x and f_y exist near (a, b) and are continuous at (a, b), then f is differentiable at (a, b).

Example 32

Show that $f(x, y) = xe^{xy}$ is differentiable at $(1, 0)$ and find its linearization there. Then use it to approximate $f(1.1, -0.1)$.

Definition 17

For a differentiable function of two variables, $z = f(x, y)$, we define the differentials dx and dy to be independent variables; that is, they can be given any values. Then the differential dz, also called the **total differential**, is defined by

$$dz = f_x(x, y)dx + f_y(x, y)dy = \frac{\partial z}{\partial x}dx + \frac{\partial z}{\partial y}dy.$$

Example 33

(a) If $z = f(x, y) = x^2 + 3xy - y^2$, find the differential dz.

(b) If x changes from 2 to 2.05 and y changes from 3 to 2.96 , compare the values of Δz and dz.

Example 34

The base radius and height of a right circular cone are measured as 10 cm and 25 cm, respectively, with a possible error in measurement of as much as 0.1 cm in each. Use differentials to estimate the maximum error in the calculated volume of the cone.

Remark 7 (Functions of Three or More Variables)

Linear approximations, differentiability, and differentials can be defined in a similar manner for functions of more than two variables. For such functions the linear approximation is

$$f(x,y,z) \approx f(a,b,c) + f_x(a,b,c)(x-a) + f_y(a,b,c)(y-b) + f_z(a,b,c)(z-c)$$

and the linearization $L(x,y,z)$ is the right side of this expression. If $w = f(x,y,z)$, then the **increment** of w is

$$\Delta w = f(x+\Delta x, y+\Delta y, z+\Delta z) - f(x,y,z).$$

The **differential** dw is defined in terms of the differentials dx, dy , and dz of the independent variables by

$$dw = \frac{\partial w}{\partial x}dx + \frac{\partial w}{\partial y}dy + \frac{\partial w}{\partial z}dz.$$

Example 35

The dimensions of a rectangular box are measured to be 75 cm, 60 cm, and 40 cm, and each measurement is correct to within 0.2 cm. Use differentials to estimate the largest possible error when the volume of the box is calculated from these measurements.

Theorem 11 (The Chain Rule (Case 1))

Suppose that $z = f(x, y)$ is a differentiable function of x and y, where $x = g(t)$ and $y = h(t)$ are both differentiable functions of t. Then z is a differentiable function of t and

$$\frac{dz}{dt} = \frac{\partial z}{\partial x}\frac{dx}{dt} + \frac{\partial z}{\partial y}\frac{dy}{dt}$$

Example 36

If $z = x^2y + 3xy^4$, where $x = \sin 2t$ and $y = \cos t$, find dz/dt when $t = 0$.

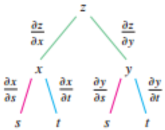
Example 37

The pressure P (in kilopascals), volume V (in liters), and temperature T (in kelvins) of a mole of an ideal gas are related by the equation $PV = 8.31T$. Find the rate at which the pressure is changing when the temperature is $300\mathbf{K}$ and increasing at a rate of $0.1\mathbf{K/s}$ and the volume is $100\mathbf{L}$ and increasing at a rate of $0.2\mathbf{L/s}$.

Theorem 12 (The Chain Rule (Case 2))

Suppose that $z = f(x, y)$ is a differentiable function of x and y, where $x = g(s, t)$ and $y = h(s, t)$ are differentiable functions of s and t. Then

$$\frac{\partial z}{\partial s} = \frac{\partial z}{\partial x}\frac{\partial x}{\partial s} + \frac{\partial z}{\partial y}\frac{\partial y}{\partial s} \text{ and } \frac{\partial z}{\partial t} = \frac{\partial z}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial z}{\partial y}\frac{\partial y}{\partial t}$$



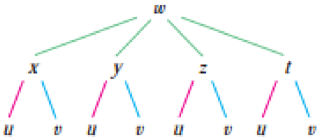
Example 38

If $z = e^x \sin y$, where $x = st^2$ and $y = s^2 t$, find $\partial z/\partial s$ and $\partial z/\partial t$.

Theorem 13 (The Chain Rule (General Version))

Suppose that u is a differentiable function of the n variables $x_1, x_2, \ldots, x_n$ and each x_j is a differentiable function of the m variables $t_1, t_2, \ldots, t_m$. Then u is a function of $t_1, t_2, \ldots, t_m$ and

$$\frac{\partial u}{\partial t_i} = \frac{\partial u}{\partial x_1}\frac{\partial x_1}{\partial t_i} + \frac{\partial u}{\partial x_2}\frac{\partial x_2}{\partial t_i} + \ldots + \frac{\partial u}{\partial x_n}\frac{\partial x_n}{\partial t_i}, \qquad \texttt{for} \quad i = 1, 2, \ldots, m.$$



Example 39

Write out the Chain Rule for the case where $w = f(x, y, z, t)$ and $x = x(u, u), y = y(u, v), z = z(u, v)$ and $t = t(u, v)$.

Example 40

If $u = x^4 y + y^2 z^3$, where $x = rse^t, y = rs^2 e^{-t}$, and $z = r^2 s \sin t$, find the value of $\dfrac{\partial u}{\partial s}$ when $r = 2, s = 1, t = 0$.

Example 41

If $g(s,t) = f(s^2 - t^2, t^2 - s^2)$ and f is differentiable, show that g satisfies the equation

$$t\frac{\partial g}{\partial s} + s\frac{\partial g}{\partial t} = 0$$

Example 42

If $z = f(x,y)$ has continuous second-order partial derivatives and $x = r^2 + s^2$ and $y = 2sr$, find $\partial z/\partial r$ and $\partial^2 z/\partial r^2$.

Theorem 14 (IMPLICIT DIFFERENTIATION)

If the equation $F(x, y) = 0$ defines y implicitly as a differentiable function of x, then

$$\frac{dy}{dx} = -\frac{F_x(x, y)}{F_y(x, y)}, \qquad F_y(x, y) \neq 0.$$

If the equation $F(x, y, z) = 0$ defines z implicitly as a differentiable function of x and y,then

$$\frac{\partial z}{\partial x} = -\frac{F_x(x, y, z)}{F_z(x, y, z)} \text{ and } \frac{\partial z}{\partial y} = -\frac{F_y(x, y, z)}{F_z(x, y, z)}, \qquad F_z(x, y, z) \neq 0.$$

Example 43

(a) Find dy/dx if $y^3 + y^2 - 5y - x^2 + 4 = 0$.

(b) Find $\partial z/\partial x$ and $\partial z/\partial y$ if $3x^2z - x^2y^2 + 2z^3 + 3yz - 5 = 0$.

Contents

Definition 18

The **directional derivative** of f at (x_0, y_0) in the direction of a unit vector $\mathbf{u} = (a, b)$ is

$$D_{\mathbf{u}} f(x_0, y_0) = \lim_{h \to 0} \frac{f(x_0 + ha, y_0 + hb) - f(x_0, y_0)}{h}$$

if this limit exists.

Remark 8

If $\mathbf{u} = \mathbf{i} = (1, 0)$, then $D_{\mathbf{i}} f = f_x$ and If $\mathbf{u} = \mathbf{j} = (0, 1)$, then $D_{\mathbf{j}} f = f_y$. In other words, the partial derivatives of f with respect to x and y are just special cases of the directional derivative.

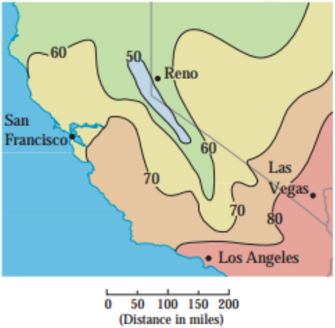


FIGURE 1

Example 44

Use the weather map in Figure 1 to estimate the value of the directional derivative of the temperature function at Reno in the southeasterly direction.

Theorem 15

If f is a differentiable function of x and y, then f has a directional derivative in the direction of any unit vector $\mathbf{u} = (a, b)$ and

$$D_{\mathbf{u}} f(x, y) = f_x(x, y)a + f_y(x, y)b.$$

Example 45

Find the directional derivative $D_{\mathbf{u}} f(x, y)$ if $f(x, y) = x^3 - 3xy + 4y^2$ and $\mathbf{u}$ is the unit vector given by angle $\theta = \pi/6$. What is $D_{\mathbf{u}} f(1, 2)$?

Definition 19

If f is a function of two variables x and y, then the **gradient** of f denoted by ∇f(read "del f") or $\mathbf{grad} f$ is the vector function defined by

$$\nabla f(x, y) = (f_x(x, y), f_y(x, y)) = \frac{\partial f}{\partial x}\mathbf{i} + \frac{\partial f}{\partial y}\mathbf{j}$$

Example 46

Find the gradient of $f(x, y) = y \ln x + xy^2$ at the point $(1, 2)$.

Theorem 16

If f is a differentiable function of x and y, then the directional derivative of f in the direction of the unit vector $\mathbf{u}$ is

$$D_{\mathbf{u}} f(x, y) = \nabla f(x, y).\mathbf{u}$$

Example 47

Find the directional derivative of the function $f(x, y) = x^2y^3 - 4y$ at the point $(2, -1)$ in the direction of the vector $\mathbf{v} = 2\mathbf{i} + 5\mathbf{j}$.

Example 48

Find the directional derivative of $f(x, y) = 3x^2 - 2y^2$ at the point $(-3/4, 0)$ in the direction from $P(-3/4, 0)$ to $Q(0, 1)$.

Definition 20 (Functions of More Than Two Variables)

Let f be a function of n variables $\mathbf{x} = (x_1, x_2, \ldots, x_n)$.

- The **directional derivative** of f at the point $\mathbf{x}_0$ in the direction of a unit vector $\mathbf{u}$ is

$$D_{\mathbf{u}} f(\mathbf{x}_0) = \lim_{h \to 0} \frac{f(\mathbf{x}_0 + h\mathbf{u}) - f(\mathbf{x}_0)}{h}$$

  if this limit exists.

- The **gradient vector** ∇f is

$$\nabla f = (f_{x_1}, f_{x_2}, \ldots, f_{x_n})$$

Theorem 17

If f is a differentiable function, then the directional derivative of f at the point $\mathbf{x}_0$ in the direction of a unit vector $\mathbf{u}$ is

$$D_{\mathbf{u}} f(\mathbf{x}_0) = \nabla f(\mathbf{x}_0).\mathbf{u}$$

Example 49

If $f(x, y, z) = x\sin(yz)$, (a) find the gradient of f and (b) find the directional derivative of f at $(1, 3, 0)$ in the direction of $\mathbf{v} = \mathbf{i} + 2\mathbf{j} - \mathbf{k}$.

Theorem 18 (Applications of Directional Derivatives)

Suppose f is a differentiable function of two or three variables.

- The function f increases most rapidly when $\mathbf{u}$ has the same direction as ∇f. The maximum value of the directional derivative $D_{\mathbf{u}}f(\mathbf{x})$ is $|\nabla f(\mathbf{x})|$.
- The function f decrease most rapidly in the direction of $-\nabla f$. The minimum value of the directional derivative $D_{\mathbf{u}}f(\mathbf{x})$ is $-|\nabla f(\mathbf{x})|$.
- Any direction $\mathbf{u}$ orthogonal to a gradient vector $\nabla f \neq 0$ is a direction of zero change in f.

Example 50

Find the directions in which $f(x, y) = (x^2/2) + (y^2/2)$

(a) Increases most rapidly at the point $(1, 1)$

(b) Decreases most rapidly at the point $(1, 1)$

(c) What are the directions of zero change in f at $(1, 1)$?

Example 51

(a) If $f(x, y) = xe^y$, find the rate of change of f at the point $P(2, 1)$ in the direction from P to $Q(1/2, 2)$.

(b) In what direction does f have the maximum rate of change? What is the maximum rate of change?

Example 52

Suppose that the temperature at a point (x, y, z) in space is given by $T(x, y, z) = 80/(1 + x^2 + 2y^2 + 3z^2)$, where T is measured in degrees Celsius and x, y, z in meters. In which direction does the temperature increase fastest at the point $(1, 1, -2)$? What is the maximum rate of increase?

Tangent Planes to Level Surfaces

Suppose S is a surface with equation $F(x, y, z) = c$,that is, it is a **level surface** of a function F of three variables, and let $P(x_0, y_0, z_0)$ be a point on S. Let C be any curve that lies on the surface S and passes through the point P. The curve $C$is described by a continuous vector function $\mathbf{r}(t) = (x(t), y(t), z(t))$. Let t_0 be the parameter value corresponding to P; that is, $\mathbf{r}(t_0) = (x_0, y_0, z_0)$.

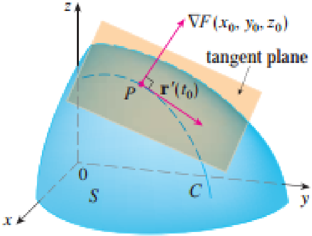


FIGURE 9

Definition 21 (Tangent Plane, Normal Line)

- The **tangent plane** at the point $P(x_0, y_0, z_0)$ on the level surface $F(x, y, z) = c$ of a differentiable function F is the plane through P normal to $\nabla F(x_0, y_0, z_0)$.
- The **normal line** of the surface at P is the line through P parallel to $\nabla F(x_0, y_0, z_0)$.

Theorem 19

- The tangent plane to $F(x, y, z) = c$ at $P(x_0, y_0, z_0)$ has the equation $F_x(x_0, y_0, z_0)(x-x_0)+F_y(x_0, y_0, z_0)(y-y_0)+F_z(x_0, y_0, z_0)(z-z_0) = 0$
- The normal line to $F(x, y, z) = c$ at $P(x_0, y_0, z_0)$ has the equation $x = x_0+F_x(x_0, y_0, z_0)t, y = y_0+F_y(x_0, y_0, z_0)t, z = z_0+F_z(x_0, y_0, z_0)t$

Example 53

Find the tangent plane and normal line of the surface

$$F(x, y, z) = x^2 + y^2 + z - 9 = 0$$

at the point $P(1, 2, 4)$.

Example 54

Find the equations of the tangent plane and normal line at the point $P(-2, 1, -3)$ to the ellipsoid

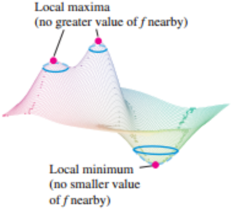
$$\frac{x^2}{4} + y^2 + \frac{z^2}{9} = 3.$$

Contents

Definition 22 (Local Maximum, Local Minimum)

Let $f(x, y)$ be a function defined on a region R containing the point (a, b). Then

1. $f(a, b)$ is a **local maximum** value of f if $f(a, b) \geq f(x, y)$ for all points (x, y) in an open disk centered at (a,b).
2. $f(a, b)$ is a **local minimum** value of f if $f(a, b) \leq f(x, y)$ for all points (x, y) in an open disk centered at (a,b).



Theorem 20 (First Derivative Test for Local Extreme Values)

If $f(x, y)$ has a local maximum or minimum value at an interior point (a, b) of its domain and if the first partial derivatives exist there, then

$$f_x(a, b) = 0 \text{ and } f_y(a, b) = 0.$$

Definition 23 (Critical Point)

An interior point of the domain of a function $f(x, y)$ where both f_x and f_y are zero or where one or both of f_x and f_y do not exist is a **critical point** of f.

Definition 24 (Saddle Point)

A differentiable function $f(x, y)$ has a **saddle point** at a critical point (a, b) if in every open disk centered at (a, b) there are domain points (x, y) where $f(x, y) > f(a, b)$ and domain points (x, y) where $f(x, y) < f(a, b)$. The corresponding point $(a, b, f(a, b))$ on the surface $z = f(x, y)$ is called a saddle point of the surface.

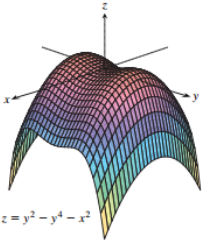


FIGURE 14.40 Saddle points at the origin.

Theorem 21 (Second Derivative Test for Local Extreme Values)

Suppose that $f(x, y)$ and its first and second partial derivatives are continuous throughout a disk centered at (a, b) and that $f_x(a, b) = f_y(a, b) = 0$. Then

1. f has a **local maximum** at (a, b) if $f_{xx} < 0$ and $f_{xx}f_{yy} - f_{xy}^2 > 0$ at (a, b).
2. f has a **local minimum** at (a, b) if $f_{xx} > 0$ and $f_{xx}f_{yy} - f_{xy}^2 > 0$ at (a, b).
3. f has a **saddle point** at (a, b) if $f_{xx}f_{yy} - f_{xy}^2 < 0$ at (a, b).
4. **The test is inconclusive** at (a, b) if $f_{xx}f_{yy} - f_{xy}^2 = 0$ at (a, b). In this case, we must find some other way to determine the behavior of f at (a, b).

The expression $f_{xx}f_{yy} - f_{xy}^2$ is called the **discriminant** or **Hessian** of f. It is sometimes easier to remember it in determinant form,

$$f_{xx}f_{yy} - f_{xy}^2 = \begin{vmatrix} f_{xx} & f_{xy} \\ f_{yx} & f_{yy} \end{vmatrix}$$

Example 55

Determine the local extrema and/or saddle points of the following functions

(a) $f(x, y) = xy - x^2 - y^2 - 2x - 2y + 4.$

(b) $f(x, y) = xy.$

(c) $f(x, y) = -x^3 + 4xy - 2y^2 + 1.$

(d) $f(x, y) = x^4 + y^4 - 4xy + 1.$

(e) $f(x, y) = 10x^2y - 5x^2 - 4y^2 - x^4 - 2y^4.$

Example 56

Find the shortest distance from the point $(1, 0, -2)$ to the plane $x + 2y + z = 4$.

Example 57

A rectangular box without a lid is to be made from $12m^2$ of cardboard. Find the maximum volume of such a box.

Theorem 22 (Extreme Value Theorem for Functions of Two Variables)

If f is continuous on a closed, bounded set D in $\mathbb{R}^2$, then f attains an absolute maximum value $f(x_1, y_1)$ and an absolute minimum value $f(x_2, y_2)$ at some points (x_1, y_1) and (x_2, y_2) in D.

Method

To find the absolute maximum and minimum values of a continuous function f on a closed, bounded set D:

1. Find the values of f at the critical points of f in D.
2. Find the extreme values of f on the boundary of D.
3. The largest values from steps 1 and 2 is the absolute maximum value; the smallest values is the absolute minimum value.

Example 58

(a) Find the absolute maximum and minimum values of the function $f(x, y) = x^2 - 2xy + 2y$ on the rectangle $D = \{(x, y)|0 \leq x \leq 3, 0 \leq y \leq 2\}$.

(b) Find the absolute maximum and minimum values of $f(x, y) = 2 + 2x + 2y - x^2 - y^2$ on the triangular region in the first quadrant bounded by the lines $x = 0, y = 0, y = 9 - x$.

Contents

Theorem 23 (The Orthogonal Gradient Theorem)

Suppose that $f(x, y, z)$ is differentiable in a region whose interior contains a smooth curve

$$C : \mathbf{r}(t) = g(t)\mathbf{i} + h(t)\mathbf{j} + k(t)\mathbf{k}.$$

If $P_0(x_0, y_0, z_0)$ is a point on C where f has a local maximum or minimum relative to its values on C, then ∇f is orthogonal to C at P_0.

Remark 9

Theorem 16 shows that $\nabla f(x_0, y_0, z_0)$ is orthogonal to the tangent vector $\mathbf{r}'(t_0)$ to every such curve C. But we already know that $\nabla g(x_0, y_0, z_0)$, is also orthogonal to $\mathbf{r}'(t_0)$ for every such curve. This means that the gradient vectors $\nabla f(x_0, y_0, z_0)$ and $\nabla g(x_0, y_0, z_0)$ must be parallel. Therefore, if $\nabla g(x_0, y_0, z_0) \neq 0$, there is a number λ such that $\nabla f = \lambda \nabla g$. The number λ is called a **Lagrange multiplier**.

Method of Lagrange Multipliers

To find the maximum and minimum values of $f(x, y, z)$ subject to the constraint $g(x, y, z) = k$ [assuming that these extreme values exist and $\nabla g \neq 0$ on the surface $g(x, y, z) = k$]:

(a) Find all values of x, y, z and λ such that

$$\nabla f(x, y, z) = \lambda \nabla g(x, y, z) \text{ and } g(x, y, z) = k$$

(b) Evaluate f at all the points (x, y, z) that result from step (a). The largest of these values is the maximum value of f; the smallest is the minimum value of f.

Example 59

A rectangular box without a lid is to be made from $12m^2$ of cardboard. Find the maximum volume of such a box.

Example 60

Find the extreme values of the function $f(x,y) = x^2 + 2y^2$ on the circle $x^2 + y^2 = 1$.

Example 61

Find the extreme values of $f(x,y) = x^2 + 2y^2$ on the disk $x^2 + y^2 \leq 1$.

Example 62

Find the minimum value of $f(x,y,z) = 2x^2 + y^2 + 3z^2$ (Objective function) subject to the constraint $2x - 3y - 4z = 49$.

Example 63

Find the points on the sphere $x^2 + y^2 + z^2 = 4$ that are closest to and farthest from the point $(3, 1, -1)$.

Two Constraints

Suppose that we want to find the maximum and minimum values of a function $f(x, y, z)$ subject to two constraints (side conditions) of the form $g(x, y, z) = k$ and $h(x, y, z) = c$. Geometrically, this means that we are looking for the extreme values of f when (x, y, z) is restricted to lie on the curve of intersection C of the level surfaces $g(x, y, z) = k$ and $h(x, y, z) = c$. (See Figure 5.) Suppose f has such an extreme value at a point $P(x_0, y_0, z_0)$. We know from the beginning of this section that ∇f is orthogonal to C at P. But we also know that ∇g is orthogonal to $g(x, y, z) = k$ and ∇h is orthogonal to $h(x, y, z) = c$, so ∇g and ∇g are both orthogonal to C. This means that the gradient vector $\nabla f(x_0, y_0, z_0)$ is in the plane determined by $\nabla g(x_0, y_0, z_0)$ and $\nabla h(x_0, y_0, z_0)$. (We assume that these gradient vectors are not zero and not parallel.) So there are numbers λ and μ (called Lagrange multipliers) such that

$$\nabla f(x_0, y_0, z_0) = \lambda \nabla g(x_0, y_0, z_0) + \mu \nabla h(x_0, y_0, z_0)$$

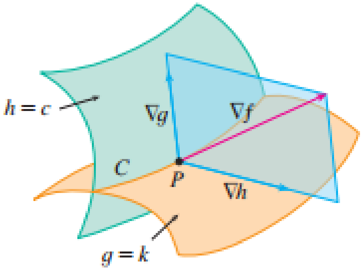


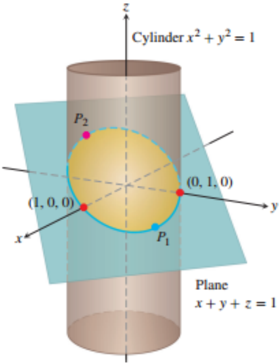
FIGURE 5

Example 64

Find the maximum value of the function $f(x, y, z) = x + 2y + 3z$ on the curve of intersection of the plane $x - y + z = 1$ and the cylinder $x^2 + y^2 = 1$.

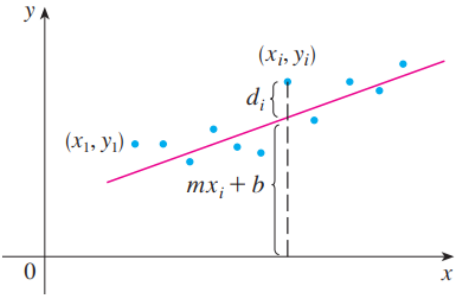
Example 65

The plane $x + y + z = 1$ cuts the cylinder $x^2 + y^2 = 1$ in an ellipse. Find the points on the ellipse that lie closest to and farthest from the origin.



Contents

Suppose that a scientist has reason to believe that two quantities x and y are related linearly, that is, $y = mx + b$, at least approximately, for some values of m and b . The scientist performs an experiment and collects data in the form of points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n),$ and then plots these points. The points don't lie exactly on a straight line, so the scientist wants to find constants m and b so that the line $y = mx + b$ "fits" the points as well as possible (see the figure).



Let $d_i = y_i - (mx_i + b)$ be the vertical deviation of the point (x_i, y_i) from the line. The **method of least squares** determines m and b so as to minimize $\sum_{i=1}^{n} d_i^2$, the sum of the squares of these deviations. Show that, according to this method, the line of best fit is obtained when

$$m \sum_{i=1}^{n} x_i + bn = \sum_{i=1}^{n} y_i$$
$$m \sum_{i=1}^{n} x_i^2 + b \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i$$

Thus the line is found by solving these two equations in the two unknowns m and b .

Theorem 24

The least squares regression line for $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ is given by $f(x) = mx + b$, where

$$m = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \quad \text{and} \quad b = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - m \sum_{i=1}^{n} x_i\right)$$

Proof

We want to find m and b such that the function

$$S(m,b) = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} \left(y_i - (mx_i + b)\right)^2 \text{ is minimized.}$$

- $$S_m(m,b) = \sum_{i=1}^{n} -2x_i\left(y_i - (mx_i + b)\right) \qquad (u^n)' = nu'u^{n-1}$$

- $$S_b(m,b) = \sum_{i=1}^{n} -2\left(y_i - (mx_i + b)\right)$$

- Solve the system $\begin{cases} S_m(m,b) = 0 \\ S_b(m,b) = 0 \end{cases}$

$$\Leftrightarrow \begin{cases} \sum_{i=1}^{n} (x_i y_i - m x_i^2 - b x_i) = 0 \\ \sum_{i=1}^{n} (y_i - m x_i - b) = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} \sum_{i=1}^{n} x_i y_i - m \sum_{i=1}^{n} x_i^2 - b \sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} y_i - m \sum_{i=1}^{n} x_i - nb = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} m\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i & (1) \\ m\sum_{i=1}^{n} x_i + bn = \sum_{i=1}^{n} y_i & (2) \end{cases}$$

$$\cdot\ \Delta = \begin{vmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{vmatrix} = n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2$$

$$\cdot\ \Delta_1 = \begin{vmatrix} \sum_{i=1}^{n} x_i y_i & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} y_i & n \end{vmatrix} = n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i$$

$$\cdot\ \Delta_2 = \begin{vmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} y_i \end{vmatrix} = \sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} x_i y_i$$

then $m = \dfrac{\Delta_1}{\Delta} = \dfrac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$

From (2), we get $b = \frac{1}{n}\left[\sum_{i=1}^{n} y_i - m\sum_{i=1}^{n} x_i\right]$.

$\cdot\ S_m(m,b) = \sum_{i=1}^{n} -2x_i\left(y_i - (mx_i + b)\right)$

$\cdot\ S_b(m,b) = \sum_{i=1}^{n} -2\left(y_i - (mx_i + b)\right)$

$\cdot\ S_{mm}(m,b) = \sum_{i=1}^{n} 2x_i^2 = 2\sum_{i=1}^{n} x_i^2$

$\cdot\ S_{mb}(m,b) = \sum_{i=1}^{n} 2x_i = 2\sum_{i=1}^{n} x_i$

$\cdot\ S_{bb}(m,b) = \sum_{i=1}^{n} 2 = 2n$

$$H = \begin{bmatrix} S_{mm} & S_{mb} \\ S_{mb} & S_{bb} \end{bmatrix} = \begin{bmatrix} 2\sum_{i=1}^{n} x_i^2 & 2\sum_{i=1}^{n} x_i \\ 2\sum_{i=1}^{n} x_i & 2n \end{bmatrix}$$

$H_1 = 2\sum_{i=1}^{n} x_i^2 > 0$, $|H| = 4n\sum_{i=1}^{n} x_i^2 - 4\left(\sum_{i=1}^{n} x_i\right)^2 = 4\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right] > 0$

$\frac{a_1}{b_1} = \cdots = \frac{a_n}{b_n}$

Cauchy-Schwarz inequality

$(a_1b_1 + \cdots + a_nb_n)^2 \leq (a_1^2 + \cdots + a_n^2)(b_1^2 + \cdots + b_n^2)$

$\Leftrightarrow \left(\sum_{i=1}^{n} a_ib_i\right)^2 \leq \sum_{i=1}^{n} a_i^2 \cdot \sum_{i=1}^{n} b_i^2$

By Cauchy-Schwarz inequality

$\left(\sum_{i=1}^{n} x_i \times 1\right)^2 \leq \sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{n} 1^2$

$\Leftrightarrow \left(\sum_{i=1}^{n} x_i\right)^2 \leq n\sum_{i=1}^{n} x_i^2$

$\Leftrightarrow n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \geq 0$

equality occures when $x_1 = \cdots = x_n$.

So $S(m,b)$ is minimized.

Example 66

The following are midterm and final examination test scores for 10 students from a calculus class, where x denotes the midterm score and y denotes the final score for each student.

x	68	87	75	91	82	77	86	82	75	79
y	74	79	80	93	88	79	97	95	89	92

(a) Calculate the least-squares regression line for these data.

(b) Plot the points and the least-squares regression line on the same graph.

Solution

(a) Find $\hat{y} = mx + b$; $m = \dfrac{\Delta_1}{\Delta} = \dfrac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$

$$b = \frac{1}{n}\left[\sum_{i=1}^{n} y_i - m\sum_{i=1}^{n} x_i\right]$$

We know $\sum_{i=1}^{n} x_i = 802$, $\sum_{i=1}^{n} y_i = 866$, $n = 10$

$$\sum_{i=1}^{n} x_i^2 = 64738, \quad \sum_{i=1}^{n} x_i y_i = 69742$$

We have $m = \dfrac{(10)(69742) - (802)(866)}{(10)(64738) - (802)^2} = 0.6915$

$$b = \frac{1}{10}\left[866 - (0.6915)(802)\right] = 31.1417$$

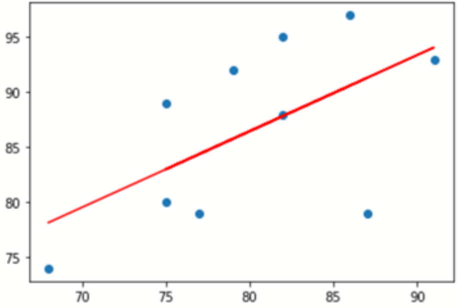
Thus $\hat{y} = 0.6915x + 31.1417$.

* Remark : If we devide the numerator and the dominator in m by n^2, we have

$$m = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\cdot\bar{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2} \quad \text{and} \quad b = \bar{y} - m\bar{x}.$$

(b).

```
Slope: 0.692
Intercept: 31.136
```



```python
import numpy as np
import matplotlib.pyplot as plt

def simple_linear_regression(x, y):
    n = len(x)
    x_mean = np.mean(x)
    y_mean = np.mean(y)
    xy_mean = np.mean(x * y)
    x_squared_mean = np.mean(x ** 2)

    slope = (xy_mean - x_mean * y_mean) / (x_squared_mean - x_mean ** 2)
    intercept = y_mean - slope * x_mean

    return slope, intercept

def predict(slope, intercept, x):
    return slope * x + intercept

# define dataset
x = np.array([68, 87, 75, 91, 82, 77, 86, 82, 75, 79])
y = np.array([74, 79, 80, 93, 88, 79, 97, 95, 89, 92])
# calculate coefficients
slope, intercept = simple_linear_regression(x, y)

# print coefficients
print('Slope: %.3f' % slope)
print('Intercept: %.3f' % intercept)

# plot dataset and line of best fit
plt.scatter(x, y)
plt.plot(x, predict(slope, intercept, x), color='red')
plt.show()
```

Suppose that we have n observations, each consisting of a y value and values of the k predictors (so each observation consists of $k+1$ numbers). Then we used the model called **multiple linear regression**:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon,$$

where $\varepsilon \sim N(0, \sigma^2)$, and the various ε's are independent of one another. Simple linear regression is the special case in which $k = 1$. We have then

$$\begin{bmatrix} y_1 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_k x_{1k} + \varepsilon_1 \\ . \\ . \\ . \\ \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \ldots + \beta_k x_{nk} + \varepsilon_n \end{bmatrix}$$

where x_{ij} is the jth independent variable for the ith observation, $i = 1, 2, \ldots, n,$ and ε_i's is are independent.

Define the following matrices:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & . & . & x_{1k} \\ 1 & x_{21} & x_{22} & . & . & x_{2k} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & x_{n1} & x_{n2} & . & . & x_{nk} \end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix}$$

Thus the n equations representing the linear equations can be rewritten in the matrix form as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

In particular, for the n observations from the simple linear model of the form

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

we can write

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdot & \\ \cdot & \\ \cdot & \\ 1 & x_n \end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix}$$

We now estimate $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ using the principle of least squares: Find $b_0, b_1, b_2, \ldots, b_k$ to minimize

$$\sum_{i=1}^{n} [y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_k x_{ik})]^2 = (\mathbf{Y} - \mathbf{Xb})^T(\mathbf{Y} - \mathbf{Xb}) = ||\mathbf{Y} - \mathbf{Xb}||^2$$

where $\mathbf{b}$ is the column vector with entries $b_0, b_1, \ldots, b_k$, and $||\mathbf{u}||$ is the length of $\mathbf{u}$.
If we equate to zero the partial derivative with respect to each of the coefficients, then it leads to the normal equations:

$$(\mathbf{X}^T\mathbf{X})\mathbf{b} = \mathbf{X}^T\mathbf{Y}$$

Assuming the matrix $(\mathbf{X}^T\mathbf{X})$ is invertible, we obtain

$$\hat{\boldsymbol{\beta}} = \mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

PROCEDURE TO OBTAIN A MULTIPLE LINEAR REGRESSION EQUATION

1. Rewrite the n observations

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_1 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i, i = 1, 2, \ldots, n$$

   in the matrix notation as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

2. Compute $(\mathbf{X}^T\mathbf{X})^{-1}$ and obtain the estimators of $\boldsymbol{\beta}$ as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

3. Then the regression equation is

$$\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

Example 67

The following data relate to the prices (Y) of five randomly chosen houses in a certain neighborhood, the corresponding ages of the houses (x_1), and square footage (x_2).

Price y in thousands of dollars	Age x_1 in years	Square footage x_2 in thousands of square feet
100	1	1
80	5	1
104	5	2
94	10	2
130	20	3

Fit a multiple linear regression model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ to the foregoing data.

Ans: $\hat{Y} = 66.12 - 0.3794x_1 + 21.4365x_2$.

Contents

The **gradient descent procedure** is an algorithm for finding the minimum of a function.

Suppose we have a function $f(x)$, where x is a tuple of several variables, i.e., $x = (x_1, x_2, \ldots, x_n)$. Also, suppose that the gradient of $f(x)$ is given by $\nabla f(x)$. We want to find the value of the variables $(x_1, x_2, \ldots, x_n)$ that give us the minimum of the function. At any iteration t, we'll denote the value of the tuple x by $x[t]$. So $x[t][1]$ is the value of x_1 at iteration t, $x[t][2]$ is the value of x_2 at iteration t, etc.

We use the following notations.

- t = Iteration number
- T = Total iterations
- n = Total variables in the domain of f (also called the dimensionality of x)
- j = Iterator for variable number, e.g., x_j represents the j-th variable
- η = Learning rate
- $\nabla f(x[t])$ = Value of the gradient vector of f at iteration t

The training method

The steps for the gradient descent algorithm are given below. This is also called the training method.

1. Choose a random initial point x_{initial} and set $x[0] = x_{\text{initial}}$
2. For iterations $t = 1, \ldots, T$

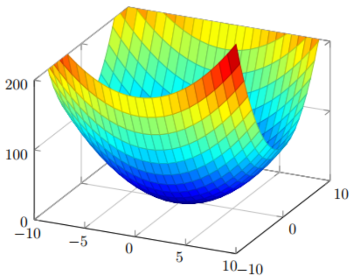
$$x[t] = x[t-1] - \eta \nabla f(x[t-1])$$

The learning rate η is a user defined variable for the gradient descent procedure. Its value lies in the range $[0, 1]$. The above method says that at each iteration we have to update the value of x by taking a small step in the direction of the negative of the gradient vector. If $\eta = 0$, then there will be no change in x. If $\eta = 1$, then it is like taking a large step in the direction of the negative of the gradient of the vector. Normally, η is set to a small value like 0.05 or 0.1. It can also be variable during the training procedure. So your algorithm can start with a large value (e.g.0.8) and then reduce it to smaller values.

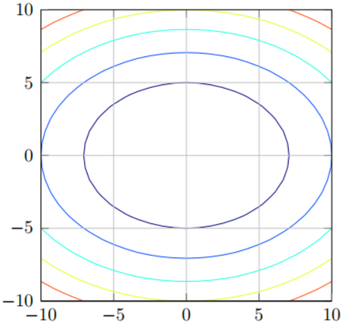
Example 68

Let's find the minimum of the following function of two variables, whose graphs and contours are shown in the figure below:

$$f(x, y) = x^2 + 2y^2$$



$f(x,y) = x^2 + 2y^2$



Contours of $f(x,y) = x^2 + 2y^2$

The general form of the gradient vector is given by:

$$\nabla f(x, y) = 2x\mathbf{i} + 4y\mathbf{j}$$

Two iterations of the algorithm, $T = 2$ and $\eta = 0.1$ are shown below:

1. Initial $t = 0$,

$$x[0] = (4, 3)$$

   (This is just a randomly chosen initial point)

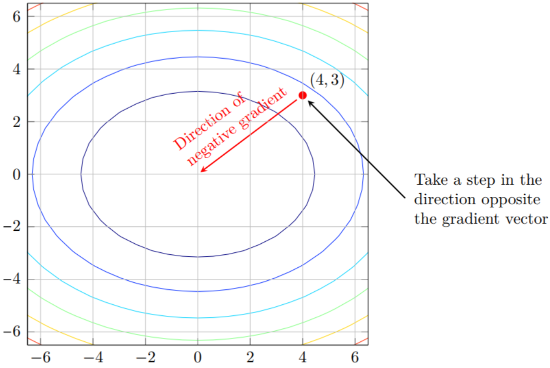
2. At $t = 1$,

$$\begin{aligned} x[1] &= x[0] - \eta \nabla f(x[0]) \\ &= (4, 3) - 0.1 \times (8, 12) \\ &= (3.2, 1.8) \end{aligned}$$

3. At $t = 2$,

$$\begin{aligned} x[2] &= x[1] - \eta \nabla f(x[1]) \\ &= (3.2, 1.8) - 0.1 \times (6.4, 7.2) \\ &= (2.56, 1.08) \end{aligned}$$

If you keep running the above iterations, the procedure will eventually end up at the point where the function is minimum, i.e., $(0, 0)$. At iteration $t = 1$, the algorithm is illustrated in the figure below:



Remark 10

1. Normally gradient descent is run till the value of x does not change or the change in x is below a certain threshold. The stopping criterion can also be a user defined maximum number of iterations (that we defined earlier as T).
2. Gradient descent can run into problems such as:
    1. Oscillate between two or more points
    2. Get trapped in a local minimum
    3. Overshoot and miss the minimum point
3. The gradient descent algorithm is often employed in machine learning problems. In many classification and regression tasks, the mean square error function is used to fit a model to the data. The gradient descent procedure is used to identify the optimal model parameters that lead to the lowest mean square error.
4. Gradient ascent is used similarly, for problems that involve maximizing a function.